Optimal Age-Energy Tradeoff in Opportunistic Access with Spectrum Handoff

Junyu Li*, Yan Lin*, Yuan-Hsun Lo†, and Yijin Zhang*
*School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China
†Department of Applied Mathematics, National Pingtung University, Pingtung, Taiwan
Email: junyuli@njust.edu.cn, yanlin@njust.edu.cn, yhlo0830@gmail.com, yijin.zhang@gmail.com.

***Abstract*—This paper considers a multichannel cognitive radio network where a secondary user (SU) aims to perform efficient spectrum handoff to achieve a good age of information (AoI)-energy tradeoff. Using the knowledge of channel dynamics and sensing outcomes, the SU needs to determine when and how to switch and when to transmit. We formulate such a spectrum handoff problem under narrow-band sensing as an infinite-horizon partially observable Markov decision process (POMDP) for obtaining optimal policies. Further, we establish the existence of structured optimal decision rules for reducing the computational complexity of optimal policies. Simulation results verify our theoretical findings and demonstrate the performance advantage of the proposed scheme over other schemes.**

I. Introduction

To address the spectrum scarcity problem, the *cognitive radio* (CR) technology [1] has enabled *secondary users* (SUs) to opportunistically access the spectrum holes not being used by *primary users* (PUs). In multichannel CR networks, SUs needs to perform efficient spectrum handoff [2] to determine switch to which channel and whether to transmit, so that SUs can enjoy good access performance with low energy cost. For example, staying at the current channel would lead to performance loss if this channel is unavailable, but switching to another channel would lead to additional energy consumption [3]. Unlike traditional access performance metrics (e.g. throughput, delay), the *age of information* (AoI) [4] was recently proposed to measure the information freshness in realtime *Internet of things* (IoT) applications. So, it is desirable to design an access scheme for SUs to achieve a good AoI-energy tradeoff by an efficient utilization of the knowledge of channel dynamics and sensing outcomes.

To minimize expected total cost (i.e., the sum of transmission cost, handoff cost, and overtime penalty) under wide-band sensing in a deadline-constrained scenario, [5] proposed an optimal policy built on the theory of Markov decision process (MDP) and proved the threshold structure of this policy for a convex penalty function. By additionally taking into account practical switching operations in [5], [6] proposed an MDP framework to obtain an optimal policy and proved some monotone optimal decision rules. For imperfect narrow-band sensing, [7] developed a partially observable Markov decision process (POMDP) framework to obtain an optimal policy that maximizes a predefined throughput-energy tradeoff and used the structure of this optimal policy to propose a low-complexity policy. Further, to achieve a good predefined delay-energy tradeoff under unknown statistics of PUs activity, [8] developed a reinforcement learning framework that allows SUs to adapt their schemes by learning from their past interactions with the environment. In a single-channel scenario, to minimize the average AoI of an energy harvesting SU, [9] used the theory of POMDP to obtain optimal policies under both perfect and imperfect sensing. Still in a single-channel scenario, to maximize a predefined AoI-energy tradeoff of an SU under an average AoI constraint of a PU, [10] obtained an asymptotically optimal policy based on the theory of constrained MDP.

However, none of these previous studies [5]–[10] considered AoI-energy tradeoff of SUs in multichannel CR networks. This paper aims to fill this gap under narrow-band sensing. The main contribution of our work lies in the following three aspects. First, we formulate the spectrum handoff problem in multichannel CR networks with AoI-energy tradeoff as an infinite-horizon POMDP for obtaining optimal policies. Second, we establish the existence of structured optimal decision rules for reducing the computational complexity of optimal policies. Third, through simulations, we verify our theoretical findings and demonstrate the performance advantage of the proposed scheme over other schemes.

Although the idea of using POMDP in the context of spectrum handoff is not new, our study is different because the consideration of AoI-energy tradeoff leads to new theoretical model properties. The method in proving structural results is similar in some aspects to that used in [5], [7], but our proof is more complicated due to the joint impact of narrow-band sensing and the evolution of AoI on the POMDP formulation. Due to the page limit, we have moved some of the technical proofs into our technical report [12].

II. System Model

A. Network Model

Consider a time-slotted CR networks with global synchronization, consisting of an SU and N PUs. Each PU is assigned a licensed channel and the index set of all such channels are denote by $\mathcal{N} \triangleq \{1, 2, \cdots, N\}$. Assume the SU generates a one-slot packet for status update at the beginning of each slot t with the probability λ, and will keep the most recent packet if it is successfully transmitted or is replaced by a fresher one.

The occupancy state of each channel $n \in \mathcal{N}$ at slot t is represented by $\theta_n[t] \in \{0(busy), 1(idle)\}$. Assume that $\theta_n[t]$

for each $n \in \mathcal{N}$ evolves independently from each other and can be modeled by a discrete-time Markov chain. Let

$$P_n \triangleq \begin{bmatrix} p_{n,00} & p_{n,01} \\ p_{n,10} & p_{n,11} \end{bmatrix}. \tag{1}$$

denote the transition probability matrix of channel n. The states of all the channels at slot t is denoted by $\theta[t] \triangleq (\theta_1[t], \theta_2[t], \cdots, \theta_N[t]) \in \{0,1\}^N$.

At the beginning of each slot t, the SU chooses a channel $m[t] \in \mathcal{N}$ and employs a narrow-band sensing technique to obtain the status of this channel, denoted by $o_{m[t]}[t] \in \{0,1\}$. Assume $o_n[t] = \varnothing$ if $n \neq m[t]$. Thus, the observation vector at slot t can be defined by $\boldsymbol{o}[t] \triangleq (\varnothing, \cdots, \varnothing, o_{m[t]}[t], \varnothing, \cdots, \varnothing) \in \mathcal{O}$, where $\mathcal{O}$ denotes the set of possible observation results. Then, the SU will determine whether to transmit on channel $m[t]$ if $o_{m[t]}[t] = 1$, and keep silent otherwise.

Denote the energy consumption for channel sensing, keeping silent, channel switching, and transmitting during a slot by E_{sen}, E_{sil}, E_{trs}, E_{swi}, respectively. Denote the total energy consumption at slot t by E_t.

B. Age of Information

Denote the local age and AoI of the SU at slot t by $x[t]$ and $y[t]$, respectively. Their evolutions can be expressed as

$$x[t+1] = \begin{cases} 0, & \text{if a packet arrives at slot } t, \\ x[t]+1, & \text{otherwise,} \end{cases} \tag{2}$$

$$y[t+1] = \begin{cases} x[t]+1 & \text{if a packet is successfully} \\ & \text{transmitted at slot } t, \\ y[t]+1, & \text{otherwise.} \end{cases} \tag{3}$$

Note that the SU has a packet to send at the beginning of slot t if and only if $x[t] < y[t]$. Following [10], considering that there is no need to count expired age, we assume that $x[t]$ and $y[t]$ are upper bounded by a sufficiently large integer δ such that $x[t], y[t] \in \mathcal{I}_\delta \triangleq \{0, 1, \cdots, \delta\}$.

C. Design Goal

At the beginning of each slot t, the SU makes an access decision for spectrum handoff, i.e., whether to stay at the current channel, switches to which channel, and whether to transmit, based on all the past available information. We aim to minimize the following long-term average cost

$$R \triangleq \limsup_{T\to\infty} \sum_{t=1}^{T} \frac{y[t+1] + \eta E[t]}{T} \tag{4}$$

where η is a tradeoff factor between the energy cost and AoI.

III. Optimal Access Policy

In this section, we formulate the problem specified in Section II as an infinite–horizon POMDP for obtaining optimal policies.

A. POMDP Formulation

The components of our POMDP formulation are described as follows.

States: At the beginning of each slot t, the state is defined as $s[t] \triangleq (x[t], y[t], m[t-1]) \in \mathcal{S}[t]$, where $\mathcal{S}[t]$ denotes the set of possible states in slot t. Thus, it can be included that the state space $\mathcal{S} \triangleq \bigcup_t \mathcal{S}[t] = \{(x,y,m) | 0 \le x \le y \le \delta, m \in \mathcal{N}\}$. Note that $|\mathcal{S}| = (\delta+1)(\delta+2) \cdot N/2$.

Actions: Given the current state $s[t] \in \mathcal{S}$, after sensing the selected channel $m[t] \in \mathcal{N}$, the SU wants to decide whether to transmit if it has a backlogged packet (i.e., $x[t] < y[t]$) and the sensed channel is idle (i.e., $o_{m[t]}[t] = 1$). So, we use $u[t] \in \mathcal{U}_{s[t],o_{m[t]}[t]}$ to represent the transmission action in slot t. Thus, it can be included that the transmission action space $\mathcal{U}_{s,o} \triangleq \bigcup_t \mathcal{U}_{s[t],o_{m[t]}[t]}$, where

$$\mathcal{U}_{s,o} \triangleq \begin{cases} \{0,1\}, & \text{if } x < y, o = 1, \\ \{0\}, & \text{otherwise.} \end{cases} \tag{5}$$

Here the SU transmits if $u[t] = 1$ and keeps silent otherwise. Then the action in slot t is defined by $a[t] \triangleq (m[t], u[t]) \in \mathcal{A}_{s,o} \triangleq \mathcal{N} \times \mathcal{U}_{s,o}$.

Observations and Beliefs: After channel switching and sensing at slot t, the observation of the SU is $\boldsymbol{o}[t]$ defined in Section II. The past information available to the SU before performing action $a[t]$ can be represented as

$$I[t] \triangleq (\boldsymbol{o}[0:t-1], a[0:t-1]). \tag{6}$$

As $\theta[t]$ is observed partially via narrow-band sensing, we define the the belief state as $b[t] \triangleq (b_1[t], b_2[t], \ldots, b_N[t]) \in \mathcal{B} \triangleq [0,1]^N$, where $b_n[t] \triangleq \Pr(\theta_n[t] = 1 \mid I[t])$ for each $n \in \mathcal{N}$. By the theory of POMDP, the update of the belief state depends only on the observation from the previous time slot and the underlying Markov chain. Thus, we have

$$b_n[t+1] = \begin{cases} p_{n,01}, & \text{if } o_n[t] = 0, \\ p_{n,11}, & \text{if } o_n[t] = 1, \\ b_n[t]p_{n,11} + (1-b_n[t])p_{n,01}, & \text{otherwise.} \end{cases} \tag{7}$$

We denote the system state in slot t by $(s[t], b[t]) \in \mathcal{S} \times \mathcal{B}$.

State Transition Function: Given current system state $(s[t], b[t]) = (s, b) = ((x, y, m), b) \in \mathcal{S} \times \mathcal{B}$, observation $\boldsymbol{o}[t] = \boldsymbol{o} \in \mathcal{O}$ and action $a[t] = (m', u) \in \mathcal{A}_{s,o}$, the transition probability to the system state $(s[t+1], b[t+1]) = (s', b') = ((x', y', m'), b') \in \mathcal{S} \times \mathcal{B}$ can be obtained by

$$\begin{aligned} &\beta_{(s',b'),(s,b),(m',u,\boldsymbol{o})} \triangleq \Pr(x[t+1] = x', y[t+1] = y', \\ &m[t] = m', b[t+1] = b' \mid (x[t] = x, y[t] = y, m[t-1] = m, \\ &b[t] = b), a[t] = (m', u), \boldsymbol{o}[t] = \boldsymbol{o}) \\ &= \Pr(x[t+1] = x' \mid x[t] = x, y[t] = y, u[t] = u) \cdot \\ &\Pr(y[t+1] = y' \mid y[t] = y, x[t] = x, u[t] = u) \cdot \\ &\Pr(m[t] = m' \mid m[t-1] = m, m[t] = m') \cdot \\ &\Pr(b[t+1] = b' \mid b[t] = b, m[t] = m', u[t] = u, \boldsymbol{o}[t] = \boldsymbol{o}). \end{aligned} \tag{8}$$

where $\Pr(m[t]=m' \mid m[t-1]=m, m[t]=m')=1$ always holds, $\Pr(b[t+1]=b' \mid (b[t]=b, m[t]=m', u[t]=u, \boldsymbol{o}[t]=\boldsymbol{o})=1$ when the belief updates itself by (7), and

$$\Pr(x[t+1]=x' \mid x[t]=x, y[t]=y, u[t]=u) = \begin{cases} \lambda, & \text{if } x<y, u=0, x'=0 \\ & \text{or } x<y, u=1, x'=0, \\ & \text{or } x=y, u=0, x'=0, \\ 1-\lambda, & \text{if } x<y, u=0, x'=x+1, \\ & \text{or } x<y, u=1, x'=x+1, \\ & \text{or } x=y, u=0, x'=y+1, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

$$\Pr(y[t+1]=y' \mid y[t]=y, x[t]=x, u[t]=u) = \begin{cases} 1, & \text{if } u=0, y'=y+1, \\ & \text{or } u=1, y'=x+1, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Cost Function: We define the cost at the state $(s[t], b[t]) = (s,b) = ((x,y,m),b) \in \mathcal{S}\times\mathcal{B}$ with the observation $o_{m'}[t] = o \in \{0,1\}$ and the action $a[t]=a=(m',u) \in \mathcal{A}_{s,o}$ as

$$C_{s,b,a} \triangleq C^{\triangle}_{(x,y,m),b,m'} + C^{\triangledown}_{(x,y,m),b,u}, \tag{11}$$

where

$$C^{\triangle}_{(x,y,m),b,m'} \triangleq \begin{cases} \eta E_{sen}, & \text{if } m'=m, \\ \eta(E_{sen}+E_{swi}), & \text{otherwise,} \end{cases} \tag{12}$$

$$C^{\triangledown}_{(x,y,m),b,u} \triangleq \begin{cases} y+1+\eta E_{sil}, & \text{if } u=0, \\ x+1+\eta E_{trs}, & \text{otherwise.} \end{cases} \tag{13}$$

Policy: During each slot t, the SU determines $a[t]$ based on a decision function $\pi[t] : \mathcal{S}\times\mathcal{B}\times\mathcal{O} \to \mathcal{A}_{s,o}$. The policy is described by a sequence of decision functions: $\pi \triangleq (\pi[1], \pi[2], \ldots)$. Let Π denote the set of all possible such policies.

B. POMDP Solution

For an initiate system state $(s[0], b[0]) = (s_0, b_0)$, the long-term average cost under a policy $\pi \in \Pi$ is defined by

$$R^{\pi}_{\infty}(s_0,b_0) \triangleq \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^{\pi}\left[\sum_{t=1}^{T} C_{s[t],b[t],a[t]} \,\middle|\, s_0, b_0\right]. \tag{14}$$

We make the following definitions for seeking an optimal policy π^* that minimizes (14).

The relative function is defined as

$$G(s,b) \triangleq V(s,b) - V(s^{\diamond}, b^{\diamond}), \tag{15}$$

where $(s^{\diamond}, b^{\diamond})$ is a reference state, and $V(s,b)$ is computed by

$$V(s,b) = \min_{m'\in\mathcal{N}} Q^{\triangle}_{m'}(s,b). \tag{16}$$

Here $Q^{\triangle}_{m'}(s,b)$ represents the Q-function for taking the sensing action $m' \in \mathcal{N}$ under the state $(s,b) \in \mathcal{S}\times\mathcal{B}$, which is defined by

$$\begin{aligned} Q^{\triangle}_{m'}(s,b) \triangleq & C^{\triangle}_{(x,y,m),b,m'} + (1-b_{m'}) Q^{\triangledown}_0(s,b,0) \\ & + b_{m'} \min_{u\in\mathcal{U}_{s,1}} Q^{\triangledown}_u(s,b,1), \end{aligned} \tag{17}$$

where

$$Q^{\triangledown}_u(s,b,o) \triangleq C^{\triangledown}_{(x,y,m),b,u} + \sum_{s'\in\mathcal{S}, b'\in\mathcal{B}} \beta_{(s',b'),(s,b),(m',u,\boldsymbol{o})} \cdot G(s',b'), \tag{18}$$

represents the Q-function for taking the transmission action $u \in \mathcal{U}_{s,o}$ under the state $(s,b) \in \mathcal{S}\times\mathcal{B}$ and the sensed channel observation $o \in \{0,1\}$.

Based on the above definitions, we summarize the monotonicity property of $G(s,b)$ in Lemmas 1–3. Their proofs are provided in Appendices A–C, respectively.

Lemma 1. *$G((x,y,m),b)$ is nondecreasing in $y \in \mathcal{I}_{\delta}$ for each $x \in \mathcal{I}_{\delta}$, $m \in \mathcal{N}$, $b \in \mathcal{B}$.*

Lemma 2. *$G((x,y,m),b)$ is nondecreasing in $x \in \mathcal{I}_{\delta}$ for each $y \in \mathcal{I}_{\delta}$, $m \in \mathcal{N}$, $b \in \mathcal{B}$.*

Lemma 3. If $G(s,b) : \mathcal{S}\times\mathcal{B} \to \mathbb{R}$ is convex in b, the function $Q^{\triangle}_{m'} : \mathcal{S}\times\mathcal{B} \to \mathbb{R}$, $\forall m' \in \{1,2,\cdots,N\}$ is convex decreasing in b.

By Lemmas 1–3, we prove in Theorem 1 that an optimal policy π^* exists based on [12, Th. 4.2]. The proof of Theorem 1 is provided in Appendix D.

Theorem 1. *When $p_{n,11} \neq p_{n,01}$ for each $n \in \mathcal{N}$, there exists $(R^*, G(s,b))$ that satisfies the following Bellman equation*

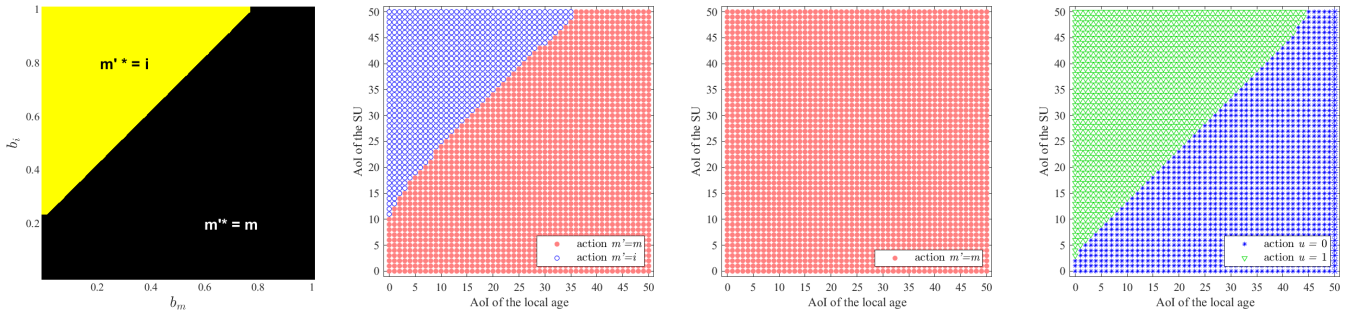
$$R^* + G(s,b) = \min_{m'\in\mathcal{N}} Q^{\triangle}_{m'}(s,b), \quad \forall (s,b) \in \mathcal{S}\times\mathcal{B}, \tag{19}$$

where $R^* \triangleq \min_{\pi\in\Pi} R^{\pi}_{\infty}(s_0,b_0)$. An optimal policy π^* exists and is obtained by

$$m'^*(s,b) \in \operatorname*{argmin}_{m'\in\mathcal{N}} Q^{\triangle}_{m'}(s,b), \tag{20}$$

$$u^*(s,b,o) \in \operatorname*{argmin}_{u\in\mathcal{U}_{s,o}} Q^{\triangledown}_u(s,b,o). \tag{21}$$

Applying the relative value iteration algorithm [15], which is summarized in Algorithm 1, to solve (19) can lead to π^*. Following [10], Algorithm 1 uses all the initial and intermediate beliefs evolved into the stationary distribution to form a simplified belief space $\tilde{\mathcal{B}}$. Typically, during each iteration step of the Algorithm 1, the required number of computations includes $(2N+3)|\mathcal{S}||\tilde{\mathcal{B}}|$ multiplications, $(2N+3)|\mathcal{S}||\tilde{\mathcal{B}}|$ additions and $|\mathcal{S}||\tilde{\mathcal{B}}|$ subtractions. Hence, the computational complexity for each iteration is $O(N|\mathcal{S}||\tilde{\mathcal{B}}|)$. If Algorithm 1 is executed for a finite number of M iterations, the computational complexity of Algorithm 1 is $O(MN|\mathcal{S}||\tilde{\mathcal{B}}|)$.



(a) (b) (c) (d)

Fig. 1: (a) case 1 in Theorem 2 for $x = 5$, $y = 30$; (b) case 2 in Theorem 2 for $b_m = 0.3$, $b_i = 0.8$; (c) case 3 in Theorem 2 for $b_m = 0.5$, $b_i = 0.5$; (d) case 4 in Theorem 2

Algorithm 1 Optimal Policy design via Relative Value Iteration Algorithm

Input: The AoI truncation upper bound value δ; the packet generation probability λ; the iteration termination criterion ϵ; the tradeoff factor η

Initialize: For all $(s,b) \in \mathcal{S}\times\tilde{\mathcal{B}}$, initialize $V_0(s,b) = 0$, choose $(s^\diamond, b^\diamond)$, set $G_0(s,b) = V_0(s,b) - V_0(s^\diamond, b^\diamond)$ and $k = 0$.

Output: An optimal policy π^*

1: **while** $\exists (s,b) \in \mathcal{S}\times\tilde{\mathcal{B}}$ s.t.$|G_{k+1}(s,b) - G_k(s,b)| > \epsilon$ **do**
2: $k = k+1$
3: **for** all $(s,b) \in \mathcal{S}\times\tilde{\mathcal{B}}$ **do**
4: compute $V_{k+1}(s,b)$ by (16)–(18) using $G_k(s,b)$,
5: set $G_{k+1}(s,b) = V_{k+1}(s,b) - V_{k+1}(s^\diamond, b^\diamond)$.
6: **end for**
7: **end while**
8: $R^* = V(s^\diamond, b^\diamond)$.
9: An optimal policy is obtained by (20) and (21).
10: **return** An optimal policy π^*.

IV. Structured Optimal Decision Rules

In this section, we establish structured optimal decision rules to reduce the computational complexity for obtaining an optimal policy π^*.

To prove the existence of the threshold structure of the action m' and u, we need to leverage on the definitions of superadditivity and subadditivity [14]. Let $g(z,v)$ be a real-valued function on $\mathcal{Z}\times\mathcal{V}$ where $\mathcal{Z}$ and $\mathcal{V}$ represent partially ordered sets. Let $z_1, z_0 \in \mathcal{Z}$ and $v_1, v_0 \in \mathcal{V}$. Then $g(z,v)$ is superadditive if

$$g(z_1,v_1) + g(z_0,v_0) \geq g(z_1,v_0) + g(z_0,v_1), \tag{22}$$

for $z_1 \geq z_0$ in $\mathcal{Z}$ and $v_1 \geq v_0$ in $\mathcal{V}$. Conversely, $g(z,v)$ is subadditive if the inequality above holds in the opposite direction. Then, we need the following properties to prove the threshold structure. The proof of Lemmas 5–6 are provided in Appendices E–F, respectively.

Lemma 4 [14, Lemma 4.7.1, Ch. 4]**.** *If $g(z,v)$ is a superadditive function on $\mathcal{Z}\times\mathcal{V}$, and $\max_{v\in V} g(z,v)$ exists for all $z \in \mathcal{Z}$, $f(z) \triangleq \max\{\arg\max_{v\in V} g(z,v)\}$ is monotone nondecreasing in z.*

Lemma 5. *$Q_u^\triangledown((x,y,m)),b,1)$ is subadditive on $(y,u) \in \mathcal{I}_\delta \times \mathcal{U}_{s,1}$ for each $x \in \mathcal{I}_\delta$, $m \in \mathcal{N}$, $b \in \mathcal{B}$.*

Lemma 6. *Define a partial order in $\mathcal{N}$:$(m'=i) > (m'=m)$ for arbitrary $i \in \mathcal{N}\setminus\{m\}$. Then, if $b_i > b_m$, $Q_{m'}^\triangle((x,y,m),b)$ is subadditive on $(y,m') \in \mathcal{I}_\delta \times \mathcal{N}$ for each $x \in \mathcal{I}_\delta$, $m \in \mathcal{N}$, $b \in \mathcal{B}$.*

Now, we establish conditions that ensure the optimality of structured decision rules by applying Lemmas 1–6.

Theorem 2. *Given an arbitrary current system state $(s,b) = ((x,y,m),b) \in \mathcal{S}\times\mathcal{B}$, there exhibits the structures in the following cases.*

1) *If $m'^*(s,b) = n \in \mathcal{N}$, then for all $0 \leq \nu \leq 1 - b_n$, $m'^*(s, b+\nu\boldsymbol{e}_n) = n$, where $\boldsymbol{e}_n$ denotes the unit vector with 1 in the nth position and 0s elsewhere,*
2) *When there exists a channel i with a higher probability of being idle than the current channel m, i.e., $\exists i \in \mathcal{N}\setminus\{m\}$, $b_i > b_m$, there exists a threshold $Y^{th1}(x,m,b) \in [x,\delta]$ such that we have*

$$m'^*(s,b) = \begin{cases} m, & \text{if } y < Y^{th1}(x,m,b), \\ i, & \text{otherwise.} \end{cases} \tag{23}$$

The threshold $Y^{th1}(x,m,b)$ can be obtained from lines 20 to 35 in Algorithm 2.

3) *When current channel m is the best channel, i.e., $b_i \leq b_m$, $i \in \mathcal{N}\setminus\{m\}$, we have*

$$m'^*(s,b) = m. \tag{24}$$

4) *When the sensing channel m' is idle, i.e., $o = 1$, there exists a threshold $Y^{th2}(x,m,b) \in [x,\delta]$ such that we have*

$$u^*(s,b,1) = \begin{cases} 0, & \text{if } y < Y^{th2}(x,m,b), \\ 1, & \text{otherwise.} \end{cases} \tag{25}$$

The threshold $Y^{th2}(x,m,b)$ can be obtained from lines 7 to 15 in Algorithm 2.

The proof of Theorem 2 is given in Appendix G.

Algorithm 2 Relative Value Iteration Based on the Threshold Structure

Input: The AoI truncation upper bound value δ; the packet arrival probability λ; the iteration termination criterion ϵ; the tradeoff factor η

Initialize: For all $(s,b) \in \mathcal{S} \times \tilde{\mathcal{B}}$, initialize $V_0(s,b) = 0$, choose $(s^\diamond, b^\diamond)$, set $G_0(s,b) = V_0(s,b) - V_0(s^\diamond, b^\diamond)$ and $k = 0$.

Output: An optimal policy π^*

while $\exists (s,b) \in \mathcal{S} \times \tilde{\mathcal{B}}$ s.t.$|G_{k+1}(s,b) - G_k(s,b)| > \epsilon$ **do**
  $k = k + 1$
  for $m \in \mathcal{N}$ and $b \in \tilde{\mathcal{B}}$ **do**
    $y = 0$
    while $y \leq \delta$ **do**
      for $x = 0$ to y **do**
        compute $Q^{\triangledown}_{u,k+1}(s,b)$ by (18)
        $u^*(s,b,o) \in \operatorname{argmin}_{u \in \mathcal{U}[t]} Q^{\triangledown}_{u,k+1}(s,b,o)$
        if $u^*(s,b,o) = 1$ **then**
          $Y^{th2}(x,m,b) = y$
          for $y = Y^{th2}(x,m,b) + 1$ to δ **do**
            $\min_{u \in \mathcal{U}_{s,o}} Q^{\triangledown}_{u,k+1}(s,b,o) = Q^{\triangledown}_{1,k+1}(s,b,o)$
            $u^*(s,b,o) = 1$
          end for
        end if
        if $b_i \leq b_m$ for each $i \in \mathcal{N} \setminus \{m\}$ **then**
          $m'^*(s,b) = m$
          $V_{k+1}(s,b) = Q^{\triangle}_{m,k+1}(s,b)$
        else
          compute $Q^{\triangle}_{m',k+1}(s,b)$ by (17)
          $m'^*(s,b) \in \operatorname{argmin}_{m' \in \mathcal{N}} Q^{\triangle}_{m',k+1}(s,b)$
          if $m'^*(s,b) = i$, $i \in \mathcal{N} \setminus \{m\}$ **then**
            $Y^{th1}(x,m,b) = y$
            for $y = Y^{th1}(x,m,b) + 1$ to δ **do**
              for $\nu = 0$ to $1 - b_i$ **do**
                $m'^*(s, b + \nu \boldsymbol{e}_i) = i$
                $V_{k+1}(s,b) = Q^{\triangle}_{i,k+1}(s,b)$
              end for
            end for
          else if $m'^*(s,b) = m$
            for $\nu = 0$ to $1 - b_m$ **do**
              $m'^*(s, b + \nu \boldsymbol{e}_m) = m$
              $V_{k+1}(s,b) = Q^{\triangle}_{m,k+1}(s,b)$
            end for
          end if
        end if
        $y = y + 1$
      end for
    end while
  end for
  set $G_{k+1}(s,b) = V_{k+1}(s,b) - V_{k+1}(s^\diamond, b^\diamond)$
end while
return An optimal policy π^*

According to Theorem 2, we know that there exists a threshold structure for decision rules in some particular states, which can be used to eliminate the necessity of minimizing (17) and (18) for all the states, thereby reducing the computational complexity of Algorithm 1. The detailed procedure is presented in Algorithm 2.

V. Results

In this section, we conduct simulations to verify the structural results as proved in Theorem 2 and compare the proposed policy against an optimal MDP-based policy (i.e., an optimal policy when all the channel states are always known to the SU), a greedy policy (i.e., the SU always selects an action to minimize the current cost), and an AoI-optimal policy (i.e., a policy that minimizes the long-term average AoI).

Unless otherwise specified, we consider the following default settings: $N = 3$ channels with $P_n = [0.2, 0.8; 0.8, 0.2]$ for each channel n, the packet arrival probability $\lambda = 0.5$, the truncated AoI value $\delta = 50$, the energy cost with $E_{sen} = 1$, $E_{sil} = 0.01$, $E_{trs} = 10$ and $E_{swi} = 5$, and the tradeoff factor $\eta = 1$.

A. Structured Property of Optimal Policies

Fig. 1(a)–(d) show the optimal actions obtained from Algorithm 1, which are indeed of the threshold structure as stated in cases 1–4 in Theorem 2, respectively. For case 1, we observe that if the previous optimal sensing action is n and only the idle probability of channel n increases, the current optimal sensing action remains unchanged. The reason is that such a decision leads to a higher probability to reduce the AoI without extra energy consumption. For case 2, we observe that, when $\exists i \in \mathcal{N} \setminus \{m\}$, $b_i > b_m$, the SU decides to sense the channel i if $y \geq Y^{th1}(x,m,b)$. This is because the SU is eager to seek an available channel to reduce the AoI when y is large. For case 3, we observe that the SU would always prefer to sense the current channel when $b_i \leq b_m$, $i \in \mathcal{N} \setminus \{m\}$, since sensing another channel would result in a switching cost without any benefit. For case 4, we observe that the SU chooses to transmit if $y \geq Y^{th2}(x,m,b)$ because the reduction of y caused by a transmission is greater than the transmission energy consumption.

B. Comparisons with Other Polices

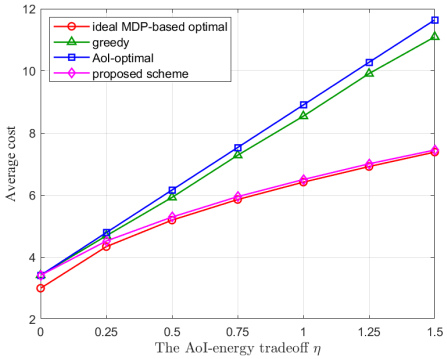


Fig. 2: Average cost versus the tradeoff factor η.

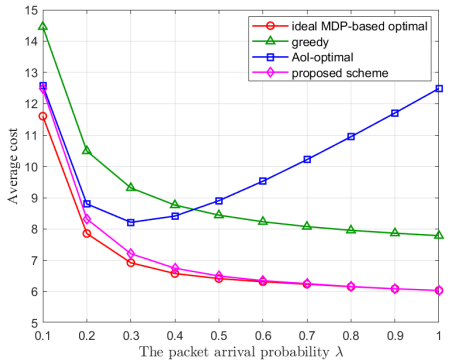


Fig. 3: Average cost versus the packet arrival probability λ.

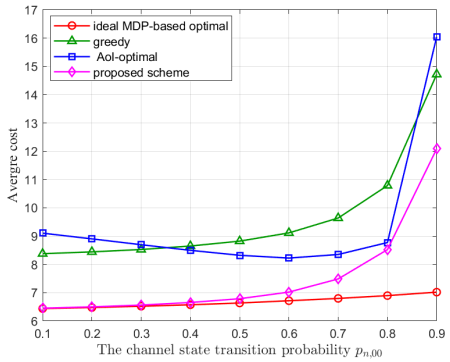


Fig. 4: Average cost versus the channel state transition probability $p_{n,00}$.

Fig. 2 compares the average cost of different policies versus the tradeoff factor η. As expected, we observe that the proposed policy performs worse than the optimal MDP-based policy because of the partial observation. The gap between these two policies narrows with η. This is because the benefit of the complete observation in determining whether to switch becomes weaker when η is larger. We also observe that the proposed policy always performs better than the greedy policy and the gap becomes more noticeable with η. This is because, when η is larger, accounting for the impact of the current action on the future cost is more effective to reduce the long-term cost than merely comparing the current energy cost and the current AoI reduction. We further observe that the proposed policy always performs better than the AoI-optimal policy and the gap becomes more noticeable with η. This is because, when η is larger, the joint considering the AoI and the energy cost is more effective to reduce the total cost than merely considering the AoI.

Fig. 3 compares the average cost of different policies versus the packet arrival probability λ. As expected, we observe that the proposed policy performs worse than the optimal MDP-based policy and the gap narrows with λ. This is because the benefit of complete observation in determining when to transmit becomes weaker when faced with fresher packets. Meanwhile, we observe that the proposed policy always performs better than the greedy policy with an almost unchanging gap. We also observe that the proposed policy always performs better than the AoI-optimal policy and the gap becomes more noticeable with λ. This is because, when λ is larger, the AoI-optimal policy prefers to make more transmissions to reduce the AoI, but the effect of a transmission on the AoI reduction becomes weaker.

Fig. 4 compares the average cost of different policies versus the channel state transition probability $p_{n,00}$. We observe that the proposed policy always performs worse than the optimal MDP-based policy and the gap becomes more noticeable with $p_{n,00}$. This is because the benefit of complete observation in switching to an idle channel becomes higher when channels are more likely to be busy. We also observe that the proposed policy always performs better than the greedy policy and the gap narrows marginally with $p_{n,00}$. This is because the greedy policy is more willing to keep silent than the proposed policy due to its myopic property, and increasing $p_{n,00}$, i.e., reducing the transmission opportunities, has a smaller negative impact on the greedy policy. We further observe that the proposed policy always performs better than the AoI-optimal policy. The gap between these two policies first narrows and then becomes noticeable when $p_{n,00}$ exceeds a certain value. This is because according to (7), under the AoI-optimal policy, a larger $p_{n,00}$ enables the SU to be more willing to stay in the current channel when $p_{n,00} \leq p_{n,10}$, but enables the SU to be more willing to switch to another channel when $p_{n,00} > p_{n,10}$, resulting in a sharp rise in the total cost.

VI. Conclusion

In this paper, we have investigated the spectrum access problem with a focus on the optimal AoI-energy tradeoff. We have formulated such a problem under narrow-band sensing as an infinite-horizon POMDP for obtaining optimal policies and proved the structure of optimal decision rules for reducing the computational complexity. Simulation results verified our theoretical findings and demonstrated the advantage of the proposed scheme over other schemes. Our future work is to extend the single SU case to multiple SUs cases.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62071236, and in part by the National Science and Technology Council, Taiwan, under Grants 112-2115-M-153-004-MY2 and 110-2115-M-110-004-MY3.

References

[1] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal," IEEE Pers. Commun., vol. 6, no. 4, pp. 13–18, Aug. 1999.

[2] I. Christian, S. Moh, I. Chung, and J. Lee, "Spectrum mobility in cognitive radio networks," IEEE Commun. Mag., vol. 50, no. 6, pp. 114– 121, Jun. 2012.

[3] S. Demirci and D. G¨oz¨upek, "Switching cost-aware joint frequency assignment and scheduling for industrial cognitive radio networks," IEEE Trans. Ind. Informat., vol. 16, no. 7, pp. 4365–4377, July 2020.

[4] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in Proc. IEEE INFOCOM, 2012, pp. 2731–2735.

[5] Y. Wu, Q. Yang, X. Liu, and K. S. Kwak, "Delay-constrained optimal transmission with proactive spectrum handoff in cognitive radio networks," *IEEE Trans. Commun*, vol. 64, no. 7, pp. 2767–2779, Jul. 2016.
[6] Z. Xue, A. Gong, Y.-H. Lo, S. Tian, and Y. Zhang, "Deadline-constrained opportunistic spectrum access with spectrum handoff," in *Proc. IEEE Glob. Commun. Conf(GLOBECOM)*, Kuala Lumpur, Malaysia, 2023, pp. 261–266.
[7] M. Santhoshkumar and K. Premkumar, "Energy-efficient opportunistic spectrum access in multichannel cognitive radio networks," *IEEE Netw. Lett.*, vol. 5, no. 1, pp. 1–5, Mar. 2023.
[8] H. Ding, X. Li, Y. Ma, and Y. Fang, "Energy-efficient channel switching in cognitive radio networks: A reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12359–12362, Oct. 2020.
[9] S. Leng and A. Yener, "Age of information minimization for an energy harvesting cognitive radio," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 2, pp. 427–439, Jun. 2019.
[10] Y. Zhao, B. Zhou, W. Saad, and X. Luo, "Age of information analysis for dynamic spectrum sharing," in *Proc. IEEE Glob Conf. Signal Inf. Process. (GlobalSIP)*, Ottawa, ON, Canada, 2019, pp. 1–5.
[11] E. Fernández-Gaucherand, A. Arapostathis, and S. I. Marcus, "On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes," *Ann. Oper. Res.*, vol. 29, no. 1, pp. 439–469, 1991.
[12] J. Li, Y. Lin, Y.-H. Lo, and Y. Zhang, "Optimal age-energy tradeoff in opportunistic access with spectrum handoff," Tech. Rep., 2025. [Online]. Available:https://modestteenager.github.io/lijunyu/reports/age-energy-spectrum-handoff.pdf.
[13] V. Krishnamurthy, *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing.* Cambridge, U.K.: Cambridge Univ. Press, 2016.
[14] D. M. Topkis, *Supermodularity and Complementarity.* Princeton, NJ, USA: Princeton Univ. Press, 1998.
[15] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. I, 3rd ed. Belmont, MA, USA: Athena Sci., 2007.

Supplementary Materials for paper "Optimal Age-Energy Tradeoff in Opportunistic Access with Spectrum Handoff"

Junyu Li, Yan Lin, Yuan-Hsun Lo, and Yijin Zhang

Proof of Lemma 1

When $u=0$, $o=0$, $y=\delta$ or $\delta-1$, by (18), we have

$$Q_0^{\triangledown}((x,\delta,m),b,0)=\delta+1+\lambda G((0,\delta,m'),b')+(1-\lambda)G((x+1,\delta,m'),b'), \tag{26}$$

$$Q_0^{\triangledown}((x,\delta-1,m),b,0)=\delta+\lambda G((0,\delta,m'),b')+(1-\lambda)G((x+1,\delta,m'),b'), \tag{27}$$

which implies

$$Q_0^{\triangledown}((x,\delta-1,m),b,0)\le Q_0^{\triangledown}((x,\delta,m),b,0). \tag{28}$$

Similarly, we have

$$Q_0^{\triangledown}((x,\delta-1,m),b,1)\le Q_0^{\triangledown}((x,\delta,m),b,1), \tag{29}$$

$$Q_1^{\triangledown}((x,\delta-1,m),b,1)\le Q_1^{\triangledown}((x,\delta,m),b,1), \tag{30}$$

which results in

$$Q_{m'}^{\triangle}((x,\delta-1,m),b)\le Q_{m'}^{\triangle}((x,\delta,m),b), \tag{31}$$

$$G((x,\delta-1,m),b)\le G((x,\delta,m),b). \tag{32}$$

When $u=0$, $o=0$, $y=\delta-2$, we have

$$Q_0^{\triangledown}((x,\delta-2,m),b,0)=\delta-1+\lambda G((0,\delta-1,m'),b')+(1-\lambda)G((x+1,\delta-1,m'),b'), \tag{33}$$

Due to (32), we have

$$Q_0^{\triangledown}((x,\delta-2,m),b,0)\le Q_0^{\triangledown}((x,\delta-1,m),b,0), \tag{34}$$

Similarly, we have

$$Q_0^{\triangledown}((x,\delta-2,m),b,1)\le Q_0^{\triangledown}((x,\delta-1,m),b,1), \tag{35}$$

$$Q_1^{\triangledown}((x,\delta-2,m),b,1)\le Q_1^{\triangledown}((x,\delta-1,m),b,1), \tag{36}$$

which results in

$$Q_{m'}^{\triangle}((x,\delta-2,m),b)\le Q_{m'}^{\triangle}((x,\delta-1,m),b), \tag{37}$$

$$G((x,\delta-2,m),b)\le G((x,\delta-1,m),b). \tag{38}$$

Thus, by the mathematical induction on (32) and (38), we obtain that $G((x,y,m),b)$ is nondecreasing in $y\in\mathcal{I}_\delta$.

VII. Proof of Lemma 2

When $y=\delta$, $x=\delta-2$ or $\delta-1$, from Lemma 1, we can derive the following inequality:

$$Q_0^{\triangledown}((\delta-2,\delta,m),b,0)\le Q_0^{\triangledown}((\delta-1,\delta,m),b,0), \tag{39}$$

$$Q_0^{\triangledown}((\delta-2,\delta,m),b,1)\le Q_0^{\triangledown}((\delta-1,\delta,m),b,1), \tag{40}$$

$$Q_1^{\triangledown}((\delta-2,\delta,m),b,1)\le Q_1^{\triangledown}((\delta-1,\delta,m),b,1), \tag{41}$$

which results in

$$Q_{m'}^{\triangle}((\delta-2,\delta,m),b)\le Q_{m'}^{\triangle}((\delta-1,\delta,m),b), \tag{42}$$

$$G((\delta-2,\delta,m),b)\le G((\delta-1,\delta,m),b). \tag{43}$$

Thus, it follows by the mathematical induction on (43) that $G((x,\delta,m),b)$ is nondecreasing in $x\in\mathcal{I}_\delta$.

When $y=\delta-1$, $x=\delta-3$ or $\delta-2$, due to (43), we have

$$Q_0^{\triangledown}((\delta-3,\delta-1,m),b,0)\le Q_0^{\triangledown}((\delta-3,\delta-1,m),b,0), \tag{44}$$

$$Q_0^{\triangledown}((\delta-3,\delta-1,m),b,1)\le Q_0^{\triangledown}((\delta-2,\delta-1,m),b,1), \tag{45}$$

$$Q_1^{\triangledown}((\delta-3,\delta-1,m),b,1) \le Q_1^{\triangledown}((\delta-2,\delta-1,m),b,1), \tag{46}$$

which results in

$$Q_{m'}^{\triangle}((\delta-3,\delta-1,m),b) \le Q_{m'}^{\triangle}((\delta-2,\delta-1,m),b), \tag{47}$$

$$G((\delta-3,\delta-1,m),b) \le G((\delta-2,\delta-1,m),b). \tag{48}$$

Similarly, $G((x,\delta-1,m),b)$ is nondecreasing in $x \in \mathcal{I}_\delta$.

Thus, we conclude inductively that $G((x,y,m),b)$ is nondecreasing in $x \in \mathcal{I}_\delta$.

VIII. Proof of Lemma 3

Due to (17)–(18), the function $Q_{m'}^{\triangle}$ can be expressed by

$$Q_{m'}^{\triangle}((x,y,m),(b_1,b_2,\cdots,b_N)) = C_{(x,y,m),b,a} + \mathbb{E}_{\phi_2(\boldsymbol{o};b)}[G((x',y',m'),({b_1}',{b_2}',\cdots,{b_N}'))], \tag{49}$$

where

$$\begin{aligned} {b_n}' &= \frac{b_n\phi_1(o_n)}{b_n\phi_1(o_n)+(1-b_n)\phi_0(o_n)}(p_{n,11}-p_{n,01})+p_{n,01} \\ &= \frac{b_n\phi_1(o_n)}{\phi_2(o_n,b_n)}(p_{n,11}-p_{n,01})+p_{n,01}. \end{aligned} \tag{50}$$

Here $\phi_2(\boldsymbol{o};b)=\prod_{i=1}^N \phi_2(o_n,b_n)$. Based on $o_n \in \{0,1,\varnothing\}$, we have

$$\phi_0(o_n) = \begin{cases} 0, & \text{if } o_n = 1, \\ 1, & \text{otherwsise,} \end{cases} \tag{51}$$

$$\phi_1(o_n) = \begin{cases} 0, & \text{if } o_n = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{52}$$

If $G(s',b')$ is convex in b', then for a given s', we have

$$G(s',b') = \sup_{(\boldsymbol{w}_\ell,q_\ell)\in J} \{\boldsymbol{w}_\ell^T b' + q_\ell\} \tag{53}$$

where $J \triangleq \{(\boldsymbol{w},q) \in \mathbb{R}^{N+1} : \boldsymbol{w}^T b' + q \le f(b'), b' \in [0,1]^N\}$. Hence, subtracting the constant terms, we have

$$\begin{aligned} &Q_{m'}^{\triangle}(s,b) - C_{(x,y,m),b,a} \\ &= \sum_{\boldsymbol{o}\in\mathcal{O}} G\left(s', \frac{b_1\phi_1(o_1)}{\phi_2(o_1,b_1)}(p_{1,11}-p_{1,01})+p_{1,01}, \ldots, \frac{b_N\phi_1(o_N)}{\phi_2(o_N,b_N)}(p_{N,11}-p_{N,01})+p_{N,01}\right)\prod_{n=1}^N[\phi_2(o_n,b_n)] \\ &= \sum_{\boldsymbol{o}\in\mathcal{O}} \left(\sup_{(\boldsymbol{w}_\ell,q_\ell)\in J}\left\{\boldsymbol{w}_\ell^T b' + q_\ell\right\}\right)\prod_{n=1}^N[\phi_2(o_n,b_n)] \\ &= \sum_{\boldsymbol{o}\in\mathcal{O}} \left(\sup_{(\boldsymbol{w}_\ell,q_\ell)\in J}\left\{q_\ell + \sum_{n=1}^N w_{\ell,n}\cdot\left(\frac{b_n\phi_1(o_n)}{\phi_2(o_n,b_n)}(p_{n,11}-p_{n,01})+p_{n,01}\right)\right\}\right)\prod_{n=1}^N[\phi_2(o_n,b_n)] \\ &= \sup_{(\boldsymbol{w}_\ell,q_\ell)}\left\{\sum_{\boldsymbol{o}\in\mathcal{O}} q_\ell\prod_{n=1}^N[\phi_2(o_n,b_n)] + \sum_{\boldsymbol{o}\in\mathcal{O}}\sum_{n=1}^N w_{\ell,n}\cdot\left(\frac{b_n\phi_1(o_n)}{\phi_2(o_n,b_n)}(p_{n,11}-p_{n,01})+p_{n,01}\right)\prod_{n=1}^N[\phi_2(o_n,b_n)]\right\} \\ &= \sup_{(\boldsymbol{w}_\ell,q_\ell)\in J}\left\{q_\ell + \sum_{n=1}^N w_{\ell,n}\cdot\Big[b_n(p_{n,11}-p_{n,01})+p_{n,01}\Big]\right\}. \end{aligned} \tag{54}$$

which is a superpremum of a collection of affine functions of b. This indicates that $Q_{m'}^{\triangle}(s,b)$ is convex decreasing in $b \in \mathcal{B}$.

IX. Proof of Theorem 1

According to [13, Th 4.2], it suffices to show that the following two conditions are satisfied: (i) $\mathcal{B}$ is a countable set; (ii) there exists a constant $L \geq 0$ such that

$$|V_\alpha(s,b) - V_\alpha(s',b')| \leq L, \forall \alpha \in (0,1), \forall (s,b),(s',b') \in \mathcal{S} \times \mathcal{B}$$

where $V_\alpha(s,b)$ is the value function of the corresponding discounted cost problem with the objective

$$R_\alpha^\pi \triangleq \lim_{T\to\infty} \frac{1}{T} \mathbb{E}^\pi \left[\sum_{t=1}^{T} \alpha^t C_{s[t],b[t],a[t]} \,\middle|\, s_0, b_0 \right], \tag{55}$$

i.e., $V_\alpha(s,b) \triangleq \min_{\pi\in\Pi} R_\alpha^\pi(s,b)$.

When $p_{n,11} \neq p_{n,01}$, we know that the matrix

$$\begin{pmatrix} 1-p_{n,01} & p_{n,01} \\ 1-p_{n,11} & p_{n,11} \end{pmatrix}$$

is nonsingular. Thus, the belief update process is an injective map based on [12, Lemma 4.2], which implies that the condition (i) holds.

Consider a system state $(\overline{s},\overline{b}) = ((\delta-1,\delta,m),(0,0\dots))$. Due to the property of consistent monotonicity between $G(s,b)$ and $V_\alpha(s,b)$ as stated in Lemmas 1–3, we have

$$0 \leq V_\alpha(s,b) \leq V_\alpha(\overline{s},\overline{b}), \quad \forall (s,b) \in \mathcal{S} \times \mathcal{B}.$$

Then it suffices to show that $V_\alpha(\overline{s},\overline{b})$ is not larger than a constant. The Q-function with discount α can be written as follows according to (17) and (18).

$$Q_{m'}^\alpha(\overline{s},\overline{b}) = C^{\triangle}_{(\delta-1,\delta,m),(0,0\dots),m'} + Q_0^\alpha(\overline{s},\overline{b},0), \tag{56}$$

$$Q_0^\alpha(\overline{s},\overline{b},0) = C^{\nabla}_{(\delta-1,\delta,m),(0,0\dots),0} + \alpha \sum_{s'\in\mathcal{S},b'\in\mathcal{B}} \beta_{(s',b'),(\overline{s},\overline{b}),(m',0,\boldsymbol{o})} \cdot V_\alpha(s',b'). \tag{57}$$

Thus, we have

$$V_\alpha(\overline{s},\overline{b}) \leq Q_{m'}^\alpha(\overline{s},\overline{b}) \leq \eta(E_{sen}+E_{swi}+E_{sil}) + \delta + 1 + \alpha V_\alpha(\overline{s},\overline{b}), \tag{58}$$

which implies that $V_\alpha(\overline{s},\overline{b})$ is not larger than a constant

$$L = \frac{\eta(E_{sen}+E_{swi}+E_{sil}) + \delta + 1}{1-\alpha}.$$

This completes the proof of Theorem 1.

X. Proof of Lemma 5

To prove the subadditivity of $Q_u^\nabla((x,y,m),b,1)$ on $(y,u) \in \mathcal{I}_\delta \times \mathcal{U}_{s,1}$, we should prove the inequality

$$\begin{aligned} W_1 \triangleq\ & Q_1^\nabla((x,y+1,m),b,1) - Q_1^\nabla((x,y,m),b,1) \\ & - Q_0^\nabla((x,y+1,m),b,1) + Q_0^\nabla((x,y,m),b,1) \leq 0, \end{aligned} \tag{59}$$

holds for each $x \in \mathcal{I}_\delta$, $y \in \mathcal{I}_\delta$, $m \in \mathcal{N}$, and $b \in \mathcal{B}$.

Due to (18), we obtain that

$$\begin{aligned} W_1 = -1 &+ \lambda \underbrace{\Big[G((0,y+1,m),b) - G((0,y+2,m),b)\Big]}_{H_1} \\ &+ (1-\lambda) \underbrace{\Big[G((x+1,y+1,m),b) - G((x+1,y+2,m),b)\Big]}_{H_2}. \end{aligned} \tag{60}$$

By Lemma 2, $H_1, H_2 \leq 0$, which implies that (59) holds. This completes the proof of Lemma 5.

XI. Proof of Lemma 6

To prove the subadditivity of $Q^{\triangle}_{m'}((x,y,m),b)$ on $(y,m') \in \mathcal{I}_\delta \times \mathcal{N}$, due to the defined partial order in $\mathcal{N}$, we should prove the following inequality for $(m'=i) > (m'=m)$

$$\begin{aligned} W_2 \triangleq & Q^{\triangle}_i((x,y+1,m),b) - Q^{\triangle}_i((x,y,m),b) \\ & - Q^{\triangle}_m((x,y+1,m),b) + Q^{\triangle}_m((x,y,m),b) \le 0, \end{aligned} \tag{61}$$

holds for each $x \in \mathcal{I}_\delta$, $y \in \mathcal{I}_\delta$, $m \in \mathcal{N}$, $i \in \mathcal{N}$ and $b \in \mathcal{B}$. Due to (17) and (18), we can obtain that

$$\begin{aligned} W_2 = & (b_i - b_m)\{Q^{\triangledown}_0((x,y,m),b,1) - Q^{\triangledown}_0((x,y+1,m),b,1) \\ & - \min_{u_1 \in \mathcal{U}_{s,1}} Q^{\triangledown}_u((x,y,m),b,1) + \min_{u_2 \in \mathcal{U}_{s,1}} Q^{\triangledown}_u((x,y+1,m),b,1)\}. \end{aligned} \tag{62}$$

Considering the subadditivity of the $Q^{\triangledown}_u((x,y,m),b,1)$ in Lemma 5, we have $u_1 \le u_2$. Thus, we only need to consider $(u_1,u_2) \in \{(0,0),(0,1),(1,1)\}$. When $(u_1,u_2) = (0,0)$ or $(0,1)$, (61) holds as $b_i > b_m$. When $(u_1,u_2) = (1,1)$, (61) holds by (59). This completes the proof of Lemma 6.

XII. Proof of Theorem 2

We shall apply Lemmas 3–6 to prove the structure in the following four cases.

Case 1: From Lemma 3, (17) and (18), we know that for $s \in \mathcal{S}$ and $m' \in \mathcal{N}$, $Q^{\triangle}_{m'}(s,b)$ is a convex decreasing function in $b \in \mathcal{B}$. Consider a b such that $V(s,b) = Q^{\triangle}_n(s,b)$. Under a given $s \in \mathcal{S}$, the decision region where the optimal action is n is defined as follows.

$$\mathcal{D}_{s,n} \triangleq \{b : V(s,b) \le Q^{\triangle}_{m'}(s,b), \forall m' \ne n\}. \tag{63}$$

Since $\mathcal{D}_{s,n}$ is a convex set, we know $m'^*(s,b) = n$, which implies that $m'^*(s, b+\nu \boldsymbol{e}_n) = n$, $0 \le \nu \le 1 - b_n$ holds true.

Case 2: From Lemma 6, we know that $Q^{\triangle}_{m'}(s,b)$ is subadditive on $\mathcal{I}_\delta \times \mathcal{N}$. From Lemma 4, the submodularity of $Q^{\triangle}_{m'}(s,b)$ in (y,m') implies that $m'^*(s,b)$ defined in (20) is nondecreasing in $y \in \mathcal{I}_\delta$. Thus, $m'^*(s,b)$ is in the form of (23).

Case 3: When $b_i \le b_m$, $i \in \mathcal{N} \setminus \{m\}$, we obtain

$$Q^{\triangle}_m(s,b) - Q^{\triangle}_i(s,b) = -\eta E_{swi} + (b_i - b_m)\{Q^{\triangledown}_u(s,b,0) - \min_{u \in \mathcal{U}_{s,1}} Q^{\triangledown}_u(s,b,1)\} < 0. \tag{64}$$

by (17) and (18). Thus, by (20), we know that $m^*(s,b)$ is in the form of (24).

Case 4: From Lemma 5, we know that $Q^{\triangledown}_u(s,b,1)$ is subadditive on $\mathcal{I}_\delta \times \mathcal{U}_{s,1}$. From Lemma 4, the submodularity of $Q^{\triangledown}_u(s,b,1)$ in (y,u) implies that $u^*(s,b,1)$ defined in (21) is nondecreasing in $y \in \mathcal{I}_\delta$. Thus, by (5) and (21), we know that $u^*(s,b,1)$ is in the form of (25).