Network Association for Cognitive Communication and Radar Co-Systems: A POMDP Formulation

Jingjing Wang[*†‡], Sanghai Guan[*‡], Chunxiao Jiang[*], Hongming Zhang[†], Yong Ren[*] and Lajos Hanzo[†]
[*]Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China
[†]School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK
[‡]Tsinghua National Laboratory for Information Science and Technology (TNList)
Email: chinaeephd@gmail.com

***Abstract*—In order to beneficially exploit wireless spectral resources, spectrum sharing between communication systems and radar systems has become a popular research topic. However, traditional network association strategies may not result in an efficient co-system. We circumvent this problem by formulating a partially observable Markov decision process (POMDP) aided network association scheme. For maximizing the network throughput, whilst minimizing the interference imposed on the radar user, communication users are capable of adaptively selecting underlay or overlay access mode. Moreover, a near-optimal reinforcement learning algorithm is proposed considering both the computational complexity and feasibility. Finally, simulations are conducted in order to evaluate the effectiveness of our proposed POMDP based network association scheme.**

***Index Terms*—Network association, POMDP, communication and radar co-system, reinforcement learning.**

I. Introduction

The critical information infrastructure has been evolving towards environmentally aware adaption, multi-functional implementations as well as towards big data handling. As one of critical information infrastructure components, radar systems are typically used for object-detection by means of analyzing the reflected radio waves to determine the range, angle or velocity of objects. By contrast, communication systems rely on the radio channels for transmitting information. In numerous civilian and military scenarios, two systems co-exist and depend on each other. For example, the object-detection information emanating from a radar system should be promptly transmitted to the command center via the communication system at a high integrity. Considering the scarce spectral resource, specific frequency bands have been invoked for spectrum sharing between the radar system and the communication system, such as the 3550–3650 MHz band. Hence, a well-designed network association scheme is beneficial in terms of mitigating the interference between the pair of systems, which requires a cognitive communication and radar co-system [1] [2].

However, traditional network association strategies face numerous challenges in hybrid communication and radar systems. Specifically, the frequency-hopping radar makes it difficult for communication users to accurately estimate the rapidly time-varying channel state information (CSI). Moreover, considering the bursty nature of wireless traffic as well as the limited power consumption, it is impractical for the communication system to incessantly sense the whole channel. Furthermore, how to set up a considerate spectrum sharing scheme is another open challenge, given the non-uniform sub-channel occupation and the presence of ambient interferences [3], [4]. The aforementioned challenges require new network association methods, which are specifically designed for the communication and radar co-system.

A range of joint optimization schemes have been conceived for the communication and radar co-system [5]–[7]. Specifically, in [5], Turlapaty *et al.* proposed a joint design of the radar transmission waveform and of the power spectral density of a multi-carrier system. This joint design was beneficial both in terms of enhancing the radar functions as well as of maintaining a high throughput for the communication system. Furthermore, a network association was formulated for maximizing the interference plus noise ratio (SINR) at the radar receiver by Li *et al.* in [6]. In [7], an adaptive radar beamforming approach was proposed by Geng *et al.*, which was capable of efficiently mitigating the wireless interferences imposed by the communication system during spectrum sharing.

However, these challenging joint optimization problems require accurate CSI and tend to suffer from a high computational complexity. In reality, the bursty nature of the traffic and the power constraint hamper accurate CSI estimation. As a popular member of the reinforcement learning family, the partially observable Markov decision process (POMDP) has been widely used in network association researches, which is capable of mapping specific situations to particular actions so as to maximize a pertinent numerical reward function [8]–[11]. Inspired by the above-mentioned open problems, in this paper, we conceive a novel POMDP-based network association scheme for communication and radar co-systems. Our original contributions are summarized as follows:

- We formulate a POMDP based network association scheme for communication and radar co-systems, which is capable of nimbly adapting to dynamically fluctuating environments, whilst of efficiently exploiting the spectral resource.
- In view of the physical reality and computational complexity, a learning aided algorithm is proposed, which provides a near-optimal solution for our network association problem.
- Simulations are conducted, which evaluate the superior performance of our proposed learning algorithm in terms of improving the network's throughput in the face of deleterious interferences.

The remainder of this article is outlined as follows. The system model is detailed in Section II. An iterative POMDP-based sensing and access decision-making strategy is formulated in Section III, followed by our simulation results in Section IV and conclusions in Section V.

II. System Model

In this context, we consider two kinds of uses, i.e. primary users (PUs) and secondary users (SUs). More specifically, PU is unaware of the existence of SUs and has unhindered access to the wireless channel. In contrast to PU, SUs firstly sense the state of the wireless channel at the beginning of each time slot and then select an access strategy relying on the outcome of their sensing results.

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

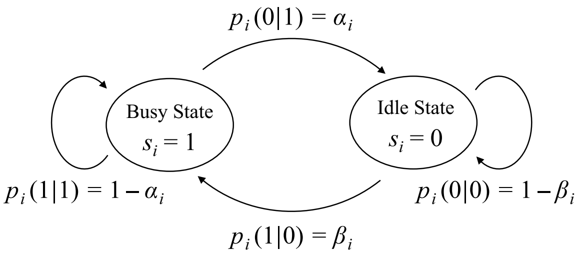


Fig. 1. The Markov process for the state transition of the sub-channel i.

A. Primary User

Radar systems detect and track objects by receiving and processing the waves reflected by the objects. In this treatise, radar systems are viewed as PUs, which constitute the primary network (PN). In order to avoid interference, frequency-hopping radar systems are considered, which are characterized by the discontinuous scanning of the time-, frequency- and spatial-resource slots. However, the frequency hopping mechanism creates spectrum holes, which may be exploited by communication systems.

B. Secondary User

In our paper, SUs are communication users. They are served by a communication base station (BS). The BS is in charge of both sensing the channel and of formulating appropriate access strategies for the SUs. The SUs as well as the BS construct the secondary network (SN) sharing the frequency band with the PU. SUs can make use of free channels and multiplex the occupied channels provided that the SINR constraint is not violated.

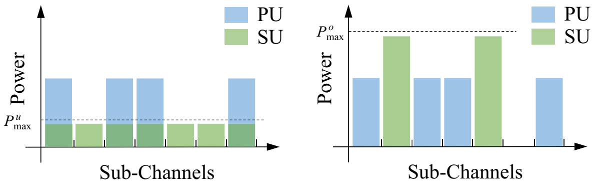
C. Co-system Model

1) System State: The total bandwidth of the co-system is denoted as W, where N sub-channels can be sensed and accessed. The subchannel bandwidths are denoted as $W_1, W_2, \ldots, W_N$. These N sub-channels are assigned to PUs, also termed as authorized users, which are capable of accessing a single or even several sub-channels. Hence, each sub-channel has two states at each time slot, i.e. the 'busy' state when PUs are transmitting signals as well as the 'idle' state when PUs are not using the sub-channel. Let $s_i(t)$ represent the state of the sub-channel i at the time slot t. Then we have $s_i(t) = 1$ if its state is busy, while $s_i(t) = 0$ if the sub-channel is idle. Therefore, the co-system's state at time slot t can be represented by $\mathbf{S}(t) = [s_1(t), s_2(t), \ldots, s_N(t)], s_i(t) \in \{0, 1\}$ and it has a total of 2^N different forms. The co-system's state set is denoted as $\mathbb{S}$ and we have $|\mathbb{S}| = 2^N$.

We model the state transitions of the sub-channel i with a Markov process, as shown in Fig. 1, where α_i represents the probability of the channel transferring from busy to idle, while β_i denotes the probability of the state evolving from idle to busy. Then, the state transition probability of sub-channel i can be formulated as:

$$p_i(s_i' \mid s_i) = \Pr\{s_i(t+1) = s_i' \mid s_i(t) = s_i\}, \ s_i, s_i' \in \{0, 1\}. \tag{1}$$

Let p_i^0 and p_i^1 represent the probability of the sub-channel i staying at the idle state and at the busy state when the above-mentioned Markov process reaches its steady state, respectively. Hence, we have $p_i^0 = \alpha_i/(\alpha_i + \beta_i)$ and $p_i^1 = \beta_i/(\alpha_i + \beta_i)$. Upon relying on the independence of sub-channels, the co-system's state transition probability can be expressed by:



(a) Underlay access scheme.

(b) Overlay access scheme.

Fig. 2. The underlay and overlay access scheme.

$$\begin{aligned} p(\mathbf{S}' \mid \mathbf{S}) &= \Pr\{\mathbf{S}(t+1) = \mathbf{S}' \mid \mathbf{S}(t) = \mathbf{S}\} \\ &= \prod_{i=1}^{N} \Pr\{s_i(t+1) = s_i' \mid s_i(t) = s_i\}, \end{aligned} \tag{2}$$

where $\mathbf{S}' = [s_1', s_2', \ldots, s_N']$ and $\mathbf{S} = [s_1, s_2, \ldots, s_N]$.

2) Access Mechanism: In our model, opportunistic spectrum management (OSM) is envisioned for SU's channel selection. The spectrum management decision-making can be divided into a pair of processes, i.e. the sensing process as well as the access process.

When considering the energy constraint, the communication BS is capable of observing at most M sub-channels during the sensing process at time slot t, where $M < N$. Upon relying on the previous observed states, we aim for selecting M sub-channels for the sake of accurately estimating the system's actual state $S(t)$, which can be viewed as the first-step action of the SN. This first-step action set is denoted by $\mathbb{A}_1 = \{\mathbf{A}_1\}$, where $\mathbf{A}_1 = [a_1^1, a_2^1, \ldots, a_N^1] \in \{0, 1\}^N$ and $|\mathbb{A}_1| = \binom{N}{M}$. If SU decides to sense the ith sub-channel, we have $a_1^i = 1$; otherwise we have $a_1^i = 0$. Moreover, a maximum of M observable sub-channels requires that $\sum_{i=1}^{N} a_i^1 \leq M$. Furthermore, the false-alarm rate and the missed-detection rate of the sensing process are denoted by ζ_f and ζ_m, respectively. Relying on M sensed sub-channels, we define the observed states of our co-system at time slot t as $\mathbf{O}(t) = [o_1(t), o_2(t), \ldots, o_N(t)]$, where $o_i(t) \in \{0, 1, \phi\}$. If the sensing result of sub-channel i is idle, we have $o_i(t) = 0$, whilst if it is busy, we have $o_i(t) = 1$. We have $o_i(t) = \phi$ when the sub-channel i is not observed at time slot t. Moreover, $\mathbb{O}$ represents the observation state set.

In the access process, based on the estimation of the actual system's state, either the underlay or the overlay mode can be selected as the access scheme by SU. The second-step action set is represented by $\mathbb{A}_2 = \{A_2\}$, and $A_2 \in \{a_u^2, a_o^2\}$. Specifically, if the underlay access scheme is invoked as the second-step action, we have $A_2 = a_u^2$, while $A_2 = a_o^2$ if SN selects the overlay access scheme. Furthermore, the power that the SN allocates to the N sub-channels is denoted by $\mathbf{P} = [P_1, P_2, \ldots, P_N]$. As shown in Fig. 2, the two aforementioned access schemes can be elaborated as follows [12] [13].

- *Underlay Scheme*: SUs access the whole channel shared with the PUs in terms of a low and equal transmission power in each sub-channel. Hence, we have $P_1^u = P_2^u = \cdots = P_N^u$ and $P_i^u \leq P_{\max}^u$, where $P_{\max}^u$ indicates the interference constraint of PUs. Moreover, the interference constraint of PUs is parameterized by the frequency-hopping radar's performance in order to guarantee its detection probability as well as false alarm probability.
- *Overlay Scheme*: SUs access L sub-channels that are deemed to be idle, relying on the channel state estimation. Furthermore, SUs are capable of using a higher transmission

power than that in the underlay mode. The transmission power in the i-th idle sub-channel can be set to P_i^o, and we have $P_i^o \leq P_{\max}^o$, where $P_{\max}^o$ depends on the transmitter's performance of the BS.

Therefore, the SN's decision-making concerning spectrum management hinges on the above-mentioned first-step action as well as on the second-step action. Thus, the action set of the whole spectrum management process can be formulated as $\mathbb{A} = \mathbb{A}_1 \times \mathbb{A}_2$, where $|\mathbb{A}| = 2\binom{N}{M}$.

3) Reward: The reward R of our proposed co-systems is the total reward of all sub-channels, i.e.

$$R = \sum_{i=1}^{N} R_i. \tag{3}$$

Specifically, if the SU does not access sub-channel i, we have $R_i = 0$. By contrast, if the SU accesses sub-channel i, the reward R_i includes two parts, i.e. the *channel capacity gain* R_{ig} as well as the *interference penalty* R_{ip}. And we have:

$$R_i = R_{ig} + R_{ip}. \tag{4}$$

To elaborate, the *channel capacity gain* R_{ig} can be expressed as:

$$R_{ig}(P_i, N_i) = \lambda_C W_i \log\left(1 + \frac{g_{sr} P_i}{(g_{pr} N_i + N_0) W_i}\right), \tag{5}$$

where λ_C denotes the weight coefficient, while W_i represents the bandwidth of sub-channel i. Furthermore, N_i and N_0 denote the average power spectral density of the radar system and of the white noise on sub-channel i, respectively. When the sub-channel i is in the idle state, we have $N_i = 0$. Moreover, g_{sr} and g_{pr} represent the receiver's power gain of SUs and the transmitter's power gain of PUs, respectively.

However, if the SU accesses a busy sub-channel, this will inevitably impose serious interference on the radar system, whilst increasing the false alarm rate or reducing the successful detection rate. In order to avoid reckless access by the SU, we formulate the *interference penalty* R_{ip} as:

$$R_{ip}(P_i, N_i) = \begin{cases} 0, & \text{if } s_i = 0, \\ -\lambda_I \cdot \dfrac{g_{sp}[P_i - P_{\max}^u]_+}{N_i W_i}, & \text{if } s_i = 1, \end{cases} \tag{6}$$

where λ_I represents a weighting coefficient, while g_{sp} denotes receiver's power gain power of PUs. We can conclude that $R_{ip} = 0$ when the SU selects the underlay scheme, while there is a risk of a detrimental interference penalty of R_{ip}, when the SU selects the overlay scheme.

III. The POMDP Approach

A. Belief State

In this subsection, first of all, we define the observation function of $z_i(o_i|s_i, a_i^1)$, which represents the probability of the sensing result of sub-channel i, namely o_i, under the condition of the first-step action a_i^1 at the system's state s_i. Hence, we have:

$$z_i(o_i \mid s_i, a_i^1) = \Pr\big(o_i(t) = o_i \mid s_i(t) = s_i, a_i^1(t) = a_i^1\big). \tag{7}$$

Considering the value of o_i, s_i and a_i^1, when we have $a_i^1 = 1$, the observation function can be formulated as:

$$\begin{cases} z_i(0 \mid 0, 1) = 1 - \zeta_f, \\ z_i(0 \mid 1, 1) = \zeta_m, \\ z_i(1 \mid 0, 1) = \zeta_f, \\ z_i(1 \mid 1, 1) = 1 - \zeta_m, \end{cases} \tag{8}$$

where ζ_f and ζ_m represent the false-alarm rate and missed-detection rate, respectively. When $a_i^1 = 0$, it can be deduced that:

$$z_i(\phi \mid \cdot, 0) = 1. \tag{9}$$

In our model, we assume that the state transition of each sub-channel is independent from each other. Hence, the co-system's observation function at time slot t can be formulated as:

$$\begin{aligned} z(\mathbf{O}|\mathbf{S}, \mathbf{A}_1) &= \Pr\big(\mathbf{O}(t) = \mathbf{O} \mid \mathbf{S}(t) = \mathbf{S}, \mathbf{A}_1(t) = \mathbf{A}_1\big) \\ &= \prod_{i=1}^{N} \Pr\big(o_i(t) = o_i \mid s_i(t) = s_i, a_i^1(t) = a_i^1\big). \end{aligned} \tag{10}$$

Given the energy and capacity constraint of the communication BS, SUs are unable to estimate the accurate state of all sub-channels. Here, we define the estimation state in order to describe the system's state after sensing. Hence, the estimation state vector of N sub-channels in our co-system can be expressed by $\mathbf{\Theta}^{\mathbf{S}}(t) = [\theta_1^{s_1}(t), \theta_2^{s_2}(t), \ldots, \theta_N^{s_N}(t)]$, where $\theta_i^{s_i}(t)$ is the probability that sub-channel i is estimated to be at state s_i at time slot t. For the convenience of deduction, we use $\theta_i^0(t)$ to present the probability that sub-channel i is idle at time slot t, while $\theta_i^1(t) = 1 - \theta_i^0(t)$ represents the probability of being busy state.

Then, the system's belief state, namely $B_{\mathbf{S}}(t)$, is defined in order to represent the conditional probability of the co-system's state being at $\mathbf{S}$ on condition that the estimation state vector is $\mathbf{\Theta}^{\mathbf{S}}(t)$ at time slot t, i.e.

$$\begin{aligned} B_{\mathbf{S}}(t) &= \Pr\big(\mathbf{S}(t) = \mathbf{S} \mid \mathbf{\Theta}^{\mathbf{S}}(t) = \mathbf{\Theta}^{\mathbf{S}}\big) \\ &= \prod_{i=1}^{N} \theta_i^{s_i}(t). \end{aligned} \tag{11}$$

Moreover, we obtain the state vector $\mathbf{B}(t) = [B_{\mathbf{S}_1}(t), B_{\mathbf{S}_2}(t), \ldots, B_{\mathbf{S}_{2^N}}(t)] \in \mathbb{B}$, where $\mathbb{B}$ represents the system's belief state set and $|B(t)| = 2^N$. Hence, we can define the belief transition function $b(\mathbf{B}' \mid \mathbf{B}, \mathbf{A}_1)$ of our proposed POMDP framework, which represents the probability that the system's belief state transfers from $\mathbf{B}$ to $\mathbf{B}'$ with the first-step action $\mathbf{A}_1$ at time slot t. And we have:

$$\begin{aligned} &b(\mathbf{B}' \mid \mathbf{B}, \mathbf{A}_1) = \Pr\big(\mathbf{B}(t) = \mathbf{B}' \mid \mathbf{B}(t-1) = \mathbf{B}, \mathbf{A}_1(t) = \mathbf{A}_1\big) \\ &= \sum_{\mathbf{O} \in \mathbb{O}} \Big(\Pr\big(\mathbf{B}(t) = \mathbf{B}' \mid \mathbf{B}(t-1) = \mathbf{B}, \mathbf{A}_1(t) = \mathbf{A}_1, \mathbf{O}(t) = \mathbf{O}\big) \\ &\quad \cdot \Pr\big(\mathbf{O}(t) = \mathbf{O} \mid \mathbf{B}(t-1) = \mathbf{B}, \mathbf{A}_1(t) = \mathbf{A}_1\big)\Big). \end{aligned} \tag{12}$$

Relying on Eq. (10) and Eq. (11), we can deduce that:

$$\begin{aligned} &\Pr\big(\mathbf{O}(t) = \mathbf{O} \mid \mathbf{B}(t-1) = \mathbf{B}, \mathbf{A}_1(t) = \mathbf{A}_1\big) \\ &= \sum_{\mathbf{S} \in \mathbb{S}} \Big(\Pr\big(\mathbf{O}(t) = \mathbf{O} \mid \mathbf{B}(t-1) = \mathbf{B}, \mathbf{S}(t) = \mathbf{S}, \mathbf{A}_1(t) = \mathbf{A}_1\big) \\ &\quad \cdot \Pr\big(\mathbf{S}(t) = \mathbf{S} \mid \mathbf{B}(t-1) = \mathbf{B}, \mathbf{A}_1(t) = \mathbf{A}_1\big)\Big) \\ &= \sum_{\mathbf{S} \in \mathbb{S}} \Big(\Pr\big(\mathbf{O}(t) = \mathbf{O} \mid \mathbf{S}(t) = \mathbf{S}, \mathbf{A}_1(t) = \mathbf{A}_1\big) \\ &\quad \cdot \sum_{\mathbf{S}' \in \mathbb{S}} p(\mathbf{S} \mid \mathbf{S}') \cdot B_{\mathbf{S}'}\Big) \\ &= \sum_{\mathbf{S} \in \mathbb{S}} \Big(z(\mathbf{O} \mid \mathbf{S}, \mathbf{A}_1) \cdot \sum_{\mathbf{S}' \in \mathbb{S}} p(\mathbf{S} \mid \mathbf{S}') \cdot B_{\mathbf{S}'}\Big). \end{aligned} \tag{13}$$

Therefore, Eq. (12) can be rewritten as:

$$b(\mathbf{B}' \mid \mathbf{B}, \mathbf{A}_1) = \\ \sum_{\mathbf{O}\in\mathbb{O}} \Big(\Pr\big(\mathbf{B}(t) = \mathbf{B}' \mid \mathbf{B}(t-1) = \mathbf{B}, \mathbf{A}_1(t) = \mathbf{A}_1, \mathbf{O}(t) = \mathbf{O}\big) \\ \cdot \sum_{\mathbf{S}\in\mathbb{S}} \Big(z(\mathbf{O} \mid \mathbf{S}, \mathbf{A}_1) \cdot \sum_{\mathbf{S}'\in\mathbb{S}} p(\mathbf{S} \mid \mathbf{S}') \cdot B_{\mathbf{S}'} \Big)\Big). \tag{14}$$

To elaborate a little further, let $I\{\cdot\}$ represent an indicator function, and we have:

$$\Pr\big(\mathbf{B}(t) = \mathbf{B}' \mid \mathbf{B}(t-1) = \mathbf{B}, \mathbf{A}_1(t) = \mathbf{A}_1, \mathbf{O}(t) = \mathbf{O}\big) \\ = I\Big\{\mathbf{B}' = [B'_{\mathbf{S}_1}, B'_{\mathbf{S}_2}, \cdots, B'_{\mathbf{S}_{2^N}}]\Big\}, \tag{15}$$

where $B'_{\mathbf{S}}$ in Eq. (13), Eq. (14) and Eq. (15) can be calculated from Eq. (16) Relying on the Markov-aided system's transition probability as shown in Fig. 1, the numerator of the fraction in Eq. (16) can be given by:

$$z_i(o_i \mid s_i, a_i^1) \cdot \sum_{s'_i\in\{0,1\}} p\big(s_i \mid s'_i\big) \cdot \theta_i^{s'_i}(t-1) \\ = \begin{cases} (1-\zeta_m)\cdot\vartheta_i^1, & \text{if } o_i = 1, s_i = 1, a_i^1 = 1, \\ \zeta_f\cdot\vartheta_i^0, & \text{if } o_i = 1, s_i = 0, a_i^1 = 1, \\ \zeta_m\cdot\vartheta_i^1, & \text{if } o_i = 0, s_i = 1, a_i^1 = 1, \\ (1-\zeta_f)\cdot\vartheta_i^0, & \text{if } o_i = 0, s_i = 0, a_i^1 = 1, \\ \vartheta_i^1, & \text{if } o_i = \phi, s_i = 1, a_i^1 = 0, \\ \vartheta_i^0, & \text{if } o_i = \phi, s_i = 0, a_i^1 = 0, \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$

where

$$\vartheta_i^1 = (1-\alpha_i)\cdot\theta_i^1(t-1) + \beta_i\cdot\theta_i^0(t-1), \tag{18}$$

as well as

$$\vartheta_i^0 = \alpha_i\cdot\theta_i^1(t-1) + (1-\beta_i)\cdot\theta_i^0(t-1). \tag{19}$$

B. Access Scheme Selection

After the first-step action $\mathbf{A}_1$ at time slot t, the SU should select either the underlay or overlay scheme according to the updated estimation state. The second-step action aims for maximizing the system's expected reward, which can be described as:

$$A_2^*(t) = \underset{A_2(t)\in\{a_u^2, a_o^2\}}{\arg\max}\ \mathbb{E}\left[R(t) \,|\, \mathbf{\Theta}^{\mathbf{S}}, A_2(t)\right]. \tag{20}$$

In other words, the SU has to compare the expected reward of both the underlay scheme and of the overlay scheme based on the given estimation state. To elaborate, if the SU selects the underlay access scheme, the system's expected reward can be rewritten as:

$$\mathbb{E}\left[R \,|\, \mathbf{\Theta}^{\mathbf{S}}, A_2 = a_u^2\right] = \sum_{i=1}^{N} \mathbb{E}\left[R_i \,|\, \theta_i^{s_i}, P_i = P_{\max}^u\right] \\ = \sum_{i=1}^{N} \Big(\theta_i^1 \, R_{ig}\left(P_{\max}^u, N_i\right) + \theta_i^0 \, R_{ig}\left(P_{\max}^u, 0\right)\Big). \tag{21}$$

If the SU selects the overlay access scheme, it will access L 'most-likely-to-be-idle' sub-channels, namely Ω. Hence, the SU may adjust the transmission power P_i on each sub-channel $i, i \in \Omega$ for maximizing the system's expected reward, which can be formulated by:

$$\begin{aligned} \max_{P_i} \quad & \mathbb{E}\left[R \,|\, \mathbf{\Theta}^{\mathbf{S}}, A_2 = a_o^2\right], \\ \text{s.t.} \quad & P_{\max}^u \le P_i \le P_{\max}^o, \ i \in \Omega. \end{aligned} \tag{22}$$

If P_i^{o*} represents the optimal power allocation of each sub-channel, the system's expected reward in terms of the overlay access scheme can be calculated as:

$$\mathbb{E}\left[R \,|\, \mathbf{\Theta}^{\mathbf{S}}, A_2 = a_o^{2*}\right] = \sum_{i\in\Omega} \mathbb{E}\left[R_i \,|\, \theta_i^{s_i}, P_i = P_i^{o*}\right] \\ = \sum_{i\in\Omega} \Big(\theta_i^1 \big(R_{ig}\left(P_i^{o*}, N_i\right) + R_{ip}\left(P_i^{o*}, N_i\right)\big) + \theta_i^0 R_{ig}\left(P_i^{o*}, 0\right)\Big). \tag{23}$$

Comparing with Eq. (21) and Eq. (23), the SU selects the better scheme as the second-step action A_2. In this paper, we assume that the second-step action A_2 does not affect the belief state $\mathbf{\Theta}^{\mathbf{S}}$. This is because the SU can only receive the total reward R after carrying out the action A_2, and still cannot accurately obtain the actual state information of each sub-channel.

C. POMDP Formulation

Therefore, our proposed POMDP network association strategy can be formulated as a quintuple, i.e. $\langle \mathbb{S}, \mathbb{B}, \mathbb{A}, b, r\rangle$. Specifically,

- *Co-system State Set:* $\mathbb{S}$ is the set of all the possible co-system states, i.e. $\mathbb{S} = \{\mathbf{S}\}$ and $\mathbf{S} = [s_1, s_2, \cdots, s_N]$;
- *Belief State Set:* $\mathbb{B} = \{\mathbf{B}\}$, where $\mathbf{B}$ is the belief probability vector reflecting the grade of similarity between each possible system state $\mathbf{S} \in \mathbb{S}$ and the partially observed estimation state $\mathbf{\Theta}^{\mathbf{S}}$;
- *Action Set:* $\mathbb{A}$ is the set of all the possible users' actions with $\mathbf{A} \in \mathbb{A}$, where $\mathbf{A} = [\mathbf{A}_1, A_2]$ represents the users' specific actions in terms of which M sub-channels they sense and which of the two available access mechanisms they select;
- *Belief Transition Function:* $b\colon \mathbb{B} \times \mathbb{A}_1 \times \mathbb{B} \mapsto [0, 1]$, where the operand '$\times$' represents the Cartesian product, while $\mathbf{B}^{'} \in \mathbb{B}$ is the belief state during the next time slot;
- *Reward Function:* $r\colon \mathbb{B} \times \mathbb{A} \mapsto \mathbb{R}$, where $r(\mathbf{B}, \mathbf{A}_1, A_2)$ indicates the immediate reward as a consequence of two sequential actions $\mathbf{A}_1$ and A_2 under belief state $\mathbf{B}$, and we have $r(\mathbf{B}, \mathbf{A}_1, A_2) = \sum_{\mathbf{S}\in\mathbb{S}} B_{\mathbf{S}} \cdot R(\mathbf{S}, \mathbf{A}_1, A_2)$.

D. Optimal Policy

As we mentioned before, we have converted the discrete POMDP problem into a continuous belief-state MDP problem. In the following, let $G(t)$ represent the total discounted accumulated reward commenced in time slot t, which can be expressed by:

$$G(t) = \sum_{k=0}^{\infty} \gamma^k \cdot r\big(\mathbf{B}(t+k+1), \mathbf{A}_1(t+k+1), A_2(t+k+1)\big), \tag{24}$$

where γ $(0 \le \gamma \le 1)$ denotes the discount rate which determines the value of the future reward to the system. As γ increases, the system becomes more 'farsighted' and pays more attention to the future reward, and vice versa. The SU's objective is to take the appropriate action $\mathbf{A}(t)$ based on the current belief state $\mathbf{B}(t)$ to maximize $G(t)$. Therefore, the policy π can be denoted as $\pi\colon \mathbb{B} \mapsto \mathbb{A}$. In terms of different possible policies, the belief value function $V^{\pi}(\mathbf{B})$ is defined in order to characterize the expected $G(t)$ relying on the policy π and the prior belief $\mathbf{B}$, which can be written as:

$$V^{\pi}(\mathbf{B}) = \mathbb{E}_{\pi}\left[G(t) | \mathbf{B}(t) = \mathbf{B}\right], \tag{25}$$

Then, we can search for the optimal policy for the SU relying on the value iteration method [14]. Specifically, in our model, it can be found that the value function $V(\mathbf{B})$ is strongly associated with the belief state. It is reasonable to assume that the strong belief of an idle sub-channel contributes a high reward. Hence, the value function $V(\mathbf{B})$ can be approximately reformulated in

$$
\begin{aligned}
B'_{\mathbf{S}}(t) &= \frac{z(\mathbf{O}\mid\mathbf{S},\mathbf{A}_1)\cdot\sum_{\mathbf{S}'\in\mathbb{S}} p(\mathbf{S}\mid\mathbf{S}')\cdot B_{\mathbf{S}'}}{\Pr\big(\mathbf{O}(t)=\mathbf{O}\mid\mathbf{B}(t-1)=\mathbf{B},\mathbf{A}_1(t)=\mathbf{A}_1\big)} = \prod_{i=1}^{N}\theta_i^{s_i}(t) \\
&= \prod_{i=1}^{N}\frac{z_i(o_i\mid s_i,a_i^1)\cdot\Big(p\big(s_i\mid s_i'=1\big)\cdot\theta_i^1(t-1)+p\big(s_i\mid s_i'=0\big)\cdot\theta_i^0(t-1)\Big)}{\sum_{s_i''\in\{0,1\}}\Big(z_i(o_i\mid s_i'',a_i^1)\cdot\Big(p\big(s_i''\mid s_i'=1\big)\cdot\theta_i^1(t-1)+p\big(s_i''\mid s_i'=0\big)\cdot\theta_i^0(t-1)\Big)\Big)}.
\end{aligned}
\tag{16}
$$

the form of a non-linear polynomial function with respect to $\mathbf{\Theta}^1$ for the convenience. And we have:

$$\tilde{V}(\mathbf{B}) = \tilde{V}(\mathbf{\Theta}^1) = \mu^{\mathrm{T}}\,\phi_N(\mathbf{\Theta}^1), \tag{26}$$

where $\mu = [\mu_0,\mu_1,\mu_2,\ldots,\mu_{\chi-1}]^{\mathrm{T}}$ represents the regression coefficient, while $\phi_N(\mathbf{\Theta}^1) = \big[1,\theta_1^1,\theta_2^1,\ldots,\theta_1^1\theta_2^1\cdots\theta_N^1\big]^{\mathrm{T}}$ denotes an N-degree expansion function of the estimate of $\mathbf{\Theta}^1$. For a N sub-channels co-system, the length of vector $\phi_N(\mathbf{\Theta}^1)$ is $\chi=\sum_{i=0}^{N}\binom{N}{i}$. For reducing the complexity, in our model, we assume that the SU can only sense M sub-channels ($M<N$) at each time slot, which results in $\chi=\sum_{i=0}^{M}\binom{N}{i}$.

As shown in Algorithm 1, we propose an improved value iteration algorithm for the POMDP formulation considered. Instead of discretizing continuous belief states and calculating the optimal policy for each state, our algorithm samples sufficient belief states, which aims for optimizing the regression coefficient μ iteratively relying on the least square method. Then, we can achieve a relatively high value function of $\tilde{V}(\mathbf{B})$ from the regression coefficient μ obtained. Hence, given the belief state $\mathbf{B}$, the expected accumulated reward in terms of the action $\mathbf{A}$ can be calculated as:

$$Q(\mathbf{B},\mathbf{A}_1,A_2) = r(\mathbf{B},\mathbf{A}_1,A_2)+\gamma\sum_{\mathbf{B}'\in\mathbb{B}} b(\mathbf{B}'\,|\,\mathbf{B},\mathbf{A}_1)\cdot\tilde{V}(\mathbf{B}'). \tag{27}$$

Thus, we can get the near-optimal policy:

$$\pi^*(\mathbf{B}) = \underset{\mathbf{A}\in\mathbb{A}}{\arg\max}\, Q(\mathbf{B},\mathbf{A}_1,A_2). \tag{28}$$

It is noteworthy that our algorithm can also be extended to the scenarios that the SU does not have any prior knowledge about sub-channels.

Algorithm 1: A value iteration based network association algorithm for the co-system

```
generate $X$ estimated values of $\mathbf{\Theta}^1_{(1)},\ldots,\mathbf{\Theta}^1_{(X)}$ randomly;
calculate the corresponding belief states $\mathbf{B}_{(1)},\ldots,\mathbf{B}_{(X)}$;
initialize $\mu \leftarrow 0$, $\mu' \leftarrow \infty$ and $\overline{V}(\mathbf{B}_{(x)}) \leftarrow 0$ for all
  $x = 1,\ldots,X$;
while $\max|\mu-\mu'| > \epsilon$ do
    $\mu' \leftarrow \mu$;
    for $x = 1,\ldots,X$ do
        $\overline{V}(\mathbf{B}_{(x)}) \leftarrow \max_{\mathbf{A}\in\mathbb{A}}\Big(r(\mathbf{B}_{(x)},\mathbf{A}_1,A_2)+\gamma\cdot$
        $\sum_{\mathbf{B}'\in\mathbb{B}} b(\mathbf{B}'\,|\,\mathbf{B}_{(x)},\mathbf{A}_1)\cdot\overline{V}(\mathbf{B}')\Big)$;
    end
    optimize
    $\mu \leftarrow \underset{\mu}{\arg\min}\sum_{x=1}^{X}\big(\mu^{\mathrm{T}}\phi(\mathbf{\Theta}^1_{(x)})-\overline{V}(\mathbf{B}_{(x)})\big)^2$;
end
return $\mu$;
```

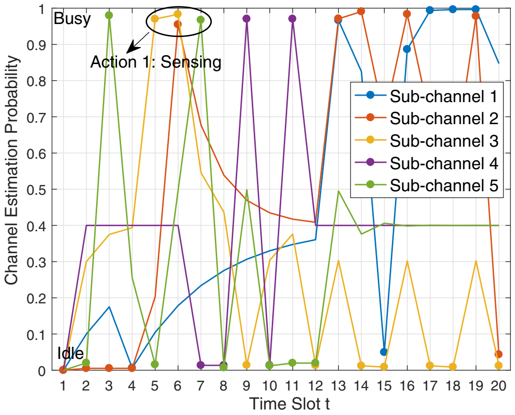


Fig. 3. Channel state estimation probability with SU's first-step action decisions during the first 20 time slots (first-step action: sensing $M=2$ sub-channels).

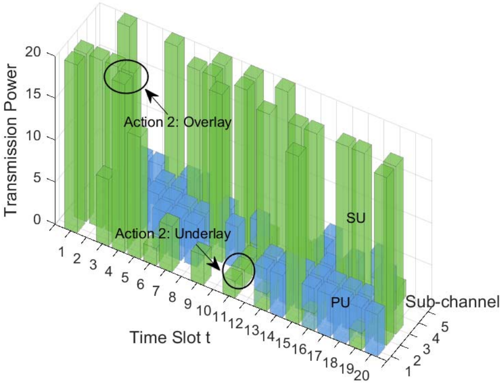


Fig. 4. Transmission power of SU with its second-step action decisions during the first 20 time slots (second-step action: selecting the accessing scheme).

IV. Numerical Simulations

In our simulations, we assume that the radar and communication co-system contains five sub-channels, i.e. $N=5$. The SU is capable of sensing and accessing at most two channels in each time slot, and we have $M=2$ as well as $L=2$. Five sub-channels have the same utilization rate $p_i^1 = 40\%, i\in\{1,2,\ldots 5\}$, while with different transition probability of $\alpha = [15\%,\ 30\%,\ 45\%,\ 60\%,\ 75\%]$ and of $\beta = [10\%,\ 20\%,\ 30\%,\ 40\%,\ 50\%]$, respectively. The false-alarm rate of SUs is $\zeta_f = 2\%$ and their miss-detection rate is $\zeta_m = 2\%$, respectively. The bandwidth of each sub-channel is $W_i = 10$ MHz with radar power spectral density $N_i = 5\times10^{-7}$ W/Hz as well as noise power spectral density $N_0 = 1\times10^{-7}$ W/Hz. The maximum SN's transmission power is $P^u_{\max} = 2$ W for underlay access scheme, while $P^o_{\max} = 20$ W for overlay access scheme. Furthermore, we set the power gain $g_{sr} = g_{pr} = g_{sp} = 1$. Let the weight coefficients $\lambda_C = 1.15 \times 10^{-7}$ (b/s)$^{-1}$ and $\lambda_I = 5$. The discount factor $\gamma = 0.8$. Hence, according to Eq. (4), if the SU accesses an idle sub-channel in the underlay scheme with transmission power $P^u_{\max}$, its reward will be around 2, while its reward will be 5 in terms of the overlay scheme with transmission power $P^o_{\max}$. By contrast, if accessing a busy sub-channel, its reward is about 0.5 and -15 for the underlay scheme and the overlay scheme with maximum transmission power, respectively.

In order to verify the feasibility of our proposed network association mechanism, a numerical simulation spanning over 100 time slots is conducted. Here, we generate $X = 5000$ samples to optimize the coefficient μ in Algorithm 1. Fig. 3 shows the result of the channel state estimation probability as well as SU's first-step action decisions during the first 20 time slots, where the dot represents the number of sub-channel that is selected by the SU for sensing. Moreover, a large value of the channel state estimation probability indicates a high possibility of the sub-channel being busy. Fig. 4 shows the result of the final transmission power of the SU as well as its second-step action decisions during the first 20 time slots. As for the underlay access scheme, SU accesses the whole channel shared with the PUs with a low and equal transmission power, while the SU can only access $L = 2$ sub-channels when it selects the overlay access scheme. Hence, simulation results have verified the feasibility of our proposed network association mechanism for radar and communication co-systems.

V. Conclusions

In this paper, we have constructed a learning assisted network association mechanism for communication and radar co-systems. Firstly, we formulated the co-system as a POMDP problem and deduced its solution, which was suitable for the channel characteristics of the co-system. Moreover, we conceived a low-complexity algorithm for solving its near-optimal policy. Finally, extensive simulations were conducted to evaluate the feasibility of our proposed POMDP formulation.

Acknowledgment

This work was funded by Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation, and also supported by the Shenzhen Science and Technology Program (Grant No. JCYJ20150831192224146). Moreover, L. Hanzo would like to acknowledge the financial support of the European Research Council's Advanced Fellow Grant.

References

[1] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 79–89, May 2007.

[2] C. Jiang, Y. Chen, K. R. Liu, and Y. Ren, "Renewal-theoretical dynamic spectrum access in cognitive radio network with unknown primary behavior," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 3, pp. 406–416.

[3] J. Liu, H. Ding, Y. Cai, H. Yue, Y. Fang, and S. Chen, "An energy-efficient strategy for secondary users in cooperative cognitive radio networks for green communications," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3195–3207, Dec. 2016.

[4] C. Jiang, Y. Chen, Y. Gao, and K. R. Liu, "Joint spectrum sensing and access evolutionary game in cognitive radio networks," *IEEE transactions on wireless communications*, vol. 12, no. 5, pp. 2470–2483, May 2013.

[5] A. Turlapaty and Y. Jin, "A joint design of transmit waveforms for radar and communications systems in coexistence," in *IEEE Radar Conference*, Cincinnati, OH, Aug. 2014, pp. 0315–0319.

[6] B. Li, H. Kumar, and A. P. Petropulu, "A joint design approach for spectrum sharing between radar and communication systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 3306–3310.

[7] Z. Geng, H. Deng, and B. Himed, "Adaptive radar beamforming for interference mitigation in radar-wireless spectrum sharing," *IEEE Signal Processing Letters*, vol. 22, no. 4, pp. 484–488, Otc. 2015.

[8] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge: MIT press, 1998.

[9] J. Wang, C. Jiang, Z. Han, Y. Ren, and L. Hanzo, "Network association strategies for an energy harvesting aided super-WiFi network relying on measured solar activity," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3785–3797, Dec. 2016.

[10] M. Hirzallah, W. Afifi, and M. Krunz, "Full-duplex-based rate/mode adaptation strategies for Wi-Fi/LTE-U coexistence: A POMDP approach," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 1, pp. 20–29, Nov. 2017.

[11] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[12] S. Senthuran, A. Anpalagan, and O. Das, "Throughput analysis of opportunistic access strategies in hybrid underlay-overlay cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2024–2035, Apr. 2012.

[13] J. Zou, H. Xiong, D. Wang, and C. W. Chen, "Optimal power allocation for hybrid overlay/underlay spectrum sharing in multiband cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 4, pp. 1827–1837, Dec. 2013.

[14] E. J. Sondik, "The optimal control of partially observable markov processes over the infinite horizon: Discounted costs," *Operations Research*, vol. 26, no. 2, pp. 282–304, Mar. 1978.