Learning-Based Spatial Reuse for WLANs With Early Identification of Interfering Transmitters

Bo Yin, *Student Member, IEEE*, Koji Yamamoto, *Member, IEEE*, Takayuki Nishio, *Member, IEEE*, Masahiro Morikura, *Member, IEEE*, and Hirantha Abeysekera, *Member, IEEE*

***Abstract*—In this paper, a reinforcement learning-based spatial reuse scheme for wireless local area networks (WLANs) is proposed and analyzed. In this scheme, when an access point (or a station) overhears an on-going transmission, it decodes the information in the frame header to identify the transmitter and decides whether or not to exploit spatial reuse accordingly. Specifically, it decides whether to stop receiving the remaining part of the frame and start its own transmission or to refrain from channel access until the detected transmission finishes. Through the repeated update Q-learning (RUQL) algorithm, the agent learns the optimal decision in the sense of reducing the media access control layer delay. Moreover, we compare the proposed scheme with the spatial reuse operation in IEEE 802.11ax, which makes the spatial reuse decision only based on a binary identification of the detected interferer, i.e., whether it is in my cell or neighboring cells. The proposed scheme, however, treats different interferers differently for exploiting spatial reuse. From a theoretical perspective, we derive a theoretical bound on the gains in the value function, i.e., the discounted sum of delay, due to making non-binary identifications. Simulation evaluations confirm that the proposed scheme achieves high throughput by reducing the time of freezing backoff counter while not increasing the time of failed transmissions.**

***Index Terms*—IEEE 802.11ax WLAN, spatial reuse, stochastic decision process, reinforcement learning, state partition.**

I. Introduction

THE SIGNIFICANT growth in the number of wireless local area networks (WLANs) devices in recent years [1]–[3] has resulted in the common occurrence of overlapping basic service sets (OBSSs), i.e., co-located WLANs cells operating in the same frequency channel. In dense OBSSs scenario, the throughput degradation becomes a severe problem because concurrent transmissions among OBSSs were not allowed in previous IEEE 802.11 standards, e.g., IEEE 802.11n, 11ac. In particular, an access point (AP) or a station (STA) has to defer its channel access when it detects the transmission of any other APs or STAs.

Manuscript received May 7, 2019; revised October 4, 2019; accepted November 19, 2019. Date of publication November 27, 2019; date of current version March 6, 2020. The associate editor coordinating the review of this article and approving it for publication was K. W. Sowerby. *(Corresponding author: Bo Yin.)*

B. Yin, K. Yamamoto, T. Nishio, and M. Morikura are with the Graduate School of Informatics, Kyoto University, Kyoto 6068501, Japan (e-mail: yin@imc.cce.i.kyoto-u.ac.jp; kyamamot@i.kyoto-u.ac.jp; nishio@i.kyoto-u.ac.jp; morikura@i.kyoto-u.ac.jp).

H. Abeysekera is with the NTT Access Network Service Systems Laboratories, NTT Corporation, Yokosuka 2390847, Japan (e-mail: hirantha.abeysekera@lab.ntt.co.jp).

Digital Object Identifier 10.1109/TCCN.2019.2956133

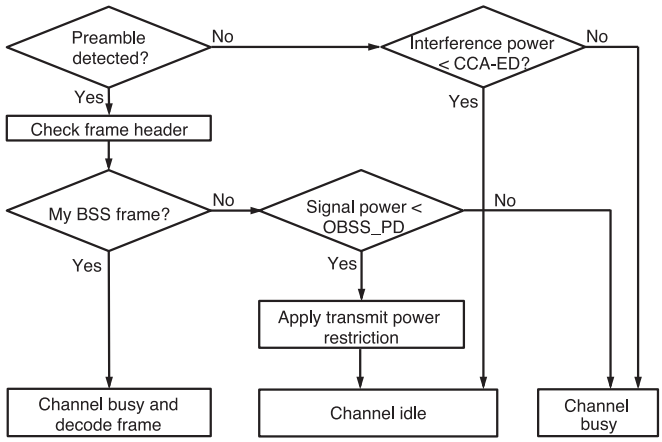

Fig. 1. The OBSS_PD-based spatial reuse operation approved in the IEEE 802.11ax. Interested readers may refer to [1]–[4] for further information.

The IEEE 802.11ax standard has approved a new operation called the *OBSS packet detect (OBSS_PD)-based spatial reuse operation* to improve spatial frequency reuse in high-density scenario [1]–[4]. This operation improves spatial reuse by allowing concurrent transmissions among OBSSs. As shown in Fig. 1, once an AP or an STA has detected an on-going transmission, it immediately identifies whether this transmission is in OBSS or not. This identification is done by checking the basic service set (BSS) color field in the frame header. If the detected transmission is in OBSS and its interference power is lower than a predefined threshold, i.e., OBSS_PD, the AP or STA stops receiving the remaining part of the frame and regards the wireless medium as idle, i.e., it is feasible to start transmission.

This OBSS_PD-based spatial reuse operation, however, has a major challenge. The challenge is that comparing the interference power with the predefined threshold OBSS_PD has limited predictive value in determining the success or failure of concurrent transmissions. In other words, a transmitter can not guarantee that its receiver receives its transmission successfully under the interference from OBSS, even if the interference is less than OBSS_PD. This results in performance degradation since if packet loss occurs, the concurrent transmission would be nonsense and should not be performed.

There are at least two reasons causing this unreliability of the OBSS_PD-based spatial reuse operation. First of all, this

TABLE I
PREVIOUS WORKS ON WLANS SPATIAL REUSE

Reference	Adaptive CCA?	Treat different interferers differently?	Learn from past experiences?
[7]–[13]	No	No	No
[14]–[18]	Yes	No	No
[19]	Yes	Yes	No
This paper	Yes	Yes	Yes

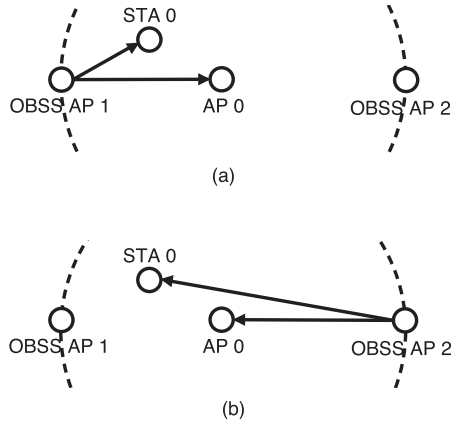

Fig. 2. The OBSS_PD-based spatial reuse operation only identifies whether the transmission is in my BSS or in OBSS. It treats interferers in different OBSSs indifferently. In this example, it is desirable if AP 0 can identify which interferer is transmitting before deciding whether or not to transmit concurrently with that interferer.

operation only identifies whether a detected transmission is in my BSS or in OBSS. It, however, treats interferers in different OBSSs indifferently. Consider the example in Fig. 2, where AP 0 tries to send a packet to STA 0. OBSS AP 1 is close to STA 0 and OBSS AP 2 is far from STA 0. In case (a), AP 0 detects the transmission of OBSS AP 1 whose interference power is I. In case (b), AP 0 detects the transmission of OBSS AP 2 whose interference power is also I. Although the measured interference at AP 0 is identical in cases (a) and (b), the receiver experiences different amounts of interference power. Concurrent transmission with OBSS AP 1 is more likely to fail. Hence, it is desirable if AP 0 can identify which interferer is transmitting before deciding whether or not to transmit concurrently with that interferer. Besides the first reason, another essential reason is that it is hard to establish an accurate and universal model of the one-to-one correspondence between interference power and packet loss probability in the actual wireless channel [5], [6].

To overcome this challenge, in this paper, we formulate the spatial reuse of WLANs as a stochastic decision process and propose a learning-based spatial reuse scheme. The proposed scheme has two distinctive features. First, the proposed scheme utilizes the information in the detected frame header to identify the interferer and makes decisions accordingly. This solves the problem in Fig. 2 in that the proposed scheme has the freedom of deciding whether or not to transmit concurrently with a particular interferer, rather than setting a common threshold to all OBSS interferers. Second, the proposed scheme learns from past experiences of success or failure in concurrent transmissions. The main merit of using learning-based scheme is that it does not assume any prior knowledge of the correspondence between the interference power and the packet loss probability. Besides, the proposed scheme does not need additional information report from the receiver other than acknowledgment (ACK).

Moreover, in this paper, we theoretically analyze the performance gains due to identifying interferers in exploiting spatial reuse. The key idea of our analysis is to partition the state space of the original decision process. Thereby, we use the partitioned decision process to model the spatial reuse operation where the agent does not identify interferers. By calculating the deviation between the partitioned decision process and the original decision process, we derive a theoretical bound on the gains in the value function due to identifying interferers.

The novelty and contributions of this work are as follows:

- A learning-based spatial reuse scheme which utilizes the information in the detected frame header to identify the interferer and makes decisions accordingly.
- We use a stochastic decision process and its partition to model the spatial reuse operations where the transmitter does and does not identify interferers, respectively. We further calculate their deviation in Markov environment.
- By using this deviation, we derive a theoretical bound on the gains in value function due to identifying interferers.

The remainder of this paper is organized as follows. Related works are summarized in Section II. Stochastic decision process formulation is presented in Section IV. The learning algorithm is presented in Section V. Section VI theoretically analyzes the gains in the value function due to identifying interferers in exploiting spatial reuse. Evaluation results are presented in Section VII. Section VIII concludes the paper.

II. RELATED WORK

Several previous works have been performed to improve spatial reuse in dense WLANs. A brief comparison between related works and this paper is presented in Table I.

A category of previous works [7]–[10] discusses the optimal setting of the clear channel assessment (CCA) threshold, i.e., OBSS_PD. Recent works [11]–[13] that use stochastic geometry approach to study the optimal CCA threshold also belong to this category. This category of works mainly aims at providing theoretical insights. They develop theoretical models of achievable throughput and study the optimal trade-off between spatial reuse and interference mitigation. The optimal CCA threshold is usually derived by considering homogeneous network density or regular topologies.

Another category of works proposes to dynamically adapt CCA threshold to network conditions. Two survey papers of adaptive CCA methods can be found in [3], [18]. An adaptive CCA method in which an AP adjusts CCA threshold according to the packet error rate (PER) is proposed in [20]. Dynamic sensitivity control (DSC) that adjusts CCA threshold based on the communication distance also belongs to this category. Evaluations of DSC-like methods are presented in [14]–[17]. All these schemes, however, treat interferers in different OBSS interferers indifferently. As shown in Fig. 2, more desirable operations can be achieved by identifying which interferer is transmitting.

Reference [19] is similar to this paper in that it proposes to adapt the CCA threshold to individual OBSS interferer. The scheme in [19], however, is model-based and requires the prior knowledge of the correspondence between interference power and packet loss probability. For example, it assumes that STAs are aware of the required signal-to-interference ratio (SIR) for successful receptions. In practice; however, it is often hard to establish an accurate and universal model of such correspondence [5], [6]. In contrast, our proposed scheme is learning-based and does not require such prior knowledge. Moreover, the scheme in [19] requires periodically information reports between the transmitter and the receiver, which is not supported in the IEEE 802.11ax standard. In contrast, our proposed scheme only needs modifications on the transmitter side whereas the receiver does not require any modifications beyond the IEEE 802.11ax standard.

III. Preliminaries: Early Identification of Interfering Transmitters

Before starting to describe the proposed scheme, in this section, we explain how to utilize the information in the detected frame header to identify the interferer [4].

In IEEE 802.11ax, the physical (PHY) preamble contains a mandatory high efficiency signal-A (HE-SIG-A) field. It lasts for 16 μs and provides some basic information about the frame. One information that we are interested in is the BSS color. It is a 6-bit numerical identifier of BSSs. An AP and all its associated STAs share the same BSS color, while co-located co-channel APs use different BSS colors. Besides the BSS color bits, the HE-SIG-A field also contains information showing whether the frame is sent in downlink or uplink.

Note that, when an AP or an STA overhears a frame, it always first decodes the PHY layer preamble. By decoding the BSS color bits in the preamble, it can make an early identification of which BSS the frame belongs to. Furthermore, if the detected frame is sent in downlink, we can uniquely identify the transmitting AP, since there is only one AP in each BSS. Note that, this identification is done before we start to receive the remaining part of the frame.

IV. A Stochastic Decision Process Formulation of Spatial Reuse

In this section, we formulate the spatial reuse in WLANs as a stochastic decision process. We focus on the spatial reuse from a single-agent perspective. In particular, we consider an AP that is sending packets to an associated STA. We treat all other interferers as a part of the environment. Note that, although we consider the AP as the transmitter and the STA as the receiver, it works in the same manner if the transmitter is an STA. Hereafter, we refer to the AP under consideration as the agent.

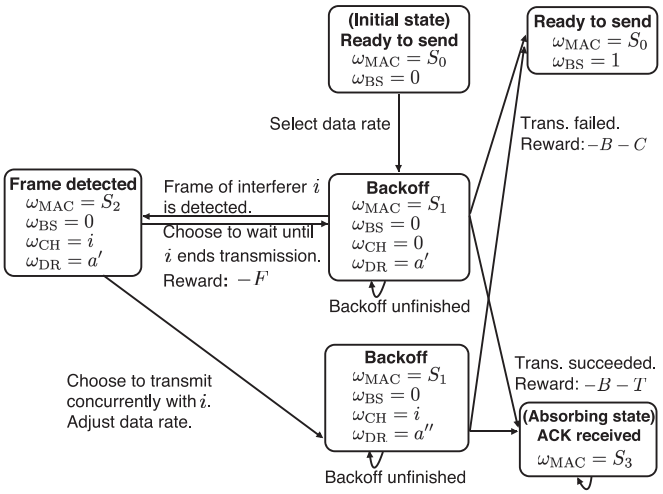

Fig. 3. State transition diagram. For illustrative simplicity, we mainly show the transitions of the MAC state ω_{MAC}.

A stochastic decision process is defined as a four-tuple $(\Omega, \mathcal{A}, q, R)$, where Ω denotes the set of states and $\mathcal{A}(\omega[t])$ denotes the set of possible actions under state $\omega[t] \in \Omega$. When the agent selects an action $a[t] \in \mathcal{A}(\omega[t])$ at instant t, the state transits to $\omega[t+1] \in \Omega$ according to a probability distribution given by q_t. The agent receives a stochastic reward $R(\omega[t], \omega[t+1], a[t])$ at the same time.

The Markov decision process (MDP) is a special type of stochastic decision process where the state transitions satisfy the Markov property. Let $q(\omega[t], \omega[t+1], a[t])$ denote the probability that the state transits to $\omega[t+1]$ when the agent selects an action $a[t] \in \mathcal{A}(\omega[t])$ at instant t.

Note that, in practice, the environment is not a stationary MDP due to many factors, e.g., topology changes, STAs arrivals or departures, the existence of multiple agents. In these cases, the probability distribution q_t is time-varying. In this work, we only assume stationary MDP for the theoretical analysis presented in Section VI. The simulation evaluations in Section VII, however, evaluate the proposed scheme in various scenarios including time-varying topology and multiple agents scenarios.

A. State

As shown in Fig. 3, the state space Ω is defined from the union and the Cartesian product of the media access control (MAC) state space Ω_{MAC}, the backoff stage state space Ω_{BS}, the channel state space Ω_{CH}, and the data rate state space Ω_{DR}.

The MAC state space $\omega_{\mathrm{MAC}}[t] \in \Omega_{\mathrm{MAC}}$ has four possible values, i.e., $\Omega_{\mathrm{MAC}} \coloneqq \{S_0, S_1, S_2, S_3\}$. State S_0 denotes that the agent is ready to contend for channel access. State S_1

denotes the state of backoff, where the agent keeps on sensing the channel while reducing the backoff counter. State S_2 denotes the state that the agent has detected the preamble of a transmission. State S_3 is an absorbing state, which denotes that the packet has been successfully received.

The backoff stage state $\omega_{\mathrm{BS}}[t] \in \Omega_{\mathrm{BS}}$ denotes the current backoff stage, i.e., the times of consecutive transmission failures at present. In IEEE 802.11 WLANs [21], the contention window size doubles after a failed transmission and is reset after a successful transmission. We consider $\Omega_{\mathrm{BS}} \coloneqq \mathcal{J} \coloneqq \{0, 1, \ldots, J_{\max}\}$, i.e., the contention window size does not grow further after $J_{\max}$ times of consecutive transmission failures [21].

The channel state $\omega_{\mathrm{CH}}[t] \in \Omega_{\mathrm{CH}}$ denotes the index of the transmitting interferer that has been identified by the agent. Note that, the agent keeps on sensing the channel during the backoff period. If the agent detects a transmission, we assume that the agent immediately identifies the interferer through checking the information in the detected frame header. Note that, if multiple interferers are transmitting preambles at the same time, there are two possibilities. Either the preamble with the strongest received power is decoded or any preamble is unable to be successfully decoded, i.e., preamble error. This depends on whether the SINR requirement for decoding the preamble is met. Note that, the preamble is modulated using the lowest modulation and coding scheme (MCS), i.e., BPSK. [21]

The channel state space is defined as $\Omega_{\mathrm{CH}} \coloneqq \{0\} \cup \mathcal{N} \coloneqq \{0, 1, 2, \ldots, N\}$, where $\omega_{\mathrm{CH}}[t] = 0$ denotes that the channel is idle or the interferer is unable to be identified.

The data rate state $\omega_{\mathrm{DR}}[t] \in \Omega_{\mathrm{DR}}$ denotes the currently chosen data rate for transmission. We consider $\Omega_{\mathrm{DR}} = \mathcal{K} \coloneqq \{1, \ldots, K\}$, where *K* denotes the number of available MCS.

As shown in Fig. 3, the backoff stage state ω_{BS} is defined if the MAC state is S_0, S_1, or S_2. The channel state ω_{CH} and data rate state ω_{DR} are defined if the MAC state is S_1 or S_2. In summary, the entire state space Ω is defined from the union and the Cartesian product of four spaces as follows:

$$\Omega \coloneqq (\{S_0\} \times \Omega_{\mathrm{BS}}) \cup (\{S_1, S_2\} \times \Omega_{\mathrm{BS}} \times \Omega_{\mathrm{CH}} \times \Omega_{\mathrm{DR}}) \cup \{S_3\}. \quad (1)$$

B. Description of State Transitions

The stochastic decision process denotes the entire process of transmitting a packet as illustrated in Fig. 3. When a new packet arrives at the head of the transmission queue, the state is initialized as $\omega_{\mathrm{MAC}}[0] = S_0$, $\omega_{\mathrm{BS}}[0] = 0$. When the packet is successfully received, the state transits to S_3. State transition happens when following events occur: the agent begins contending for the channel access, the agent detects an ongoing transmission, the agent makes a CCA decision, or the agent transmits.

- When the MAC state is S_0, i.e., when the agent is ready to contend for channel access, the agent selects a data rate from $\mathcal{A}(\omega_{\mathrm{MAC}} = S_0)$ for transmission. Then, the MAC state transits to S_1, the channel state transits to 0, and the data rate state transits to the chosen data rate.
- When the MAC state is S_1, i.e., during the backoff period, the agent keeps on carrier sensing while reducing the backoff counter. If the agent detects the preamble of a transmission, the MAC state transits to S_2 and the channel state transits to the index of the detected interferer.
- When the MAC state is S_2, the agent decides whether to wait until the detected transmission ends or to transmit concurrently with the identified interferer *i*. If the agent chooses to wait, the agent freezes the backoff counter until the detected transmission ends. After that, the MAC state transits from S_2 to S_1. The channel state ω_{CH} is reset to 0. On the other hand, if the agent chooses to transmit concurrently with interferer *i*, the MAC state transits from S_2 to S_1, the channel state ω_{CH} stays to be the index of the identified interferer. In this case, the agent also adjusts the data rate for concurrent transmission. The data rate state transits to the re-selected data rate.
- The agent starts transmission when the backoff counter is reduced to 0. According to the transmission results, the MAC state transits to either S_0 or S_3. If the transmission has failed, the MAC state transits to S_0 and the backoff stage $\omega_{\mathrm{BS}}[t]$ transits to $\min(\omega_{\mathrm{BS}}[t] + 1, J_{\max})$. If the packet is successfully transmitted, i.e., the agent has received the ACK from the receiver, the MAC state transits to S_3.

The MAC state S_3 is an absorbing state with zero reward. It indicates that the packet has been received successfully. The learning episode ends when the MAC state reaches S_3 and there is no further state transition. When a new packet arrives at the head of the transmission queue, another learning episode begins and the state is initialized as $\omega_{\mathrm{MAC}}[0] = S_0$, $\omega_{\mathrm{BS}}[0] = 0$.

C. Action

The set of available actions depends on the MAC state. First of all, when the MAC state is S_0, i.e., when the agent is ready to contend for channel access, the agent needs to select a data rate for transmission, i.e., $\mathcal{A}(\omega_{\mathrm{MAC}} = S_0) = \mathcal{K}$.

Second, when the MAC state is S_2, i.e., when a transmission is detected, the agent chooses whether or not to ignore the detected transmission. If the agent chooses to ignore the detected transmission, i.e., to transmit concurrently with interferer *i*, the agent also adjusts the data rate for concurrent transmission. Hence, $\mathcal{A}(\omega_{\mathrm{MAC}} = S_2) = \{0\} \cup \mathcal{K}$, where $a = 0$ means to wait until the interferer ends its transmission.

Finally, when the MAC state is S_1, i.e., when the backoff counter has not been reduced to zero, the only available action is to continue carrier sensing. We denote it as $\mathcal{A}(\omega_{\mathrm{MAC}} = S_1) = \{0\}$.

D. Metric and Reward

The metric that we use to evaluate a spatial reuse operation is the MAC layer service time [22] of a packet. Formally, the MAC service time is defined as the duration from the instant when a packet arrives at the head of the transmission queue and the agent begins contending for the channel, to the instant

when the agent has received an ACK from the receiver [22]. It comprises four parts. Given that the agent has successfully transmitted a packet after *J* times of consecutive packet transmission failures, where $J \in \mathbb{N}_{\geq 0}$, the MAC service time of a packet includes:

- the duration of *J* times of failed transmissions,
- the duration of the successful transmission,
- the backoff duration before each transmission attempt,
- the duration that the agent freezes its backoff counter.

The MAC service time of a packet is formulated as follows [22]:

$$D = \underbrace{\sum_{j=0}^{J-1} C_j}_{\substack{J \text{ times of failed} \\ \text{transmissions}}} + \underbrace{T_J}_{\substack{\text{Successful} \\ \text{transmission}}} + \underbrace{\sum_{j=0}^{J} B_j}_{\substack{\text{Backoff} \\ \text{countdown}}} + \underbrace{\sum_{i=1}^{Y} F_i}_{\text{Freeze}}, \quad (2)$$

where C_j, T_J, B_j, *Y*, and F_i are stochastic values. Here, C_j denotes the duration of the unsuccessful transmission in backoff stage *j*, which includes the duration of transmitting the header, data, ACK timeout, and DIFS [21]. T_J denotes the duration of the successful transmission, which includes the duration of transmitting the header, data, ACK, SIFS, and DIFS [21]. B_j denotes the backoff countdown duration in backoff stage *j*. *Y* denotes the number of times that the agent has freezed its backoff counter. F_i denotes the duration that the agent freezes its backoff counter.

Note that, the MAC service time *D* of a packet is exactly the duration from the initial state to the absorbing state as shown in Fig. 3. Since we are interested in reducing the MAC service time, in this paper, we consider the reward as the negative value of the MAC service time.

The agent measures the duration of each event and calculates the corresponding reward when state transition happens. In particular, the agent receives a reward $-B_j - C_j$ when the transmission failed, i.e., when the MAC state transited from S_1 to S_0. The agent receives a reward $-B_J - T_J$ when the transmission succeeded, i.e., when the MAC state transited from S_1 to S_3. The agent receives a reward $-F_i$ when it has frozen the backoff counter to wait until the detected transmission ends, i.e., when $a = 0$ and the MAC state transits from S_2 to S_1.

V. Learning-based Spatial Reuse Operation

The ultimate goal of the learning algorithm presented in this section is to find the policy of making spatial reuse decisions such that the maximum discounted sum of rewards can be achieved. Intuitively speaking, we hope the agent can learn to transmit concurrently with those OBSS interferers whose interference is tolerable at the receiver. On the other hand, we hope the agent can learn to refrain from transmitting concurrently with those interferers whose interference is not tolerable at the receiver. Since we have already formulated the problem as a stochastic decision process, we can apply reinforcement learning algorithms to solve this problem.

A. Learning Algorithm

A policy π is a solution concept of the stochastic decision process. It is a mapping from the state space Ω to the action space $\mathcal{A}(\omega[t])$. Given a state ω_0 and a policy π, let $V^\pi(\omega_0)$ denote the expectation of the discounted sum of reward the agent would receive within one episode, i.e.,

$$V^\pi(\omega_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R[t] \Big| \omega[0] = \omega_0\right], \quad (3)$$

where $\gamma \in [0, 1)$ is the discounted factor. Note that, in the considered stochastic decision process, the physical meaning of $V^\pi(\omega_0)$ is the negative discounted value of the MAC service time, which we hope to maximize. The discounted factor γ indicates how important future rewards are to the current state. Generally, a large γ yields a better outcome after convergence, but it requires longer time to converge [23]. The γ is often set to a value close to one in related studies, e.g., [23]. In this paper, we consider $\gamma = 0.99$.

It is known that the Q-learning (QL) algorithm derives the optimal policy in a stationary MDP environment [24]. The proposed scheme, however, does not directly apply the conventional QL as the learning algorithm. The reason is that the real environment is not always stationary due to many factors, e.g., changes in communication distance, new interferers arrivals, the existence of multiple agents.

To tackle the non-stationarity, we make modifications to the QL algorithm. First, we incorporate the technique of repeated update Q-learning (RUQL) [25] to solve the policy bias problem of the conventional QL algorithm. As pointed in [25], the policy bias problem causes performance degradation in noisy non-stationary environments. The policy bias problem refers to the problem that, those optimal actions with temporal lower values are executed less often during the learning process in the conventional QL algorithm. As a result, the values of those actions are updated less often. This leads to performance degradation since the environment may have already changed before the agent learns the optimal action.

The basic idea of RUQL is to adjust the learning rate in the conventional QL algorithm so that less-chosen actions have a higher learning rate. Let $Q(\omega, a)$ denote the state-action value corresponding to state ω and action *a*. Given the current state $\omega[t]$, the selected action *a*[*t*], the new state $\omega[t+1]$, and the associated reward *R*[*t*], the RUQL updates the state-action value $Q(\omega, a)$ according to the following expression [25], i.e.,

$$\begin{aligned} Q(\omega[t], a[t]) \leftarrow\ & (1 - z_n) Q(\omega[t], a[t]) \\ & + z_n \left[R[t] + \gamma \max_{a'} Q\big(\omega[t+1], a'\big) \right], \end{aligned} \quad (4)$$

where γ is the discounted factor and z_n is the learning rate at episode *n*. The learning rate z_n is given as follows [25]:

$$z_n = 1 - [1 - \alpha_n]^{\frac{1}{\pi_n(\omega[t], a[t])}}, \quad (5)$$

where α_n denotes the learning rate in the conventional QL algorithm and $\pi_n(\omega[t], a[t])$ denotes the probability of choosing action *a*[*t*] at state $\omega[t]$ at episode *n*.

In this paper, we consider the ε-greedy exploration policy, i.e.,

$$\pi_n(\omega, a) = \begin{cases} 1 - \varepsilon, & a = \arg\max_{a' \in \mathcal{A}(\omega)} Q\big(\omega, a'\big); \\ \varepsilon, & a \neq \arg\max_{a' \in \mathcal{A}(\omega)} Q\big(\omega, a'\big), \end{cases} \quad (6)$$

where ε is a small constant that denotes the exploration rate. Note that, ε controls the trade-off between exploration and exploitation. Given a higher ε, the agent explores the action space more aggressively. On the other hand, it cannot guarantee acceptable run-time performance since non-greedy actions will be taken frequently [23]. In this paper, we consider $\varepsilon = 0.1$.

Moreover, in non-stationary environment, it is also important that the agent can respond quickly to environment changes. The learning rate $0 < \alpha < 1$ plays an important role in determining the learning speed of the agent. In fact, there is a trade-off between the stability and the speed of the learning algorithm [25], [26]. If α is small, the agent cannot respond quickly to the environment changes. If α is large, the algorithm may not be robust and stable under stationary environment.

In this paper, we consider that the learning rate α satisfies the following two conditions: $\sum_0^\infty \alpha_n = \infty$ and $\sum_0^\infty \alpha_n^2 = 0$. Specifically, we consider $\alpha_n = 1000/(1000+n)$ in simulation evaluations, where *n* is the learning episode. It is worth mentioning that, if this condition is satisfied, the RUQL guarantees stability, i.e., convergence, in stationary MDP environment [25]. On the other hand, we also let the agent reset the learning episode $n = 1$, if *n* is large and the agent detects any arrivals of new interferers or significant changes in the topology. This can be detected by periodically measuring the received signal strength indicator (RSSI) of neighbors [27].

B. Transmit Power Restriction

We consider that the transmit power of the agent is restricted when the agent starts a concurrent transmission. The reason of restricting transmit power is to protect the on-going transmission in OBSS from being corrupted by the newly issued concurrent transmission of the agent. Note that, the transmit power restriction is also considered in the OBSS_PD-based spatial reuse operation as shown in Fig. 1 [1]–[3].

When the agent decides to transmit concurrently with an on-going transmission, we consider the transmit power of the concurrent transmission is given as follows:

$$p = \min\left(P_{\text{ref}}, \frac{P_{\text{ref}}\Theta_{\min}}{I}\right), \tag{7}$$

where P_{ref} denotes the maximum possible transmit power of the agent, $\Theta_{\min} = -82\,\text{dBm}$ [21] denotes the default CCA threshold of legacy devices, and *I* denotes the measured interference strength. The intuition of this rule is to adjust the transmit power inversely proportional to the detected interference strength. Note that if the agent does not detect any on-going transmissions, it transmits with its maximum possible transmit power.

VI. Analysis of Gains Due to Identifying Interferers

As shown in Fig. 2, it is desirable if the agent can identify which interferer is transmitting before deciding whether or not to transmit concurrently with that interferer. In this section, we introduce the concept of state aggregation and use this concept to analyze the gains due to identifying interferers.

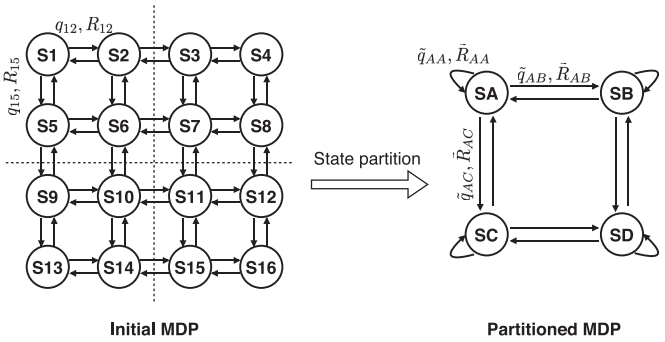

Fig. 4. The concept of state partition (or called state aggregation).

A. Preliminary: State Partition

We first introduce the concept of state partition (or called state aggregation). The state partition is originated as an approximate method to solve the MDP whose state space is large [28], [29]. The intuition of the state partition is to aggregate multiple similar states into meta-states to create smaller state space. An illustration of the state partition is given in Fig. 4, where 16 states are aggregated into 4 states. Thereby, the optimal policy derived on the partitioned MDP can be considered as an approximate solution of the initial MDP. Since the state space is smaller, the derivation is generally easier.

The definition of state partition is given as follows:

Definition 1 [29]: An MDP $\tilde{\Psi} = (\tilde{\Omega}, \mathcal{A}, \tilde{q}, \tilde{R})$ is an δ-homogenous partition of MDP $\Psi = (\Omega, \mathcal{A}, q, R)$ if there exists a mapping $\phi : \Omega \to \tilde{\Omega}$, such that ϕ is surjective and for all $\omega \in \Omega$ and $a \in \mathcal{A}$ the following conditions hold:

$$\left|\sum_{\omega'' \in \tilde{\Omega}} \left| \sum_{\omega' : \phi(\omega') = \omega''} q(\omega, \omega', a) - \tilde{q}(\phi(\omega), \omega'', a) \right| \right| \leq \delta, \tag{8}$$

$$\max_{\omega' : \phi(\omega') = \omega} \frac{\left| R(\omega', a) - \tilde{R}(\omega, a) \right|}{|R|_{\max}} \leq \delta, \tag{9}$$

where $|R|_{\max}$ denotes the maximum achievable absolute value of reward.

The metric $\delta \in [0, 1]$ describes how much the partitioned MDP $\tilde{\Psi}$ deviates from the original Ψ. In this paper, we refer to δ as the *deviation* between two MDPs. The constant $|R|_{\max}$ is a normalization factor such that $\frac{|R|}{|R|_{\max}} \in [0, 1]$.

Let $\tilde{\pi}^*$ denote the optimal policy derived on the partitioned MDP $\tilde{\Psi}$. It induces a policy on the original MDP Ψ as follows:

$$\hat{\pi}^*(\omega) \coloneqq \tilde{\pi}^*(\phi(\omega)). \tag{10}$$

The induced policy $\hat{\pi}^*$ is an approximate solution of the real optimal policy π^* on the original MDP Ψ. It has been shown in [29] that the value functions of $\hat{\pi}^*$ and π^* satisfies the following property.

Theorem 1 [29]: Let $\tilde{\Psi}$ be a δ-homogeneous partition of MDP Ψ, then the optimal policy in $\tilde{\Psi}$ induces an $\frac{2\delta|V|_{\max}}{1-\gamma}$-optimal policy in Ψ, i.e., $\forall \omega \in \Omega$

$$\left| V^{\hat{\pi}^*}(\omega) - V^{\pi^*}(\omega) \right| \leq \frac{2\delta|V|_{\max}}{1-\gamma}, \tag{11}$$

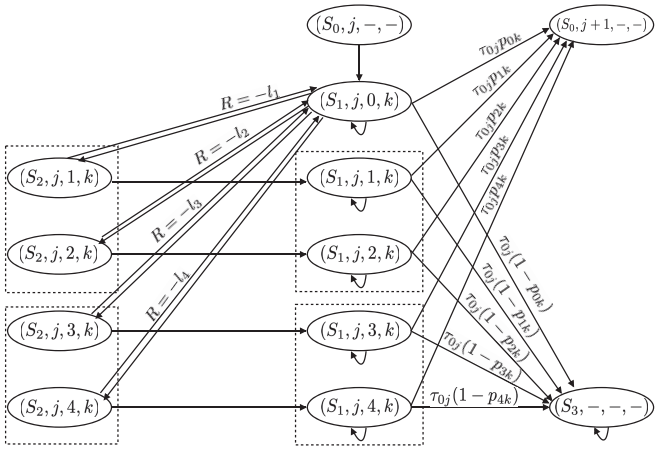

Fig. 5. The MDP Ψ that represents the spatial reuse operation where agent identifies interferers and treats different OBSS interferers differently. The channel state space is $\Omega_{\mathrm{CH}} = \{0, 1, \ldots, N\}$. In this illustration, there are 4 OBSS interferers, i.e., $N = 4$. For notational simplicity, we use a vector $(\omega_{\mathrm{MAC}}, \omega_{\mathrm{BS}}, \omega_{\mathrm{CH}}, \omega_{\mathrm{DR}})$ to represent the state. We omit some states for illustrative simplicity.

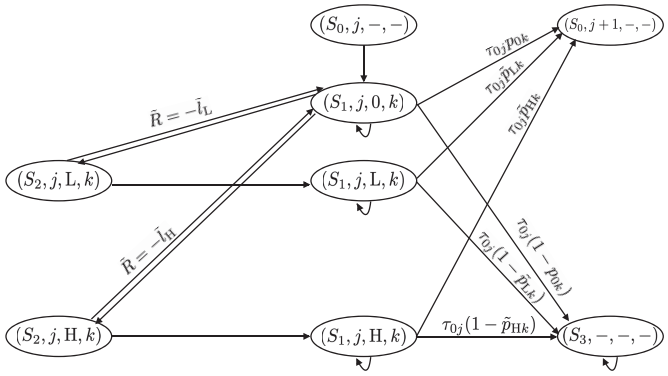

Fig. 6. This partitioned MDP $\tilde{\Psi}$ represents the spatial reuse operation where agent does not identify interferers and only distinguishes whether the interference power of the detected OBSS interferer is above or below the OBSS_PD. The original channel state $\Omega_{\mathrm{CH}} = \{0, 1, \ldots, N\}$ is partitioned into $\tilde{\Omega}_{\mathrm{CH}} = \{0, \mathrm{L}, \mathrm{H}\}$. The states inside the same dashed boxes in Fig. 5 are partitioned together.

where $|V|_{\max}$ is the maximal achievable absolute value of the value function.

B. Analysis of Gains Due to Identifying Interferers

The main idea of our analysis is that we use an MDP $\Psi = (\Omega, \mathcal{A}, q, R)$ to represent the decision process where the agent identifies interferers and treats different OBSS interferers differently. On the other hand, we use its partition $\tilde{\Psi} = (\tilde{\Omega}, \mathcal{A}, \tilde{q}, \tilde{R})$ to represent the decision process where the agent does not identify interferers and only distinguishes whether the interference power of the detected OBSS interferer is above or below the OBSS_PD. Note that, we assume the environment is stationary MDP only in this section. We do not assume the environment to be stationary MDP in the learning algorithm and in the simulation evaluations. The illustrations of Ψ and $\tilde{\Psi}$ are shown in Figs. 5 and 6. For notational simplicity, we use a vector $(\omega_{\mathrm{MAC}}, \omega_{\mathrm{BS}}, \omega_{\mathrm{CH}}, \omega_{\mathrm{DR}})$ to represent the state. We omit some states for illustrative simplicity.

As shown in Figs. 5 and 6, the original channel state $\Omega_{\mathrm{CH}} = \{0, 1, \ldots, N\}$ is partitioned into $\tilde{\Omega}_{\mathrm{CH}} = \{0, \mathrm{L}, \mathrm{H}\}$, where $\tilde{\omega} = 0$ denotes that the channel is idle or the preamble of the detected frame is unable to be decoded, $\tilde{\omega} = \mathrm{H}$ denotes that the agent has detected the transmission of an OBSS interferer whose interference power is above the OBSS_PD, and $\tilde{\omega} = \mathrm{L}$ denotes that the interference is below the OBSS_PD. Hence, the partitioned state space $\tilde{\Omega}$ is defined as follows:

$$\begin{aligned}\tilde{\Omega} \coloneqq\ & (\{S_0\} \times \Omega_{\mathrm{BS}}) \\ & \cup \left(\{S_1, S_2\} \times \Omega_{\mathrm{BS}} \times \tilde{\Omega}_{\mathrm{CH}} \times \Omega_{\mathrm{DR}}\right) \cup \{S_3\}. \quad (12)\end{aligned}$$

Let $\mathcal{N}_{\mathrm{L}}$ and $\mathcal{N}_{\mathrm{H}}$ denote the set of interferers whose average interference power to the agent is below or above the OBSS_PD, respectively. Note that, $\mathcal{N}_{\mathrm{L}}$ and $\mathcal{N}_{\mathrm{H}}$ are two disjoint subsets of $\mathcal{N}$, where $\mathcal{N} = \mathcal{N}_{\mathrm{L}} \cup \mathcal{N}_{\mathrm{H}}$ and $\mathcal{N}_{\mathrm{L}} \cap \mathcal{N}_{\mathrm{H}} = \emptyset$. Hence, the mapping $\phi_{\theta,\mathrm{CH}} : \Omega_{\mathrm{CH}} \to \tilde{\Omega}_{\mathrm{CH}}$ that maps the original channel state space to the partitioned channel state space is given as follows:

$$\phi_{\theta,\mathrm{CH}}(\omega_{\mathrm{CH}}) = \begin{cases} 0, & \omega_{\mathrm{CH}} = 0; \\ \mathrm{L}, & \omega_{\mathrm{CH}} \in \mathcal{N}_L; \\ \mathrm{H}, & \omega_{\mathrm{CH}} \in \mathcal{N}_H, \end{cases} \quad (13)$$

where θ denotes the OBSS_PD. Therefore, the mapping ϕ_θ:$\Omega \to \tilde{\Omega}$ that maps the original state space to the partitioned state space is given as follows:

$$\begin{aligned}&\phi_\theta(\omega) \\ &= \begin{cases} (\omega_{\mathrm{MAC}}, \omega_{\mathrm{BS}}, \phi_{\theta,\mathrm{CH}}(\omega_{\mathrm{CH}}), \omega_{\mathrm{DR}}), & \omega_{\mathrm{MAC}} \in \{S_1, S_2\}; \\ \omega, & \omega_{\mathrm{MAC}} \in \{S_0, S_3\}. \end{cases} \end{aligned} \quad (14)$$

Next, we introduce a series of parameters and analyze how much the partitioned MDP $\tilde{\Psi}$ deviates from the original MDP Ψ, i.e., the deviation δ.

Let $\tau_{0j} \in [0, 1]$ denote the probability that the backoff counter has been reduced to zero and the agent starts transmission in an idle slot time in backoff stage j. Given the current contention windows size CW_j, this probability can be approximately calculated as follows [30]:

$$\tau_{0j} = \frac{2}{1 + \mathrm{CW}_j}. \quad (15)$$

Let $p_{ik} \in [0, 1]$ denote the expected packet error probability when the agent transmits concurrently with interferer $i \in \mathcal{N}$ using data rate $k \in \mathcal{K}$. Specifically, let $p_{0k} \in [0, 1]$ denote the expected packet error probability when the agent transmits with data rate $k \in \mathcal{K}$ when channel is idle or the preamble of the detected frame is unable to be decoded. Hence, the probability that the agent fails a transmission at backoff stage j is given as follows:

$$q((S_1, j, i, k), (S_0, \min(j + 1, J_{\max})), 0) = \tau_{0j} p_{ik}, \quad (16)$$

where $i \in \{0, 1, \ldots, N\}$, $j \in \mathcal{J}$, and $k \in \mathcal{K}$. On the contrary, the probability that the agent succeeds in transmitting the packet and reaches the absorbing state is given as follows:

$$q((S_1, j, i, k), S_3, 0) = \tau_{0j}(1 - p_{ik}), \quad (17)$$

where $i \in \{0, 1, \ldots, N\}$, $j \in \mathcal{J}$, and $k \in \mathcal{K}$. Let l_i denote the expected duration of the transmission of interferer $i \in \mathcal{N}$. Hence, the reward that the agent receives when it chooses to wait until the transmission of interferer *i* ends is given as follows:

$$R((S_2, j, i, k), (S_1, j, 0, k), 0) = -l_i, \tag{18}$$

where $i \in \mathcal{N}$, $j \in \mathcal{J}$, and $k \in \mathcal{K}$.

On the other hand, in the partitioned MDP, the probability that the agent fails or succeeds in a transmission are given as follows:

$$\tilde{q}_\theta((S_1, j, i, k), (S_0, \min(j+1, J_{\max})), 0) = \tau_{0j}\tilde{p}_{ik}, \tag{19}$$

$$\tilde{q}_\theta((S_1, j, i, k), S_3, 0) = \tau_{0j}(1 - \tilde{p}_{ik}), \tag{20}$$

respectively, where $i \in \{0, \mathrm{L}, \mathrm{H}\}$, $j \in \mathcal{J}$, and $k \in \mathcal{K}$. Here, $\tilde{p}_{0k} = p_{0k}$ while $\tilde{p}_{\mathrm{L}k}$ and $\tilde{p}_{\mathrm{H}k}$ denote the expected packet error probability when the agent transmits concurrently with an interferer whose interference power is below or above the OBSS_PD, respectively. The reward that the agent receives when it chooses to wait until the detected transmission ends is given as follows:

$$\tilde{R}_\theta((S_2, j, i, k), (S_1, j, 0, k), 0) = -\tilde{l}_i, \tag{21}$$

where $i \in \{\mathrm{L}, \mathrm{H}\}$. Here, $\tilde{l}_\mathrm{L}$ and $\tilde{l}_\mathrm{H}$ denote the expected durations of the transmission of an interferer whose interference power is below or above the OBSS_PD, respectively.

The following theorem derives the deviation of the considered two MDPs.

Theorem 2: $\tilde{\Psi}$ is a δ_θ-homogenous partition of Ψ, where δ_θ is given as follows:

$$\delta_\theta = \max\left(\max_{(i,k)}(2\tau_{00}|p_{ik} - \tilde{p}_{i'k}|), \max_i\left(\frac{\left|l_l - \tilde{l}_{i'}\right|}{|R|_{\max}}\right)\right), \tag{22}$$

where $i \in \mathcal{N}$, $i' = \phi_{\theta,\mathrm{CH}}(i)$, $k \in \mathcal{K}$, and $|R|_{\max}$ denotes the maximum achievable absolute value of reward.

Proof: Let us calculate δ_θ based on (8) and (9) for every $\omega \in \Omega$.

First of all, when $\omega_{\mathrm{MAC}} = S_0$, the agent selects a data rate and ω_{MAC} transits from S_0 to S_1. This state transition is deterministic and does not depend on the channel state. The reward associated with this state transition is zero. Hence, there is no deviation between Ψ and $\tilde{\Psi}$ when $\omega_{\mathrm{MAC}} = S_0$.

Secondly, when $\omega_{\mathrm{MAC}} = S_1$ and the agent fails a transmission, ω_{MAC} transits from S_1 to S_0. The reward associated with the state transition does not depends on the channel state. The state transition probability, however, has a deviation between Ψ and $\tilde{\Psi}$ as follows:

$$\begin{aligned}&|q((S_1, j, i, k), (S_0, \min(j+1, J_{\max})), 0)\\ &\quad - \tilde{q}_\theta\big((S_1, j, i', k), (S_0, \min(j+1, J_{\max})), 0\big)|\\ &= \tau_{0j}|p_{ik} - \bar{p}_{i'k}|,\end{aligned}$$

where $i \in \{0, 1, \ldots, N\}$, $i' = \phi_{\theta,\mathrm{CH}}(i)$, $j \in \mathcal{J}$, and $k \in \mathcal{K}$. Similarly, when the agent succeeds in a transmission, the deviation between Ψ and $\tilde{\Psi}$ is given as follows:

$$\begin{aligned}&\big|q((S_1, j, i, k), S_3, 0) - \tilde{q}_\theta\big((S_1, j, i', k), S_3, 0\big)\big|\\ &\quad = \tau_{0j}|(1 - p_{ik}) - (1 - \tilde{p}_{i'k})| = \tau_{0j}|p_{ik} - \tilde{p}_{i'k}|,\end{aligned}$$

where $i \in \{0, 1, \ldots, N\}$, $i' = \phi_{\theta,\mathrm{CH}}(i)$, $j \in \mathcal{J}$, and $k \in \mathcal{K}$. When $\omega_{\mathrm{MAC}} = S_1$ and the agent detects a transmission of interferer *i*, ω_{MAC} transits from S_1 to S_2. The reward associated with this state transition is zero. The deviation of the state transition probability between Ψ and $\tilde{\Psi}$ is calculated as follows:

$$\begin{aligned}&\left|\sum_{i:\phi_{\mathrm{CH}}(i)=i'} q((S_1, j, 0, k), (S_2, j, i, k), 0)\right.\\ &\qquad \left. - \tilde{q}_\theta\big((S_1, j, 0, k), \big(S_2, j, i', k\big), 0\big)\right|,\end{aligned}$$

where $i \in \mathcal{N}$, $i' = \phi_{\theta,\mathrm{CH}}(i)$, $j \in \mathcal{J}$, and $k \in \mathcal{K}$. Since the events of detecting the transmission of interferers are mutually exclusive, the first term is equal to the second term. Hence, for $\omega_{\mathrm{MAC}} = S_1$, the deviation between Ψ and $\tilde{\Psi}$ is $2\tau_{0j}|p_{ik} - \tilde{p}_{i'k}|$, where $i \in \{0, 1, \ldots, N\}$, $i' = \phi_{\theta,\mathrm{CH}}(i)$, $j \in \mathcal{J}$ and $k \in \mathcal{K}$.

Thirdly, when $\omega_{\mathrm{MAC}} = S_2$ and the agent chooses to wait, ω_{MAC} transits from S_2 to S_1. The state transition probability is deterministic and does not depends on the channel state. The associated reward, however, has a deviation as follows:

$$\begin{aligned}&\big|R((S_2, j, i, k), (S_1, j, 0, k), 0) - \tilde{R}_\theta\big(\big(S_2, j, i', k\big), (S_1, j, 0, k), 0\big)\big|\\ &\quad = \left|l_i - \tilde{l}_{i'}\right|,\end{aligned}$$

where $i \in \mathcal{N}$, $i' = \phi_{\theta,\mathrm{CH}}(i)$, $j \in \mathcal{J}$ and $k \in \mathcal{K}$.

Finally, note that $\tau_{00} = \max_{j\in\mathcal{J}} \tau_{0j}$ and $p_{0k} = \tilde{p}_{0k}$, the deviation between Ψ and $\tilde{\Psi}$ for every state $\omega \in \Omega$ and action $a \in \mathcal{A}$ is calculated as (22). ∎

Plugging (22) into Theorem 2 and noticing the fact that $|V|_{\max}$ is bounded by $\frac{|R|_{\max}}{1-\gamma}$, the following lemma is derived.

Lemma 1: Let V^{π^*} denote the optimal value function where the agent identifies the interferer and treats different OBSS interferers differently. Let $V^{\hat{\pi}^*}$ denote the optimal value function where the agent does not identify the interferer and only distinguishes whether the interference power is above or below θ. Then, $\forall\omega \in \Omega$,

$$\left|V^{\hat{\pi}^*}(\omega) - V^{\pi^*}(\omega)\right| \leq \frac{2\delta_\theta|R|_{\max}}{(1-\gamma)^2}, \tag{23}$$

where $|R|_{\max}$ denotes the maximum achievable absolute value of reward.

It can be seen from this theorem that, identifying interferes is more attractable if δ_θ is large. This is more likely to happen if the agent and its associated STA are apart from each other as shown in Fig. 2, where p_{ik} may deviate largely from $\tilde{p}_{i'k}$.

VII. Numerical Evaluation

A. Evaluation Settings

We evaluate the proposed learning-based spatial reuse scheme through MATLAB based simulations. The simulation evaluations are dividend into three scenarios: single

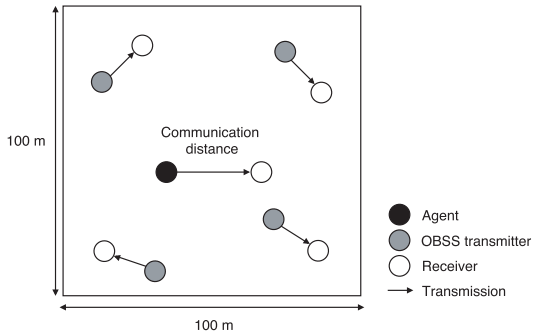

Fig. 7. Evaluation topology.

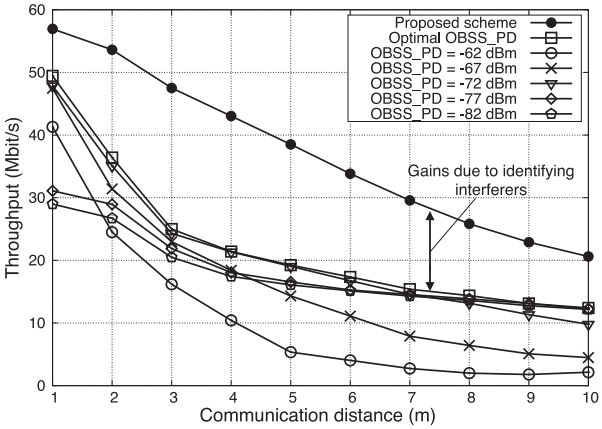

Fig. 8. Throughput of the agent with the communication distance. The number of OBSS transmitters: 4. The agent adopts the schemes in the legend.

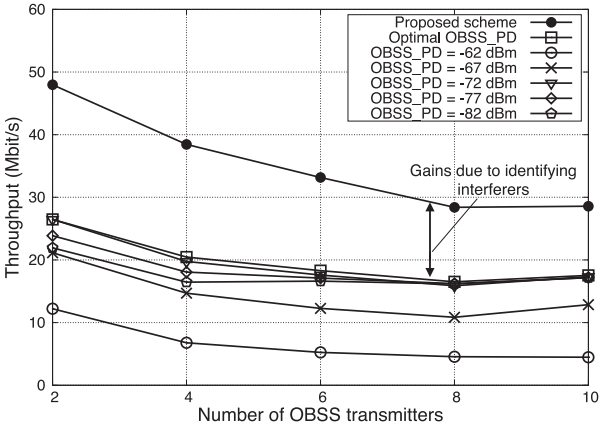

Fig. 9. Throughput of the agent with the number of OBSS transmitters. Communication distance: 5 m. The agent adopts the schemes in the legend.

agent static topology, single agent time-varying topology, and multiple agents static topology scenarios.

Firstly, the simulation to be shown in Figs. 8 to 11 and Fig. 13 evaluate the proposed scheme in a single agent and static topology scenario. In these simulations, the agent and other OBSS transmitters are randomly placed in a square region with side length 100 m as shown in Fig. 7. The agent and each OBSS transmitter have one associated receiver. The distance between them is called communication distance. The agent and each OBSS transmitter are assumed to have saturated downlink traffic, i.e., they are always backlogged with

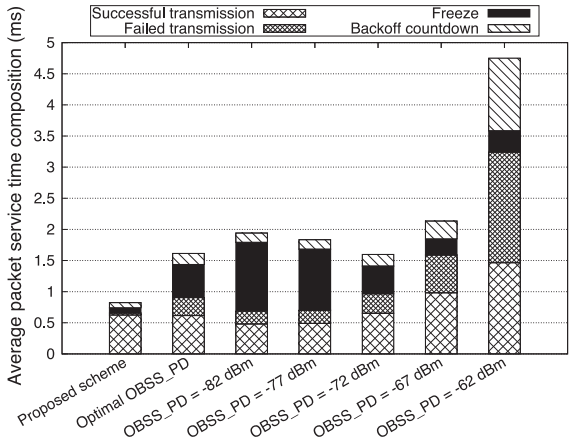

Fig. 10. MAC service time composition. Communication distance: 5 m. Number of OBSS transmitters: 4. The agent adopts the schemes in the horizontal axis.

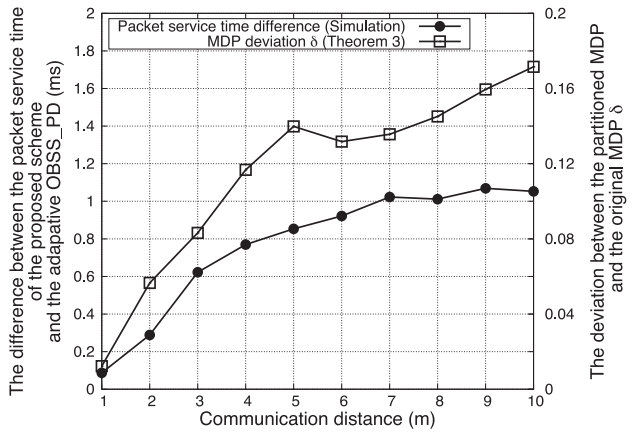

Fig. 11. Performance gains due to identifying interferers. The number of OBSS transmitters: 4. The agent adopts the proposed scheme.

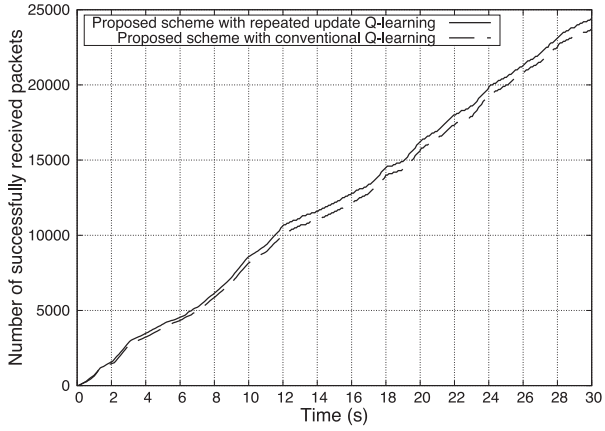

Fig. 12. Number of successfully transmitted packets. Communication distance: 5 m. Number of OBSS transmitters: 4. The locations each OBSS transmitter change once a second.

packets to send. The number of OBSS transmitters and the communication distance of the agent are stated in each evaluation. Unless otherwise stated, each OBSS transmitter adopts a fixed OBSS_PD of -82 dBm and the communication distance

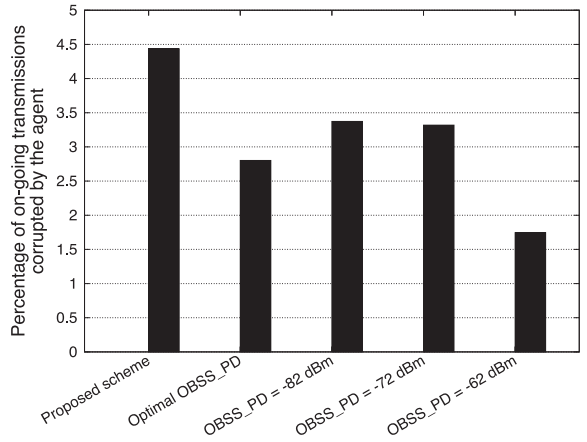

Fig. 13. Percentage of on-going transmissions corrupted by the agent. Number of OBSS transmitters: 4. The communication distance of each AP is randomly distributed from 1 m to 10 m. The agent adopts the schemes in horizontal axis.

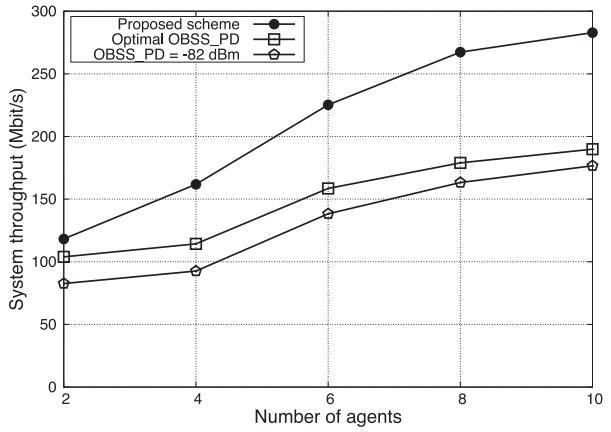

Fig. 14. System throughput in the multiple agents case. The communication distance of each agent is randomly distributed from 1 m to 10 m. All agents adopt the same scheme in the legend.

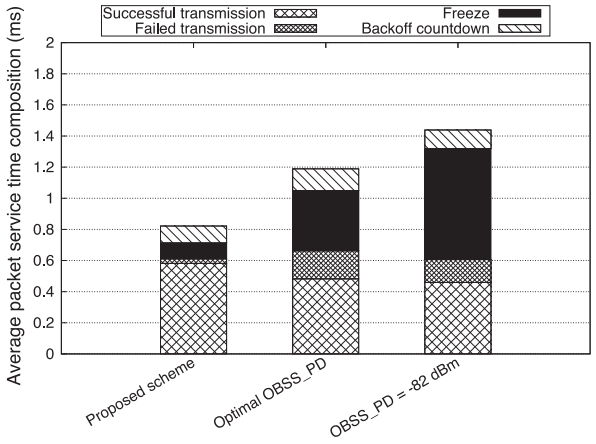

Fig. 15. Packet service time composition in the multiple agents case. The communication distance of each agent is randomly distributed from 1 m to 10 m. Number of agents: 5. All agents adopt the same scheme in horizontal axis.

TABLE II
EVALUATION PARAMETERS [4], [33], [34]

Parameters	Value
Slot time	9 μs
DIFS	34 μs
SIFS	16 μs
ACK	44 μs
ACK timeout	60 μs
Maximum transmit power	21 dBm
Noise power	−101 dBm
Discount factor γ	0.99
Exploration factor ε	0.1

TABLE III
REQUIRED SINR AND TRANSMISSION TIME FOR IEEE 802.11AX [4], [33], [34]

Data rate	Required SINR	Transmission time (header+ data)
8.6 Mbit/s	1 dB	3844 μs
17.2 Mbit/s	4 dB	1937 μs
25.8 Mbit/s	6 dB	1302 μs
34.4 Mbit/s	9 dB	984 μs
51.6 Mbit/s	13 dB	666 μs
68.8 Mbit/s	17 dB	508 μs
77.4 Mbit/s	18 dB	455 μs
86 Mbit/s	19 dB	412 μs
103.2 Mbit/s	24 dB	349 μs
114.7 Mbit/s	26 dB	317 μs
129 Mbit/s	29 dB	285 μs
143.4 Mbit/s	31 dB	260 μs

of each OBSS transmitter is 1 m. In these simulations, we generate 100 different random patterns of locations. We evaluate each pattern of locations for 10 s and take the average of the evaluation results.

Secondly, the simulation to be shown in Fig. 12 evaluate the proposed scheme in a single agent and time-varying topology scenario. The evaluation conditions are basically the same as the first scenario, whereas the locations of each OBSS transmitter changes once a second randomly. This simulation runs for 30 s.

Finally, the simulation to be shown in Figs. 14 and 15 evaluate the proposed scheme in a multiple agents scenario. The evaluation conditions are basically the same as the first scenario, whereas all the transmitters are non-cooperative and independent agents. Same as the first scenario, we generate 100 different random patterns of locations. We evaluate each pattern of locations for 10 s and take the average of the evaluation results.

The simulation program that we used to evaluate the proposed scheme is an event-driven simulator written in MATLAB. The program mainly attempts to simulate the MAC layer operation of APs and STAs, including random back-off procedure. Note that this kind of simulation model is widely used in related studies such as [30]. Evaluations parameters are as shown in Table II and III. We consider a distance-based path loss model in [31], where the center frequency f_c is 5200 MHz and the path loss coefficient $N_{\mathrm{P}L}$ is 30 for indoor residential scenario, i.e.,

$$PL_{\mathrm{dB}}(d) = 20\log_{10}(f_c) - 28 + N_{\mathrm{P}L}\log_{10}(\min(d, 1)). \quad (24)$$

We assume that a packet will be successfully decoded if the required SINR is met, where the payload is 4096 B [4].

We consider that the contention window size updates according to the binary exponential backoff (BEB) algorithm in the IEEE 802.11 [21]. The contention window size of after j times of consecutive packet loss is given as follows:

$$\mathrm{CW}_j = \begin{cases} 2^j(\mathrm{CW}_{\min}+1)-1, & j=0,1,\ldots,6; \\ 2^6(\mathrm{CW}_{\min}+1)-1, & j>6, \end{cases} \tag{25}$$

where $\mathrm{CW}_{\min}=15$ is the minimum contention window size.

We consider several performance comparison benchmarks in the following evaluations. The benchmark *optimal OBSS_PD* represents the highest performance under an exhaustive search of all the integer values of OBSS_PD within [-82 dBm, -62 dBm]. In other words, the optimal OBSS_PD can be seen as the performance upper bound of threshold-based CCA policies. In the comparison benchmarks, the rate selection scheme auto-rate feedback (ARF) [32] is used, which tunes up the data rate after two consecutive successful transmissions and tunes it down after one unsuccessful transmission.

B. Evaluations Results

1) Throughput: In Figs. 8 and 9, we evaluate the throughput of the proposed scheme under different conditions of communication distance and the number of OBSS transmitters.

From Fig. 8, we first notice that the comparison benchmark optimal OBSS_PD outperforms other comparison benchmarks of fixed OBSS_PD under different conditions of communication distance. The throughput gain of the optimal OBSS_PD compared to fixed OBSS_PD schemes comes from the ability to adjust the OBSS_PD according to communication distance. When the communication distance is short, a high value of OBSS_PD achieves good performance. This is because the received signal strength is high and conservative concurrent transmission policy decreases transmission opportunities. When the communication distance increases, the OBSS_PD value that achieves the highest throughput decreases. This is because the received signal strength is low and aggressive concurrent transmission policy increases the possibility of transmission failures.

In Fig. 8, the proposed scheme outperforms the comparison benchmark optimal OBSS_PD. This demonstrates the performance gains of identifying interferers in learning concurrent transmissions under different conditions of communication distance. Although the optimal OBSS_PD allows to adjust the OBSS_PD to communication distance, it treats different OBSS interferers indifferently.

We have also conducted one-tail paired *t*-test to evaluate the statistical accuracy of this result. The *t*-test is a statistic method that is commonly used to determine if the means of two sets of data are significantly different from each other [35]. We consider the null hypothesis that the proposed scheme does not achieve higher throughput than the benchmark optimal OBSS_PD. Given the simulation data generated from 100 different random patterns of locations, the null hypothesis is rejected at a *p*-value of 0.05. In other words, the result that the proposed scheme achieves higher throughput than the optimal OBSS PD is at least 95% confident.

It can also be confirmed in Fig. 9 that the optimal OBSS_PD outperforms other comparison benchmarks of fixed OBSS_PD and the proposed scheme outperforms the optimal OBSS_PD. We have also conducted one-tail paired *t*-test where the null hypothesis is that the proposed scheme does not achieve higher throughput than the benchmark optimal OBSS_PD. The *t*-test shows that the null hypothesis is rejected at a *p*-value of 0.05. This demonstrates the performance gains of interferer identification in facilitating concurrent transmissions under different number of OBSS transmitters.

2) MAC Service Time Composition: The composition of MAC service time is evaluated in Fig. 10. Remember that the MAC service time is the duration from the instant when the packet arrives at the head of the transmission queue to the instant when the agent has received the ACK from the receiver. As shown in (2), the MAC service time comprises four components: the duration of counting down backoff counter, freezing backoff counter, failed transmissions, and successful transmission.

The evaluation results in Fig. 10 reveal that the proposed scheme reduces the time of freezing backoff counter compared to the fixed OBSS_PD of -82 dBm while keeping the percentage of failed transmission low. The proposed scheme achieves the shortest packet service time of less than 0.85 ms. In the comparison benchmark with -82 dBm OBSS_PD, the average packet service time is larger than 1.9 ms where about 57% of the MAC service time is freezing backoff timer. This indicates that the wireless channel is under high contention.

From Fig. 10, we also confirm the negative effects of increasing the OBSS_PD for facilitating concurrent transmissions. Note that increasing the OBSS_PD indeed reduces the time of freezing backoff counter; however, the other three components also increase. This is because, as the OBSS_PD increases, infeasible concurrent transmissions are more likely to occur. Transmission failures increase the contention window size and hence increase the backoff countdown duration. Transmission failures result in the agent choosing a slower data rate. Hence, the time of successful transmission also increases.

3) Performance Gains Due to Identifying Interferers: In Fig. 11, we evaluate the performance gains due to identifying interferers from both the simulation and the theoretical perspective.

From the simulation perspective, we evaluate the difference between the packet service time of the proposed scheme and the optimal OBSS_PD. Note that, this difference corresponds to the gaps in Figs. 8 and 9.

From the theoretical perspective, we calculate the deviation between the partitioned MDP and the original MDP. Here, we show how δ is calculated based on (22). Note that, there is no difference in the expected transmission durations of each interferer according to our considered simulation setting. Hence, we calculate δ_θ as $\delta_\theta = \max_{ik} 2\tau_{00}|p_{ik}-\tilde{p}_{i'k}|$. Furthermore, in the optimal OBSS_PD scheme, the agent searches for the optimal value of OBSS_PD that maximizes throughput. Hence, we calculate δ as follows:

$$\delta = \min_{\theta} \delta_\theta = \min_{\theta} \max_{ik} 2\tau_{00}|p_{ik}-\tilde{p}_{i'k}|. \tag{26}$$

Remember that p_{ik} denotes the packet error probability when the agent transmits concurrently with i by using data rate k.

The value of p_{ik} can be calculated from the locations and parameters in Tables II and III.

We first notice from Fig. 11 that the deviation δ increases as the distance increases. This is because when the communication distance increases, the interference power measured by the agent is more likely to deviate largely from that measured by its receiver. As a consequence, comparing the interference power with the OBSS_PD has limited predictive value in determining the success or failure of transmissions. We also notice that the theoretical deviation δ and the packet service time difference evaluated through simulations have similar shapes. This shows that, given node locations and necessary parameters, calculating δ provides a quantitative method of analyzing the gains due to identifying interferers.

4) Time-Varying Topology: All above mentioned evaluations evaluate the proposed scheme in the static topology environment. In Fig. 12, we present the evaluation result of the proposed scheme under time-varying topology. In this evaluation, the locations of each OBSS transmitter change once a second. The agent does not reset the learning algorithm within the evaluation period. Simulation result reveals that the proposed scheme achieves less throughput in the time-varying topology than that in the static topology in Figs. 8 and 9. The reason is that the relative distance between the interferer and the agent may have already changed significantly before the agent learns the optimal action. Note that, the performance degradation due to the time-varying environment is in a sense inevitable as long as the learning-based scheme is utilized.

On the other hand, simulation results also confirm that the proposed scheme with RUQL algorithm outperforms the conventional QL algorithm in the time-varying topology. In the RUQL, the value functions of those optimal actions with temporal lower values are updated more frequently than that in the conventional QL algorithm. Hence, it can be inferred that the agent learns the optimal action faster by using RUQL. As a result, the RUQL outperforms the conventional QL in the time-varying topology.

5) Impact to Legacy Transmitters: All abovementioned simulations focus on the evaluations of the performance gains of the proposed scheme. Notice that, there is a concern if the concurrent transmission of the agent can corrupt on-going transmissions in OBSSs. In this evaluation, we evaluate the percentage of packets transmitted by the OBSS transmitters that are corrupted by the transmission of the agent. These packets are defined as failed transmissions which would be received successfully if the agent is not transmitting. Note that we assume that all the OBSS transmitters are legacy devices with a fixed OBSS_PD of -82 dBm. The communication distance of each AP is randomly distributed from 1 m to 10 m.

The simulation result in Fig. 13 shows that the proposed scheme corrupts approximately 4% of the packets transmitted by legacy devices. This indicates that, by applying the transmit power restriction rule (7), the agent does not cause serious corruptions to legacy devices. Although the comparison benchmark of a fixed OBSS_PD of -62 dBm causes fewer corruptions to legacy devices, it restricts concurrent transmission power by 20 dB and causes low throughput performance.

6) Multiple Agents: All abovementioned evaluations evaluate the proposed scheme from a single agent perspective. In Figs. 14 and 15, we evaluate the multiple agents case where there are multiple agents using the proposed scheme and each agent wishes to maximize its own value function. We consider that the communication distance of each agent is randomly distributed from 1 m to 10 m.

We confirm that the proposed scheme achieves better performance than comparison benchmarks in the multiple agents case. Fig. 14 evaluates the system throughput in multiple agents case, where all the transmitters adopt the same scheme as shown in the legend. The case where all the APs use the proposed scheme achieves the highest area throughput. Fig. 15 evaluates the packet service time composition in multiple agents case. The case where all the APs use the proposed scheme achieves the shortest packet service time. The gain of identifying interfering transmitters is also confirmed in the multiple agents case. In Fig. 14, we have conducted one-tail paired t-tests. The null hypothesis is that the proposed scheme does not achieve higher throughput than the benchmark optimal OBSS_PD. The t-tests show that the null hypothesis is rejected at a p-value of 0.05.

We observe that there is no obvious performance degradation in the multiple-agent scenario compared to that in the single-agent case. One reason is that the transmit power restriction rule (7) has alleviated the interactions among agents. When one agent decides to transmit concurrently with another on-going transmission, it does not generate strong interference to that on-going transmissions.

C. Practical Implications

From the simulation evaluations presented above, we may find some insights into designing spatial reuse operation of WLANs. First of all, OBSSs should be avoided as much as possible, especially when the communication between AP and STA is large. Since if the communication distance is large, AP does not have an accurate understanding of the interference at a remote STA by performing carrier sensing. Second, if the existence of OBSSs is inevitable and the communication distance between AP and STA is large, it is unreasonable to set a common CCA threshold to all the OBSS interferers. It is desirable that the AP can treat different interferers differently for deciding whether or not to exploit spatial reuse.

VIII. Conclusion

In this work, we proposed a reinforcement learning-based spatial reuse scheme. When the agent overhears an on-going transmission, it utilizes the information in the detected frame header to identify the interferer and decides whether or not to freeze the backoff counter accordingly. We have evaluated the proposed scheme under various scenarios through simulations. Specifically, we analyzed the composition of MAC layer service time. We found that the proposed scheme reduces the time

of freezing the backoff counter while keeping the number of failed transmissions low. This confirms that, on the one hand, the agent learns to transmit concurrently with those OBSS interferers whose interference is tolerable at the receiver. On the other hand, the agent learns to refrain from transmitting concurrently with those interferers whose interference is not tolerable at the receiver. Moreover, we also utilized the concept of state partition in MDP to study the performance gains due to making non-binary identifications of interferers on exploiting spatial reuse in WLANs. A theoretical bound on the gains in value function due to identifying interferers is obtained.

REFERENCES

[1] D.-J. Deng *et al.*, "IEEE 802.11ax: Highly efficient WLANs for intelligent information infrastructure," *IEEE Commun. Mag*, vol. 55, no. 12, pp. 52–59, Dec. 2017.

[2] M. S. Afaqui, E. Garcia-Villegas, and E. Lopez-Aguilera, "IEEE 802.11ax: Challenges and requirements for future high efficiency WiFi," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 130–137, Jun. 2017.

[3] H. A. Omar, K. Abboud, N. Cheng, K. R. Malekshan, A. T. Gamage, and W. Zhuang, "A survey on high efficiency wireless local area networks: Next generation WiFi," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2315–2344, 4th Quart., 2016.

[4] 802.11 Working Group of the 802 Committee, *TM/D4.0 Amendment 6: Enhancements for High Efficiency WLAN*, IEEE Standard P802.11AX, Feb. 2019.

[5] T. Joshi, D. Ahuja, D. Singh, and D. P. Agrawal, "SARA: Stochastic automata rate adaptation for IEEE 802.11 networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 11, pp. 1579–1590, Nov. 2008.

[6] D. Aguayo, J. Bicket, S. Biswas, G. Judd, and R. Morris, "Link-level measurements from an 802.11b mesh network," in *Proc. Conf. Appl. Technol. Architect. Protocols Comput. Commun.*, Portland, OR, USA, Aug. 2004, pp. 121–132.

[7] Y. Zhu, Q. Zhang, Z. Niu, and J. Zhu, "On optimal QoS-aware physical carrier sensing for IEEE 802.11 based WLANs: Theoretical analysis and protocol design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1369–1378, Apr. 2008.

[8] H. Ma, R. Vijayakumar, S. Roy, and J. Zhu, "Optimizing 802.11 wireless mesh networks based on physical carrier sensing," *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1550–1563, Oct. 2009.

[9] L. Fu, S. C. Liew, and J. Huang, "Effective carrier sensing in CSMA networks under cumulative interference," *IEEE Trans. Mobile Comput.*, vol. 12, no. 4, pp. 748–760, Apr. 2013.

[10] J. Deng, B. Liang, and P. K. Varshney, "Tuning the carrier sensing range of IEEE 802.11 MAC," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dallas, TX, USA, Dec. 2004, pp. 2987–2991.

[11] D. M. Kim and S.-L. Kim, "An iterative algorithm for optimal carrier sensing threshold in random CSMA/CA wireless networks," *IEEE Commmu. Lett.*, vol. 17, no. 11, pp. 2076–2079, Nov. 2013.

[12] H. ElSawy and E. Hossain, "A modified hard core point process for analysis of random CSMA wireless networks in general fading environments," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1520–1534, Apr. 2013.

[13] K. Yamamoto, X. Yang, T. Nishio, M. Morikura, and H. Abeysekera, "Analysis of inversely proportional carrier sense threshold and transmission power setting," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2017, pp. 13–18.

[14] Y. Wen, H. Fujita, and D. Kimura, "Throughput-aware dynamic sensitivity control algorithm for next generation WLAN system," in *Proc. IEEE Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Feb. 2017, pp. 1–7.

[15] I. Selinis, M. Filo, S. Vahid, J. Rodriguez, and R. Tafazolli, "Evaluation of the DSC algorithm and the BSS color scheme in dense cellular-like IEEE 802.11ax deployments," in *Proc. IEEE Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–7.

[16] M. S. Afaqui, E. Garcia-Villegas, E. Lopez-Aguilera, and D. Camps-Mur, "Dynamic sensitivity control of access points for IEEE 802.11ax," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.

[17] W. Afifi, E.-H. Rantala, E. Tuomaala, S. Choudhury, and M. Krunz, "Throughput-fairness tradeoff evaluation for next-generation WLANs with adaptive clear channel assessment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[18] C. Thorpe and L. Murphy, "A survey of adaptive carrier sensing mechanisms for IEEE 802.11 wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1266–1293, Quart., 2014.

[19] S. Kim, S. Yoo, J. Yi, Y. Son, and S. Choi, "FACT: Fine-grained adaptation of carrier sense threshold in IEEE 802.11 WLANs," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1886–1891, Feb. 2017.

[20] J. Zhu, B. Metzler, X. Guo, and Y. Liu, "Adaptive CSMA for scalable network capacity in high-density WLAN: A hardware prototyping approach," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Barcelona, Spain, Apr. 2006, pp. 1–10.

[21] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard 802.11, Mar. 2012.

[22] T. Sakurai and H. L. Vu, "MAC access delay of IEEE 802.11 DCF," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1702–1710, May 2007.

[23] N. Mastronarde and M. van der Schaar, "Joint physical-layer and system-level power management for delay-sensitive wireless communications," *IEEE Trans. Mobile Comput.*, vol. 12, no. 4, pp. 694–709, Apr. 2013.

[24] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[25] S. Abdallah and M. Kaisers, "Addressing environment non-stationarity by repeating Q-learning updates," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1582–1612, Apr. 2016.

[26] K. Levy, F. J. Vazquez-Abad, and A. Costa, "Adaptive stepsize selection for online Q-learning in a non-stationary environment," in *Proc. 8th Int. Workshop Discrete Event Syst.*, Ann Arbor, MI, USA, Jul. 2006, pp. 372–377.

[27] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Comput. Surveys*, vol. 46, no. 2, p. 25, Nov. 2013.

[28] R. Ortner, "Pseudometrics for state aggregation in average reward Markov decision processes," in *Proc. Int. Conf. Algorithmic Learning Theory*, Sendai, Miyagi, Japan, Oct. 2007, pp. 373–387.

[29] E. Even-Dar and Y. Mansour, "Approximate equivalence of Markov decision processes," in *Proc. Conf. Learn. Theory (COLT)*, Washington, DC, USA, Aug. 2003, pp. 581–594.

[30] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[31] "Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 450 GHz," Int. Telecommun. Union, Geneva, Switzerland, ITU-Recommendation P. 1238–10, Aug. 2019.

[32] A. Kamerman and L. Monteban, "WaveLAN-II: A high-performance wireless LAN for the unlicensed band," *Bell Labs Tech. J.*, vol. 2, no. 3, pp. 118–133, Aug. 1997.

[33] E. Perahia and R. Stacey, *Next Generation Wireless LANs: 802.11n and 802.11ac*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[34] Y. Inoue, K. Saitoh, T. Sakata, M. Morikura, and H. Matsue, "A study on the rate switching algorithm for IEEE 802.11 wireless LANs," *IEEJ Trans. Electron. Inf. Syst.*, vol. 124, no. 1, pp. 33–40, Apr. 2004.

[35] D. C. Montgomery, G. C. Runger, and N. F. Hubele, *Engineering Statistics*. New York, NY, USA: Wiley, 1998.

Bo Yin received the B.E. degree in electrical and electronic engineering from Kyoto University in 2016 and the M.E. degree from the Graduate School of Informatics, Kyoto University in 2018, where he is currently pursuing the Ph.D. degree. He received the VTS Japan Young Researcher's Encouragement Award in 2017.

Koji Yamamoto (S'03–M'06) received the B.E. degree in electrical and electronic engineering and the master's and Ph.D. degrees in informatics from Kyoto University in 2002, 2004, and 2005, respectively. From 2004 to 2005, he was a Research Fellow with the Japan Society for the Promotion of Science. Since 2005, he has been with the Graduate School of Informatics, Kyoto University, where he is currently an Associate Professor. From 2008 to 2009, he was a Visiting Researcher with Wireless@KTH, Royal Institute of Technology, Sweden. His research interests include radio resource management, game theory, and machine learning. He received the PIMRC 2004 Best Student Paper Award in 2004, the Ericsson Young Scientist Award in 2006, the Young Researcher's Award, the Paper Award, the SUEMATSU-Yasuharu Award from the IEICE of Japan in 2008, 2011, and 2016, respectively, and the IEEE Kansai Section GOLD Award in 2012. He serves as an Editor for the *IEEE Wireless Communications Letters* and *Journal of Communications and Information Networks*, a Track Co-Chair of the APCC 2017, CCNC 2018, APCC 2018, and CCNC 2019, and a Vice Co-Chair of the IEEE ComSoc APB CCC. He was a Tutorial Lecturer in ICC 2019. He is a Senior Member of the IEICE and the Operations Research Society of Japan.

Takayuki Nishio (M'87) received the B.E. degree in electrical and electronic engineering and the master's and Ph.D. degrees in communications and computer engineering from the Graduate School of Informatics from Kyoto University, Kyoto, Japan, in 2010, 2012 and 2013, respectively. From 2012 to 2013, he was a Research Fellow (DC1) with the Japan Society for the Promotion of Science. Since 2013, he is an Assistant Professor with the Communications and Computer Engineering, Graduate School of Informatics, Kyoto University. From 2016 to 2017, he was a Visiting Researcher with Wireless Information Network Laboratory, Rutgers University, U.S. His current research interests include mmWave networks, wireless local area networks, application of machine learning, and sensor fusion in wireless communications. He received the IEEE Kansai Section Student Award in 2011, the Young Researcher's Award from the IEICE of Japan in 2016, and the Funai Information Technology Award for Young Researchers in 2016.

Masahiro Morikura received B.E., M.E., and Ph.D. degrees in electronic engineering from Kyoto University, Kyoto, Japan, in 1979, 1981, and 1991, respectively. He joined NTT in 1981, where he was engaged in the research and development of TDMA equipment for satellite communications. From 1988 to 1989, he was with the communications Research Centre, Canada, as a Guest Scientist. From 1997 to 2002, he was active in standardization of the IEEE 802.11a based wireless LAN. He is currently a Professor with the Graduate School of Informatics, Kyoto University. He received Paper Award and Achievement Award from the IEICE in 2000 and 2006, the Education, Culture, Sports, Science and Technology Minister Award in 2007, the Maejima Award from the Teishin association in 2008, and the Medal of Honor with Purple Ribbon from Japan's Cabinet Office in 2015.

Hirantha Abeysekera received the B.Eng., M.Eng., and Ph.D. degrees in communications engineering from Osaka University, Japan, in 2005, 2007, and 2010, respectively. He joined NTT Network Innovation Laboratories, Yokosuka, Japan in 2010, where he was involved in the research and development of next-generation wireless LAN systems. He is currently working as a Senior Research Engineer with NTT Access Service Systems Laboratories, NTT Corporation. His research interests include resource allocation in wireless LANs. He received the IEEE VTS Japan Student Paper Award, in 2009.