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

Hsien-Po Shiang and Mihaela van der Schaar, *Senior Member, IEEE*

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 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 implicitly or explicitly model 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 multicarrier 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 with schemes that perform myopic best response.

Index Terms—Delay-sensitive applications, information feedback, interactive learning, 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 noncooperative 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 that are 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 multiuser case to maximize the overall throughput, the resource allocation problem becomes very complicated since the wireless mutual interference among users results in a nonconvex 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], noncooperative power control games were constructed, where each user possesses an energy-efficient utility function. The existence and the uniqueness of Nash equilibrium in such a noncooperative game were studied extensively. In [9] and [10], other than maximizing the throughput, users maximize a ratio of throughput over the transmitted power (measured in bits/joule). In [11] and [12], a pricing mechanism was employed to provide Pareto-efficient solutions [20] by adopting an additional penalty term that is associated with the power consumption in the utility function. In [13], a reinforcement learning approach for the noncooperative game is proposed, and the convergence property of the reinforcement approach was studied.

In short, previous research mainly concentrates on studying the existence and the performance of Nash equilibrium in
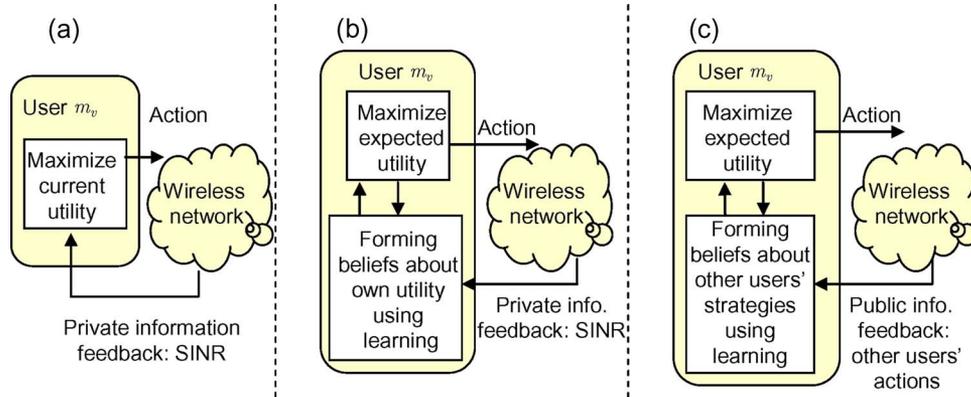
Fig. 1.   (a) Conventional distributed power control. (b) Payoff-based interactive learning with private information feedback. (c) Model-based interactive learning with public information feedback.

noncooperative 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 modeled accurately. Note that without a central coordinator, multiple users that are 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 noncooperative 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 noncooperative 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-plus-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 noncooperative 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 efficiently model the other users and, hence, improve the accuracy of the utility evaluation

resulting from taking different actions. Even when users are noncooperative, they can still reveal their action information to others 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, which means 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. Fig. 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.

We focus on the problem of delay-sensitive applications that are 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 the source traffic characteristics as well, including the source rates and the delay deadlines of the applications.

---

[1]In this paper, we will assume that the public information is transmitted accurately. However, if it is believed that malicious users are presented in the system, a mechanism design can be used to compel users to truthfully declare their information (see [6]).
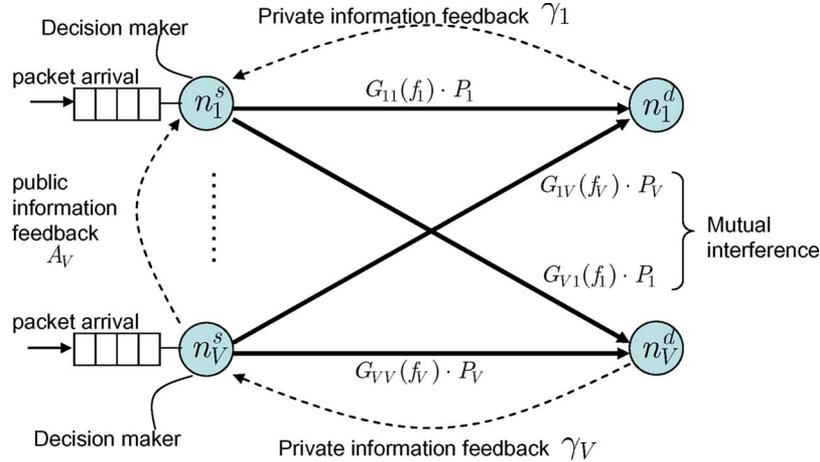
Fig. 2. System diagram of the dynamic joint power-spectrum resource allocation.

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 1) from whom (i.e., from which transmitters or receivers) this information is obtained and 2) 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 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 myopically optimizing 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].

This 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 that are 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 this 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 Fig. 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 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 $A_v \in \mathcal{A}_v$.

---

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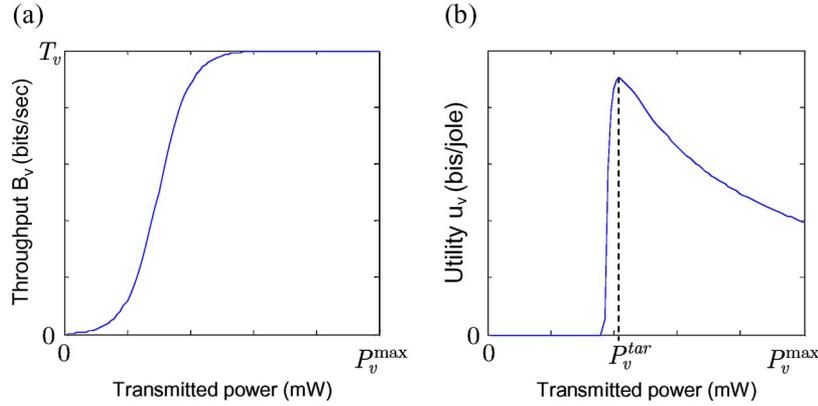
(a)

(b)

Fig. 3. (a) Throughput B_v versus P_v in a selected frequency channel f_v with fixed interference. (b) Utility u_v versus P_v in a selected frequency channel f_v with fixed interference.

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 user m_v. The SINR γ_v that is 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, which are denoted as A_{-v}, i.e.,

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

where N_{f_v} represents the additive white Gaussian 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 that is 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 the packet error rate of user m_v using the frequency channel f_v, respectively.

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 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 that cannot be transmitted 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 arrival rate λ_v and effective throughput $B_v(A_v, A_{-v})$. The packet loss rate is defined as the probability of when this delay exceeds the packet delay deadline, i.e., $\text{Prob}\{D_v(\lambda_v, B_v(A_v, A_{-v})) > d_v\}$. Therefore, the rate of successfully received packets is $\lambda_v \text{Prob}\{D_v(\lambda_v, B_v(A_v, A_{-v})) \leq d_v\}$.

D. Utility Function Definition

We assume that 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

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

The utility function reflects the expected number of packets that are successfully received (rather than transmitted as in [9]) per joule of energy that is 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 A. Fig. 3 illustrates the utility function of a user m_v using different power $0 \leq P_v \leq P_v^{\max}$ 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 $P_v^{\text{tar}}(f_v)$.

E. Problem Formulation

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

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

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 $\mathbf{A}_{-v}^{\text{fors}}(A_v)$, better performance can be achieved [20]. Let $\mathbf{A}_{-v}^{\text{fors}}(A_v)$ represent the actions of the other users given that action A_v is taken by user m_v. The optimization that is performed by a foresighted user can be formulated as [20]

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

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 $\mathbf{A}_{-v}^{\text{fors}}(A_v)$, the foresighted decision making based on the complete information of the other
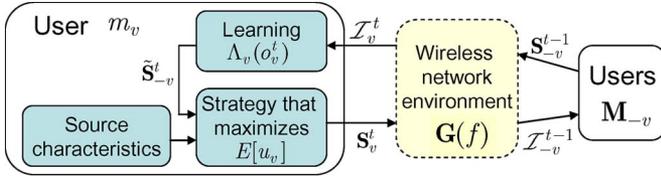
Fig. 4. Interactions among users and the foresighted decision making based on information feedback.

users will converge to the Stackelberg equilibrium [20], and the optimal utility is denoted as

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

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 $\mathbf{A}_{-v}^{\text{fors}}(A_v)$. Hence, accurately modeling the actions $\mathbf{A}_{-v}^{\text{fors}}(A_v)$ 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 of 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 $\mathbf{A}_{-v}^{\text{fors}}$ are not available to user m_v in real time, user m_v estimates $\mathbf{A}_{-v}^{\text{fors}}$ by building a *belief* on the other users' strategies $\mathbf{S}_{-v}^t$. The belief of user m_v is defined as $\tilde{\mathbf{S}}_{-v}^t(A_v) = \{\tilde{S}_{-v}^t(A_{-v}|A_v), \text{ for all } A_v \in \mathcal{A}_v\}$, where $\tilde{S}_{-v}^t(A_{-v}|A_v)^3$ 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' strategies $\tilde{\mathbf{S}}_{-v}^t(A_v)$ for each of its action $A_v \in \mathcal{A}_v$.[4]

Definition 3: Assume that Λ_v represents the interactive learning scheme that is 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 to estimate the actions of other users $\mathbf{A}_{-v}^{\text{fors}}$.

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 determines its own best response strategy $\mathbf{S}_v^t$. Fig. 4 illustrates how a delay-sensitive user makes decisions based on the observed information history o_v^t and the mutual

---

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

[4] Based on the 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 details.

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

---

interference coupling in the dynamic wireless environment. The problem in (4) can be now reformulated as

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

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

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 $\mathbf{A}_{-v}^{\text{fors}}$. A more accurate prediction of $\mathbf{A}_{-v}^{\text{fors}}$ can lead to 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.,

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

where

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

The notation $\tilde{S}_{-v}^t(A_{-v}|A_v)$ 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 $\mathbf{A}_{-v}^{\text{fors}}$, 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] $U_v(\mathbf{A}_{-v}^{\text{fors}})$ (where 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.,

$$\Delta_P \left(\Lambda_v \left(o_v^t \right) \right) \triangleq U_v \left(\mathbf{A}_{-v}^{\text{fors}} \right) - J_v \left(\Lambda_v \left(o_v^t \right) \right). \quad (9)$$

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

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

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.

1) *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 $\mathcal{I}_v^{t,\text{priv}} = \{\gamma_v^{t-1}\}$ or the public information feedback $\mathcal{I}_{-v}^{t,\text{pub}} = \{\mathbf{G}_u^{t-1}, A_u^{t-1}, \text{ for } m_u \in \mathbf{M}_{-v}\}$. Recall that $o_v^t = \{\mathcal{I}_v^t, o_v^{t-1}\}$ in Definition 1.

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

3) *Information feedback frequency.* In our problem formulation in (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 slot ($\tau_v \in \mathbb{Z}^+$). 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 user m_v obtains information and makes decisions, i.e., $\mathbf{T}_v^t = \{(s_v^0 + k\tau_v), k = 0, 1, \ldots, K_v^t\}$, where s_v^0 is the initial time slot that a user m_v obtains information and starts making decisions. The number of decisions that are made by user m_v up to time t is equal to $K_v^t = \lfloor (t - s_v^0)/\tau_v \rfloor$, 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, large information overhead $\sigma_v(\omega_v, |\mathbf{V}_v|)$ 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 that is dedicated to the packet transmission, and it is a decreasing function of σ_v. Given $\mathbf{A}_{-v}^{\text{fors}}$, the utility function in (2) can be derived as (see Appendix A for more details) given in (11) and (12), shown at the bottom of the page, where 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 nonincreasing function of σ_v.

Intuitively, if σ_v is large, the belief $\tilde{\mathbf{S}}_{-v}^t$ provides an accurate model on $\mathbf{A}_{-v}^{\text{fors}}$. On the other hand, if σ_v is small, the belief $\tilde{\mathbf{S}}_{-v}^t$ provides an inaccurate model on $\mathbf{A}_{-v}^{\text{fors}}$. 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

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

Proof: Note that minimizing Δ_P is the same as maximizing $J_v(\Lambda_v(o_v^t))$. Since $0 \le J_v(\Lambda_v(o_v^t)) \le U(\mathbf{A}_{-v}^{\text{fors}})$ 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. Fig. 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 detail.

---

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

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

$$F_v(\sigma_v, \gamma_v(A_v, A_{-v})) \equiv \exp\left(B_v'(\sigma_v, \gamma_v(A_v, A_{-v})) \frac{d_v}{L_v} - \lambda_v d_v \right) \tag{12}$$
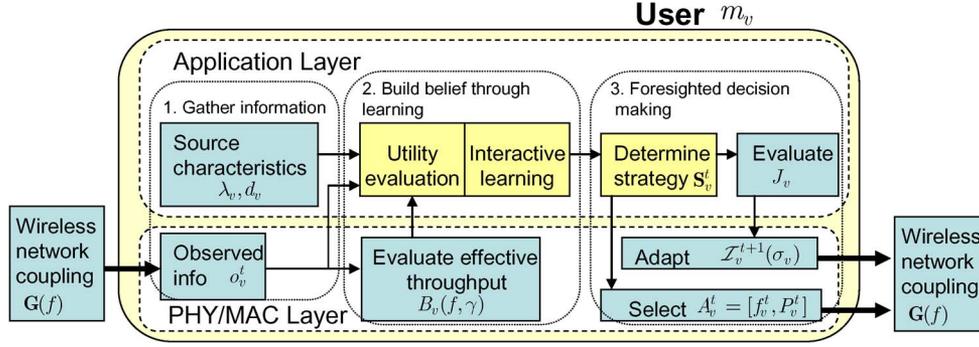
Fig. 5. System block diagram for the adaptive interactive learning for dynamic resource management.

IV. INTERACTIVE LEARNING WITH PRIVATE INFORMATION FEEDBACK

In the case where user m_v only observes the private information feedback $\mathcal{I}_v^{t,\mathrm{priv}}$, it can only model the aggregate effect of other users' actions through the experienced SINR value γ_v. Hence, it cannot explicitly model the exact response actions of the other users $\mathbf{A}_{-v}^{\mathrm{fors}}$. 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 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 explicitly predicting the exact response actions $\mathbf{A}_{-v}^{\mathrm{fors}}$, 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 $\tilde{S}_{-v}^t(A_{-v}|A_v)$ in (8). Instead, user m_v directly builds 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 $\sum_{A_{-v} \in \mathcal{A}^{V-1}} \tilde{S}_{-v}^t(A_{-v}|A_v) u_v(A_v, A_{-v})$ in (8).

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

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

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

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

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

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

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 of learning 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 (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 payoff-based learning with private information feedback, if the information feedback is costless, the upper bound of the learning efficiency $J_v(\Lambda_v^{\mathrm{priv}})$ is $(1 - \varepsilon_v(\Lambda_v^{\mathrm{priv}}))U_v(\mathbf{A}_{-v}^{\mathrm{fors}})$, with $0 \leq \varepsilon_v(\Lambda_v^{\mathrm{priv}}) < 1$, and

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

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

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

To increase the learning efficiency $J_v(\Lambda_v^{\mathrm{priv}})$, 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

Here, let us assume that $\omega_v = 1$. 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, i.e., $u_v(A_v^{t-1}, \gamma_v^{t-1})$, when the action A_v^{t-1} is taken at time slot $t-1$. Here, we adopt the cumulative payoff matching [19], i.e.,

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

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

$$I\left(A = A_v^t\right) = \begin{cases} 1, & \text{if } A = A_v^t \\ 0, & \text{if } A \neq A_v^t. \end{cases} \tag{20}$$

B. Adaptive Reinforcement Learning

The reinforcement learning in the previous section 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 $J_v(\Lambda_v^{\text{priv}})$. 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 section. 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}, i.e.,

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

$$u_v^{\text{base}} = \max\left(u_v^{\text{base}}, u_v\left(A_v^{t-1}, \gamma_v^{t-1}\right)\right). \tag{22}$$

Finally, user m_v evaluates the learning efficiency $J(\Lambda_v(o_v^t))$ and changes the information feedback frequency ω_v by $\Delta\omega_v$ 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 (AR).

Algorithm 1. Adaptive reinforcement learning with private information feedback

For user m_v at time slot t, assume that $\mathbf{U}(0, 1)$ represents a uniform distribution from 0 to 1.

Initialization: set $J_v^{\text{prev}} = 0$, $\omega_v = 1$, $\Delta\omega_v = 0.05$.

Step 1. If $Rand(\mathbf{U}(0, 1)) < 1 - \omega_v$, keep using action $A_v^t = A_v^{\text{base}}$, $t \leftarrow t + 1$, 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 $\mathcal{I}_v^{t,\text{priv}} = \{\gamma_v^{t-1}\}$.

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 the baseline payoff value u_v^{base} as in (21) and (22).

---

[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 Algorithm 1 (AR) scheme employs reinforcement learning *while* considering the cost of the information feedback and adapts the information feedback frequency to maximize the user's utility.

Step 6. Evaluate J_v. If $J_v > J_v^{\text{prev}}$, then
if $\omega_v - \Delta\omega_v > 0$, $\omega_v \leftarrow \omega_v - \Delta\omega_v$,
else if $\omega_v - \Delta\omega_v \leq 0$, keep ω_v.
Otherwise, if $\omega_v + \Delta\omega_v \leq 1$, $\omega_v \leftarrow \omega_v + \Delta\omega_v$,
else if $\omega_v - \Delta\omega_v > 1$, keep ω_v.

Step 7. Set $J_v^{\text{prev}} \leftarrow J_v$, $t \leftarrow t + 1$, and go back to step 1.

V. INTERACTIVE LEARNING WITH PUBLIC INFORMATION FEEDBACK

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

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

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

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

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

Model-based learning [19] provides a method of building the belief on $\tilde{S}_{-v}^t(\sigma_v)$ 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 (24) as an example in Section V-A.

Similarly, if the public information feedback is costless [i.e., $B_v' = B_v$ in (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 $J_v(\Lambda_v^{\text{pub}})$ is $U_v(\mathbf{A}_{-v}^{\text{fors}})$.

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

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

∎

The reason why the model-based learning with public information feedback has a higher upper bound compared with 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 to build the belief $\tilde{\mathbf{S}}_{-v}^t$ from $o_v^t = \{\mathcal{I}_{-v}^{s,\mathrm{pub}}, s \in \mathbf{T}_v^t\}$, user m_v maintains a set of strategy vectors $\tilde{S}_{-v}^t(A_{-v}|A_v) = \{\tilde{S}_u^t(A_u \in \mathcal{A}_u|A_v),$ for $m_u \in \mathbf{M}_{-v}\}$ for all possible actions $A_v \in \mathcal{A}_v$, where $\tilde{S}_u^t(A_u \in \mathcal{A}_u|A_v) = [\tilde{S}_u^t(A_u|A_v),$ for $A_u \in \mathcal{A}_u]$ represents the estimated strategy of user $m_u \in \mathbf{M}_{-v}$ given that user m_v takes action A_v at time slot t. Hence, in the action learning, whenever action A_v is taken by user m_v, we set

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

where $r_u^t(A_u|A_v)$ 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 action A_v is taken by user m_v, vector $\mathbf{r}_u^t(A_u \in \mathcal{A}_u|A_v) = [r_u^t(A_u|A_v),$ for all $A_u \in \mathcal{A}_u]$ is updated by

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

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

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

Proof: See Appendix B. ∎

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

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

Proposition 5. Optimal Actions Given the Target SINR Values: Let $F_v^{\mathrm{tar}}(f) = F_v(f, \gamma_v^{\mathrm{tar}}(f))$ in (12). Given the corresponding target $P_v^{\mathrm{tar}}(f)$, the optimal action A_v^* of a user m_v is

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

Proof: From Proposition 4, maximizing $u_v = (\lambda_v/P_v)(1 - (1/F_v))$ leads to (28). ∎

In summary, user m_v selects frequency channel f_v^* and power level P_v^* to support target SINR $\gamma_v^{\mathrm{tar}}(f_v^*)$, which maximizes the utility function in (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 action A_v is taken. Given $\tilde{\mathbf{S}}_{-v}^t$, $\Omega_v(A_v)$ can be computed as

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

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 target power $P_v^{\mathrm{tar}}(f)$ in different frequency channels, i.e.,

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

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 section, the public information feedback $\mathcal{I}_{-v}^{t,\mathrm{pub}} = \{\mathbf{G}_u^{t-1}, A_u^{t-1}, m_u \in \mathbf{M}_{-v}\}$ is required from every user in the network during each time slot. This results in heavy information overhead. Moreover, 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 $U_v(\mathbf{A}_{-v}^{\mathrm{fors}})$ 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 ($1 \leq H \leq |\mathbf{M}_{-v}|$) and assign different information feedback frequency ω_v^i to different groups (i.e., $1 \geq \omega_v^1 \geq \omega_v^2 \geq \cdots \geq \omega_v^H \geq 0$). For the dynamic power/spectrum management problem in this paper, the neighboring users can be classified based on their average channel gains $\overline{G}_{uv}$ over the frequency channels, i.e., $\overline{G}_{uv} = (1/|\mathcal{F}|) \sum_{f \in \mathcal{F}} G_{uv}(f)$ (from the transmitter of neighboring user m_u to the receiver of foresighted user m_v) since these channel gains directly impact the user's utility [see (1) and (2)]. For instance, a neighboring user m_u with a larger channel gain $\overline{G}_{uv}$ will have more impact on u_v.

Let X_v^i represent the number of users in the group H_i, $i = 1, \ldots, H$. Assume that the neighboring users are relabeled according to their average channel gain value, i.e., $\overline{G}_{[1]v} \geq \overline{G}_{[2]v} \geq \cdots \geq \overline{G}_{[V-1]v}$. Then

$$m_{[u]} \in H_i, \quad \text{iff} \sum_{j=1}^{i-1} X_v^j \leq [u] \leq \sum_{j=1}^{i} X_v^j. \tag{32}$$

In Algorithm 2 (AA), we provide our adaptive action learning approach for the extreme case when $H = 2$, as an example.

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 $|\mathbf{V}_v|$	$U_v(\mathbf{A}_{-v}^{fors})$

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, $\omega_v = 0$. That is, user m_v only needs to model the users in set $\mathbf{V}_v$ based on $\mathcal{I}_{-v(\mathbf{V}_v^t)}^{t,\text{pub}} = \{\mathbf{G}_u^{t-1}, A_u^{t-1}, m_u \in \mathbf{V}_v\}$. In Table I, we compare the two proposed interactive learning algorithms.

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

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

VI. SIMULATION RESULTS

We simulate an ad hoc wireless network environment, which is shown in Fig. 6, with five users (distinct transmitter–receiver pairs) and three 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 user m_v to the receiver of user $m_{v'}$, and $K_0 = 5 \times 10^{-4}$, $N_f = 1 \times 10^{-5}$, $dis_0 = 10$, and $\alpha = 2$ are constants. For the application layer parameters, we set the average packet length $L_v = 1000$ bytes, input rate $R_v = 500$ kb/s ($\lambda_v = R_v/L_v$), and delay deadline $d_v = 200$ ms for all the users. 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 A).
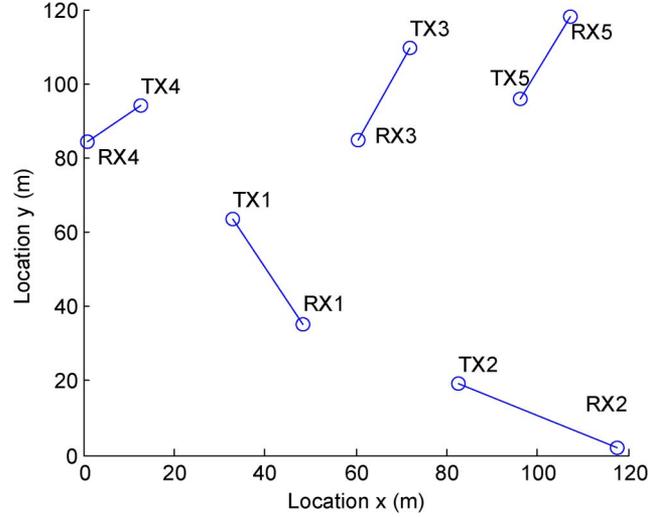

Fig. 6. Topology settings for the simulation.

A. Comparison Among Different Learning Approaches Based on Information Feedback

We show the simulation results using five different schemes when the physical transmission rates are $T = 700$ and 2100 kb/s in Tables II and III, respectively. The five schemes are 1) the centralized optimal (CO), 2) the theoretical upper bound $U(\mathbf{A}_{-v})$ (UB), 3) the myopic best response without learning (NE), 4) the user m_1 adopting adaptive reinforcement learning with private information feedback in AR, and 5) the user m_1 adopting adaptive action learning with public information feedback in AA. The CO scheme provides the global optimal results for the overall utilities. In the NE scheme, each user attempts to maximize its current utility function based on the actions that it observes in the previous time slot, as in (3). The UB is computed from (4) for m_1 given the exact response of the other four users ($u_1 = U(\mathbf{A}_{-1}^{\text{fors}})$). Since 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 (1).

Table II shows that user m_1 stays in channel 1 in the CO and UB schemes, whereas the other four users stay using the rest of the two channels. However, since users are self-interested, the NE scheme shows that user m_5 also attempts to transmit in channel 1, and, hence, 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 on 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 a significant information overhead, which results in worse performance at low bandwidth, i.e., when $T = 700$ kb/s. 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 noncooperative setting, this foresighted user actually benefits the overall system performance.

When $T = 2100$ kb/s, Table III shows that users are now selecting lower power levels since the physical transmission

TABLE II
SIMULATION RESULTS OF THE FIVE SCHEMES WHEN $T = 700$ kb/s

$T = 700$ 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 ($\omega_v = 0.7$)
	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 ($|\mathbf{V}_v| = 2$)
	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	

TABLE III
SIMULATION RESULTS OF THE FIVE SCHEMES WHEN $T = 2100$ kb/s

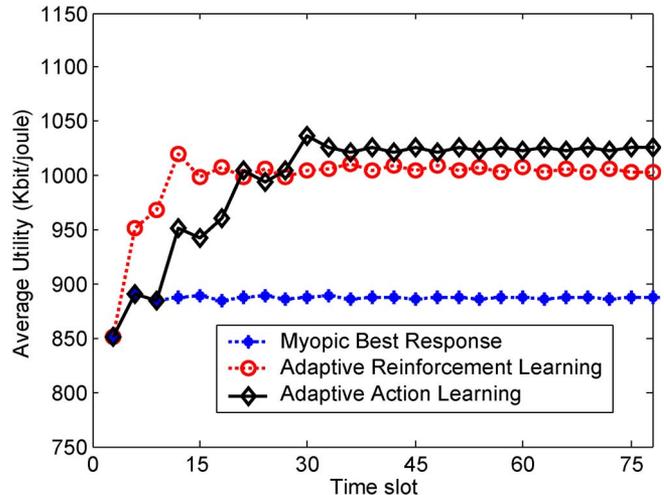| $T = 2100$ Kbps		Actions $A_v = [f_v, P_v]$ (or mixed strategies $\mathbf{S}_v$)	u_v (Kbit/joule)	$\sum_{v=1}^{5} u_v / 5$
1) Centralized Optimal (CO)	m_1	[1,2]	1562.2	1718.7
	m_2	[2,4]	781.2	
	m_3	[3,2]	1562.5	
	m_4	[2,1]	3125.0	
	m_5	[2,2]	1562.5	
2) Theoretical Upper Bound (UB)	m_1	[1,2]	$U_1(\mathbf{A}_{-v}^{fors})=$ 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 ($\omega_v = 1$)
	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 ($|\mathbf{V}_v| = 4$)
	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 versus time slot of the proposed algorithms when $T = 700$ kb/s (a time slot is considered to be 10 ms).

bandwidth is sufficient. Using the AR scheme, user m_1, again, occupies channel 1 by using a higher power level compared to that of 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 $T = 2100$ kb/s. Again, the average utilities of the adaptive interactive learning schemes outperform that of the myopic NE scheme. The higher T gives a better learning environment for user m_1 using the AA scheme to approach the theoretical upper bound $U_1(\mathbf{A}_{-v}^{\mathrm{fors}})$ than when using the AR scheme. Since all the users are selfish (including user m_1 who is learning), 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 T is small.

B. Convergence of the Learning Schemes

To show the convergence of the proposed learning schemes, in Fig. 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 in Table II when $T = 700$ kb/s. It is shown that the two proposed learning schemes outperform the myopic best response scheme in terms
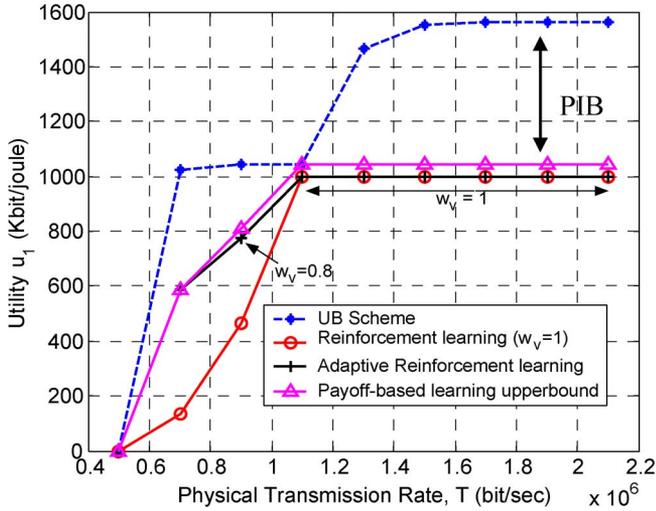
Fig. 8. Performance of user m_1 adopting adaptive reinforcement learning with private information feedback using different ω_1.

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

of the average utility. The convergence speed of the AR scheme is about three times slower than that of the myopic best response (which converges to Nash equilibrium in about five time slots), whereas that of the AA scheme is about six times slower. The convergence speed of the AR scheme is faster than that of the AA scheme since the AR scheme only needs 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.

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 user m_1's initial strategy by performing a myopic best response in the first 20 time slots. Then, we simulate the reinforcement learning with different values of ω_v in Fig. 8 for different T. Since the input rates of the applications are fixed to 500 kb/s, the utility will saturate as the bandwidth increases. The UB scheme has another saturation when T becomes larger than 1.1 Mb/s since the larger bandwidth enables another set of actions for the users. Note that when $\omega_1 = 1$, the reinforcement learning learns 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 that of $\omega_1 = 1$ when the physical bandwidth is lower than 1 Mb/s since learning at a slower pace can reduce the overhead of the private information feedback. The results in Fig. 8 show that the proposed adaptive reinforcement learning operates on the envelope of the solutions that are 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.

D. Adaptive Action Learning From Different Neighboring Users

In Fig. 9, we also simulate the case that the action learning models the strategy of the nearest $|\mathbf{V}_v| = 2$ users instead of
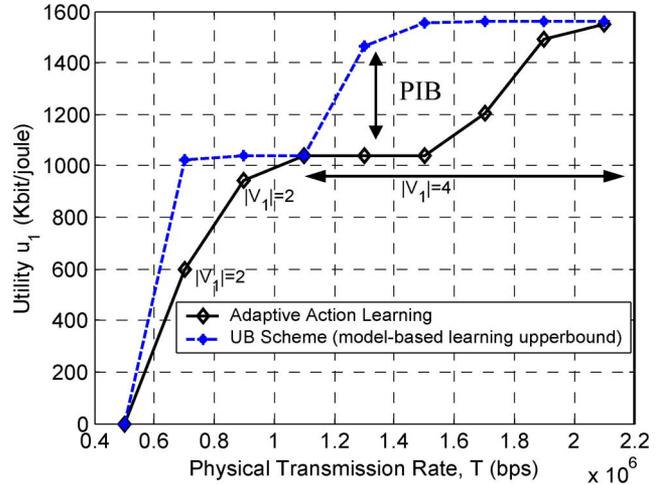
$|\mathbf{M}_{-v}| = 4$ users. With smaller $|\mathbf{V}_v|$, fewer neighbors need to feed back information and, hence, results in less information overhead. The simulation results show that modeling users from public information feedback can improve the performance of user m_1. However, when the physical transmission rate is lower than 1.1 Mb/s, the required information overhead significantly degrades the performance, 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.

E. Mobility Effect on the Interactive Learning Efficiency

In the previous sections, all the simulation results are based on the fixed topology that is shown in Fig. 6. Here, we simulate the case that all five receivers move according to the 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 Fig. 6, Fig. 10 shows the learning efficiency over time of the AR, AA, and NE schemes for $\nu = 0.5, 1, 2$ (m/s) with $T = 2100$ kb/s. It is shown that, on average, the AA scheme has higher learning efficiency 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. In particular, for the reinforcement learning without explicit channel gain information, the results show that the performance can be worse than that of the 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. CONCLUSION

In this paper, we have provided an adaptive interactive learning framework for delay-sensitive users to adapt their frequency channel selections and power levels in wireless networks in
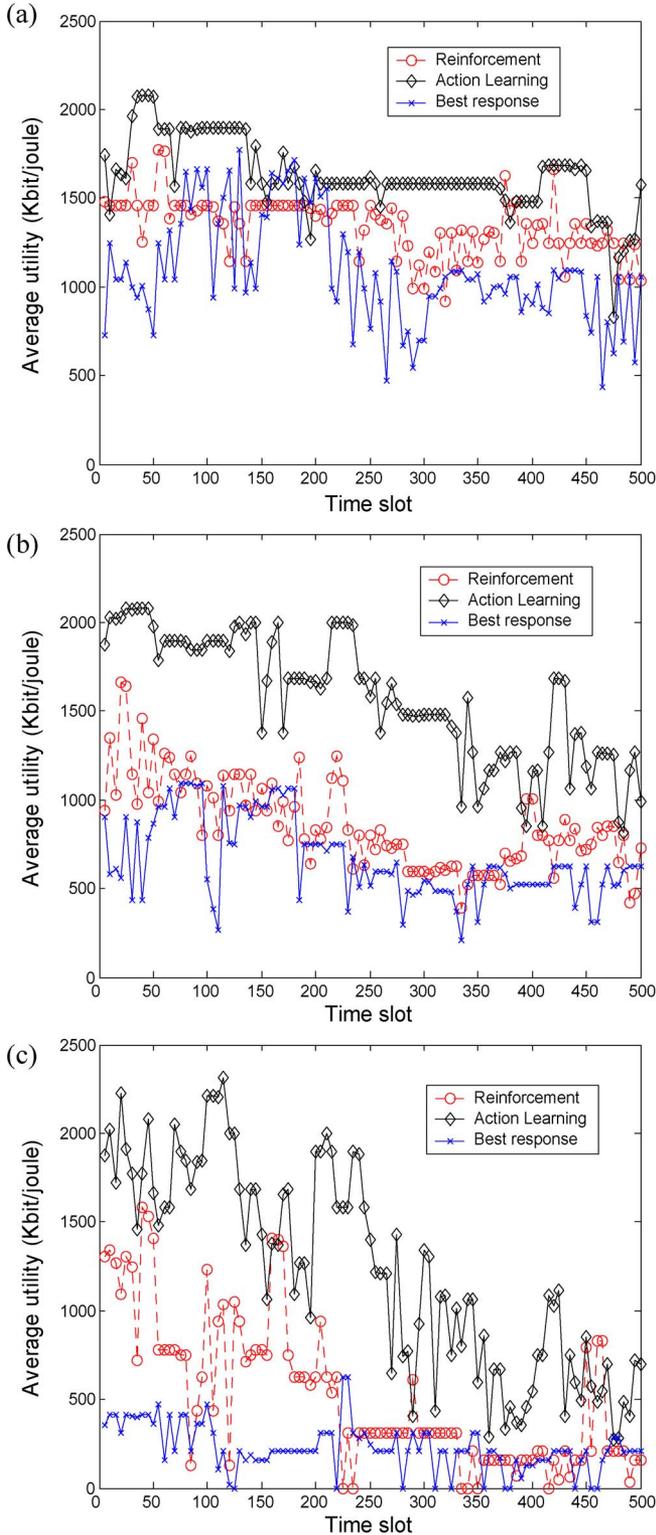
(a)

(b)

(c)

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

a decentralized manner. We have shown that a foresighted user can significantly improve its utility by learning from the information feedback. We have determined the 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 has been shown that even when only one user learns from its information feedback, the overall performance can be better than that of Nash equilibrium resulting from the myopic best response. In particular, if the available system bandwidth is not limited, the proposed adaptive action learning with public information feedback approaches the utility upper bound.

APPENDIX A

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

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

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

$B_v'(\sigma_v, A_v, A_{-v}) = B_v(A_v, A_{-v})\theta(\sigma_v)$, 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 user m_v through the network with average input rate R_v (bits per second). Assume that user m_v maintains a queue with an infinite buffer size in the application layer. We model the packet arrival process using the Poisson process. The packet arrival rate is assumed as $\lambda_v = R_v/L_v$ (packets per second). Considering the packet protection scheme that is 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 $\mu_v(\sigma_v, A_v, A_{-v}) = B_v'(\sigma_v, A_v, A_{-v})/L_v$ (packets per second). 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

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

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

$$\text{Prob}\left\{D_v(\sigma_v, A_v, A_{-v}) \leq d_v\right\} =$$
$$\begin{cases} 1 - \exp\left(-\frac{d_v}{E[D_v(\sigma_v, A_v, A_{-v})]}\right), & \text{for } \mu_v(\sigma_v, A_v, A_{-v}) > \lambda_v \\ 0, & \text{otherwise.} \end{cases}$$
$$(36)$$

The utility function in (2) is equal to zero, 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 $\text{Prob}\{D_v(\sigma_v, A_v, A_{-v}) \leq d_v\} > 0$ (see Fig. 2). Substituting (35) and (36) into (2), we have (11). Since $B'_v(\sigma_v)$ is a decreasing function of σ_v, the utility function is a nonincreasing function of σ_v.

APPENDIX B

Proof of Proposition 4: Given channel model $B_v(f, \gamma)$ for frequency channel f in (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 (35) and (36), we have $\text{Prob}\{D_v \leq d_v\} = 1 - (1/F_v(\gamma_v))$. The optimality condition of $(\partial u_v/\partial P_v) = 0$ becomes $-P_v \times (\partial/\partial P_v)(1/F_v(\gamma_v)) = 1 - (1/F_v(\gamma_v))$. The left-hand side can be derived as $\gamma_v(\partial B_v(\gamma_v)/\partial \gamma_v) \times (d_v/L_v)(1/F_v(\gamma_v))$ since $P_v(\partial \gamma_v/\partial P_v) = \gamma_v$. 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 J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 89–103, Jan. 2005.

[2] S. G. Kiani, G. E. Oien, and D. Gesbert, "Maximizing multicell capacity using distributed power allocation and scheduling," in *Proc. IEEE WCNC*, Mar. 2007, pp. 1690–1694.

[3] S. Singh and D. Bertsekas, "Reinforcement learning for dynamic channel allocation in cellular telephone systems," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1997, pp. 974–980.

[4] J. Zhao, H. Zheng, and G.-H. Yang, "Distributed coordination in dynamic spectrum allocation networks," in *Proc. IEEE DySPAN*, 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," *Proc. 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 and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 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 multicarrier CDMA systems," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 6, pp. 1115–1129, Jun. 2006.

[10] D. J. Goodman and N. B. Mandayam, "Power control for wireless data," *IEEE Pers. Commun.*, vol. 7, no. 11, 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. 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 Trans. Netw.*, vol. 11, no. 2, 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 J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 1101–1112, Aug. 2007.

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

[15] A. G. Konheim, "A queuing analysis of two ARQ protocols," *IEEE Trans. Commun.*, vol. COM-28, no. 7, pp. 1004–1014, Jul. 1980.

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

[17] H. Shiang and M. van der Schaar, "Informationally decentralized video streaming over multihop wireless networks," *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1299–1313, Sep. 2007.

[18] S. Lal and 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*. New York: Oxford Univ. Press, 2004.

[20] Y. Su and M. van der Schaar, "A new look at multi-user power control games," in *Proc. ICC*, Jun. 2008, pp. 1072–1076.

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

[22] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wirel. Commun. Mob. Comput.*, vol. 2, no. 5, pp. 483–502, Sep. 2002.

Hsien-Po Shiang is currently working toward the Ph.D. degree with the Department of Electrical Engineering, University of California, Los Angeles.

During his graduate study, in 2006, he was with Intel Corporation, Folsom, CA, researching the overlay network infrastructure over wireless mesh networks. He has authored several journal papers and conference papers and has recently been selected as one of the eight Ph.D. students for the 2007 Watson Emerging Leaders in Multimedia Workshop, which was organized by IBM Research. His research interests include cross-layer optimizations/adaptations for multimedia transmission over wireless mesh networks and dynamic resource allocation based on collaborative information exchange for delay-sensitive applications.

Mihaela van der Schaar (SM'04) received the Ph.D. degree from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2001.

She is currently an Associate Professor with the Department of Electrical Engineering, University of California, Los Angeles. Since 1999, she has been an active participant with the ISO MPEG standard, to which she has made more than 50 contributions and for which she has received three ISO recognition awards. She is a coeditor (with P. Chou) of the book *Multimedia over IP and Wireless Networks: Compression, Networking, and Systems* (Elsevier, Mar. 2007). She is the holder of 30 granted U.S. patents.

Dr. van der Schaar was the recipient of the 2004 National Science Foundation Career Award, the 2004–2006 Most Cited Paper Award from the *EURASIP Journal of Signal Processing: Image Communication*, the 2005 Best IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Paper Award, the 2005 and 2007 IBM Faculty Award, and the 2006 Okawa Foundation Award.