Ding *et al. EURASIP Journal on Wireless Communications and Networking* (2015) 2015:241
DOI 10.1186/s13638-015-0470-7

EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Structured optimal transmission control in network-coded two-way relay channels

Ni Ding*, Parastoo Sadeghi and Rodney A. Kennedy

Abstract

This paper considers a transmission control problem in network-coded two-way relay channels (NC-TWRC), where the relay buffers randomly arrived packets from two users, and the channels are assumed to be fading. The problem is modeled by a discounted infinite horizon Markov decision process (MDP). The objective is to find an adaptive transmission control policy that minimizes the packet delay, buffer overflow, transmission power consumption and downlink error rate simultaneously and in the long run. By using the concepts of submodularity, multimodularity and $L^\natural$-convexity, we study the structure of the optimal policy searched by dynamic programming (DP) algorithm. We show that the optimal transmission policy is nondecreasing in queue occupancies and/or channel states under certain conditions such as the chosen values of parameters in the MDP model, channel modeling method, and the preservation of stochastic dominance in the transitions of system states. Based one these results, we propose to use two low-complexity algorithms for searching the optimal monotonic policy: monotonic policy iteration (MPI) and discrete simultaneous perturbation stochastic approximation (DSPSA). We show that MPI reduces the time complexity of DP, and DSPSA is able to adaptively track the optimal policy when the statistics of the packet arrival processes change with time.

Keywords: Cross-layer optimization, Discounted Markov decision process, Discrete stochastic approximation, Dynamic programming, $L^\natural$-convexity, Multimodularity, Network coding, Submodularity

1 Introduction

Network coding (NC) was proposed in [1] to maximize the information flow in a wired network. It was introduced in multicast wireless communications to optimize the throughput and has attracted significant interest recently due to the rapid growth in multimedia applications [2]. It was shown in [3] that the power efficiency in wireless transmission systems could be improved by NC. For example, in a 3-node network system, called the network-coded two-way relay channels (NC-TWRC) [4] as shown in Fig. 1, the messages m_1 and m_2 are XORed at the relay and broadcast to the end users. This method, compared to the conventional store-and-forward transmission, reduces the total number of transmissions from 4 to 3 so that the transmission power is saved by 25 %. Since then, numerous optimization problems have been studied in NC-TWRC, e.g., the precoding scheme design proposed in [5], the optimal achievable sum-rate problem studied in [6] and the optimal beamforming method proposed in [7].

In [8], Katti et al. pointed out the importance of being opportunistic in practical NC scenarios. It was suggested that the assumptions in the related research work should comply with the practical wireless environments, e.g., decentralized routing and time-varying traffic rate. This suggestion highlighted a problem in the existing literature; the majority of the studies (e.g., [9, 10]) consider static environments (e.g, synchronized traffic) while ignoring the stochastic nature of the packet arrivals in the data link layer. On the other hand, the randomness of traffic in Fig. 1 poses the problem of how to make an optimal decision in a dynamic environment with a power-delay tradeof;: when there are packet inflows in the relay but no coding opportunities or XORing pairs (e.g., one packet arrives from one user, but no packet arrives from the other), waiting for coding opportunities by holding packets saves transmission power but increases packet delay and results in more packets to be transmitted in the

*Correspondence: ni.ding@anu.edu.au
Research School of Engineering, College of Engineering and Computer Science, Australian National University (ANU), 2601 Canberra, Australia

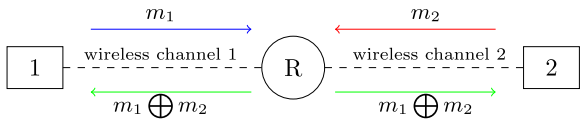

Fig. 1 NC-TWRC [4]. Two users exchange information (m_1 and m_2) via the center node *R* (stands for relay)

future. Since a decision made at any instant affects both the immediate and future costs, the decision-making is a dynamic, instead of a one-time, process, i.e., the objective is to determine a decision rule that is optimal over time. In [11, 12], this problem was studied and solved by a cross-layer design, NC-TWRC with buffering. The optimal policy by Markovian process formulation was shown to minimize the transmission power and packet delay simultaneously and in the long run. In [13], the buffer-assisted NC-TWRC was extended to include the dynamics of wireless channels (Fig. 2). In this system, a transmission policy that solves power-delay tradeoff may not be the best decision rule because it does not consider the possible loss in throughput due to the downlink transmission errors. For this reason, the scheduler is required to make an optimal decision that simultaneously minimizes the transmission power, packet delay, downlink BER in the long run by considering current queue and channel states and their expectations in the future. In [13], this problem was formulated by a discounted infinite horizon Markov decision process (MDP) [14] with channels modeled by finite-state Markov chains (FSMCs) [15]. The optimal transmission policy was shown to be superior to [11, 12] in terms of enhancing the QoS (quality of service, evaluated by packet delay and overflow in the data link layer, and power consumption and error rate in the physical layer) in a practical wireless environment, e.g., Rayleigh fading channels.

The optimal policy of a discounted infinite horizon MDP can be found by dynamic programming (DP) [16], value or policy iterations. However, the DP algorithm is burdened with high complexity. In Fig. 2, the system state is a 4-tuple (two channels and two queues), and the decision/action is a 2-tuple (each associated with the departure control of one queue). In such a high dimensional MDP, *the curse of dimensionality*[1] becomes more evident [17]; the computation load grows quickly if the cardinality of any tuple in the state variable is large. To relieve the curse, one solution is to qualitatively understand the model and prove the existence of a monotonic optimal policy [18]. Then, a low complexity algorithm or a model-free learning method can be proposed, e.g., simultaneous perturbation stochastic approximation (SPSA) [19, 20]. But, monotonic optimal policy does not exist in general. Most often, optimal policy exists, but it varies with the state variable irregularly. In order to prove the existence of certain feature in the optimal policy, we need to extensively analyze the MDP model and the recursive functions in DP algorithm. The basic approach in the existing literature is to show by induction that the submodularity is preserved in each iterative optimization process (maximization/minimization) in DP, e.g., [19, 21]. We adopt the same method in this paper but consider a submodularity in high dimensional cases. Moreover, we use $L^\natural$-convexity and multimodularity, two concepts that were originally defined in discrete convex analysis [22, 23], to describe the joint submodularity and integral convexity in a high dimensional space.

The aim of our work is to prove the existence of a monotonic optimal transmission policy in the NC-TWRC system in Fig. 2. By observing the $L^\natural$-convexity and submodularity of DP function, we derive the sufficient conditions for the optimal policy to be nondecreasing in queue and/or channel states. These structured results are used to derive two low complexity algorithms: monotonic policy iteration (MPI) and discrete simultaneous perturbation stochastic approximation (DSPSA). We compare the time complexity of MPI to that of DP and show the convergence performance of DSPSA algorithm. The main results in this paper are:

- We prove that each tuple in the optimal policy is nondecreasing in the queue state that is controlled by that tuple if the chosen values of unit costs in immediate cost function give rise to an $L^\natural$-convex or multimodular DP. Moreover, we show that the same results found in [19, 21] can also be explained by

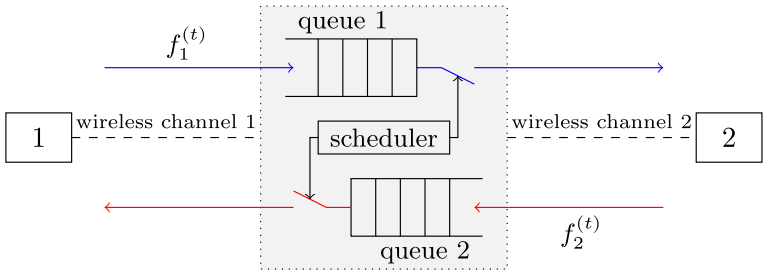

Fig. 2 NC-TWRC with random packet arrivals and fading channels [13]. The incoming packets are buffered by two finite length first-in-first-out (FIFO) queues. The outflows are controlled by a scheduler

$L^\natural$-convexity or multimodularity by a unimodular coordinate transform.

- By thinking of each iteration in DP as a one-stage pure coordination supermodular game, we show that equiprobable traffic rates and certain conditions on unit costs guarantee that each tuple in the optimal policy is monotonic in not only the queue state that is controlled by that tuple but also the queue state that is associated with the information flow of the opposite direction, i.e., the one that is not under the control of that tuple.
- By observing the submodularity of DP, we show the sufficient conditions for an optimal policy to be nondecreasing in both queue and channel states in terms of unit costs, channel statistics, and FSMC models.
- Based on the submodularity, multimodularity, and $L^\natural$-convexity of DP, we show that the optimal transmission control problem in Fig. 2 can be solved by two low-complexity algorithms. One is MPI, a modified DP algorithm with the action searching space progressively shrinking with the increasing indices of queue and/or channel states. It is shown that the time complexity of MPI is much less than that of DP when the cardinality of system state is large. The other algorithm is a stochastic optimization method. We formulate the optimal policy searching problem by a minimization problem over a set of queue thresholds and use the DSPSA algorithm to approximate the minimizer. We show that DSPSA is able to adaptively track the optimal values of queue thresholds when the statistics of packet arrival processes change with time. We run simulations in NC-TWRC with Rayleigh fading channels to show that the average cost incurred by the policy approximated by DSPSA is similar to that incurred by the optimal policy searched by DP.

The rest of this paper is organized as follows. In Section 2, we state the optimization problem in NC-TWRC with random packet arrivals and FSMC modeled channels and clarify the assumptions. In Section 3, we describe the MDP formulation, state the objective, and present the DP algorithm. In Section 4, we investigate the structure in the optimal transmission policy found by DP algorithm in queue and channel states. Section 5 presents MPI and DPSA algorithms.

2 System

Consider the NC-TWRC shown in Fig. 2. User 1 and 2 randomly send packets to each other via the relay. The relay is equipped with two finite-length FIFO queues, queue 1 and 2, to buffer the incoming packets from user 1 and 2, respectively. The outflows of queues are controlled by a scheduler. The scheduler keeps making decisions as to whether or not to transmit packets from queues. If the decision results in a pair of packets in opposite directions transmitted at the same time, they will be XORed (coded) and broadcast. Otherwise, the packet will be simply forwarded to the end user. The objective is to minimize packet delay, queue overflow, transmission power (saved by utilizing the coding opportunities), and downlink transmission errors simultaneously and their expectations in the future. Obviously, the optimization concerns are contradictory to each other: (1) If there does not exist a pair of packets for XORing, waiting for coding opportunity by holding packets results in a high packet delay on average, while transmitting a packet without coding results in one more packet to be transmitted in the future, i.e., more transmission power on average; (2) If the SNR of one channel is low, waiting for high SNR transition by holding packets results in higher packet delay but lower transmission error rate. Therefore, the scheduler must seek an optimal decision rule that solves this *power-delay-error tradeoff*.

It should be pointed out that the problem under consideration is a cross-layer multi-objective optimization one; we want to optimize both the power consumption and transmission error rate in the physical layer and the packet delay in the data link layer. As discussed above, since there are tradeoffs among these optimization metrics, it is not possible to get all of them optimized simultaneously. Therefore, in this paper, we are actually seeking the Pareto optimality of these optimization metrics.[2]

2.1 Assumptions

We consider a discrete-time decision-making process, where the time is divided into small intervals, called *decision epochs* and denoted by $t \in \{0, 1, \ldots, T\}$. Let $i \in \{1, 2\}$ and assume the following:

A1 (*i.i.d.* incoming traffic) Denote random variable $f_i^{(t)} \in \mathcal{F}_i$ as the number of incoming packets to queue i at decision epoch t. Let the maximum number of packets arrived per decision epoch be no greater than 1, i.e., $\mathcal{F}_i = \{0, 1\}$. Assume that $\left\{f_1^{(t)}\right\}$ and $\left\{f_2^{(t)}\right\}$ are two independent *i.i.d.* random processes with $Pr\left(f_i^{(t)} = 1\right) = p_i$ and $Pr\left(f_i^{(t)} = 0\right) = 1 - p_i$ for all t.

A2 (modulation scheme) Packets are of equal length. The packets arrived at the relay are decoded and stored in the queues. The relay transmits packets by BPSK modulation. Denote L_P the packet length in bits. Since the maximum information flow is one packet per decision epoch, each decision epoch lasts for L_P symbol durations. The relay can set a certain field in the header of a packet so as to notify the receivers whether the packet is XORed or not.

A3 (finite-length queues) Queue i can store maximum L_i packets. At each t, the scheduler makes a decision and incurs an immediate cost before the event $\mathbf{f}^{(t)} = \left(f_1^{(t)}, f_2^{(t)}\right)$. Denote $b_i^{(t)} \in \mathcal{B}_i$ as the occupancy of queue i at the beginning of decision epoch t, then $\mathcal{B}_i = \left\{0, 1, \ldots, L_i + \max\left\{f_i^{(t-1)}\right\}\right\} = \{0, 1, \ldots, L_i + 1\}$. If the relay's decision results in queue occupation $L_i + 1$, the newly arrived packet will be dropped. We call it packet lost due to the *queue overflow*.

A4 (Markovian channel modeling) Let the full variation range of $\gamma_i^{(t)}$, the instantaneous SNR of channel i, be partitioned into K_i non-overlapping regions $\{[\Gamma_1, \Gamma_2), [\Gamma_2, \Gamma_3), \ldots, [\Gamma_{K_i}, \infty)\}$, called channel states. Here, the SNR boundaries satisfy $\Gamma_1 < \Gamma_2 < \ldots < \Gamma_{K_i}$. Denote $\mathcal{G}_i = \{1, 2, \ldots, K_i\}$ as the state set of channel i and $g_i^{(t)}$ as the state of channel i at decision epoch t. We say that $g_i^{(t)} = k_i$ if $\gamma_i^{(t)} \in [\Gamma_{k_i}, \Gamma_{k_i+1})$. Each channel is modeled by a finite-state Markov chain (FSMC) [15], where the state evolution of channel i is governed by the transition probability $P_{g_i^{(t)} g_i^{(t+1)}} = Pr\left(g_i^{(t+1)} \mid g_i^{(t)}\right)$.

A5 (downlink channel state information) Let $\left\{g_1^{(t)}\right\}$ and $\left\{g_2^{(t)}\right\}$ be two independent and *i.i.d.* random processes. The relay has the channel state information (the value of channel state and its transition probabilities) of both channels before the decision making at t.

3 Markov decision process formulation

Based on A1, A4, and A5, we know that the statistics of the incoming traffic flow and channel dynamics associated with user 1 or 2 are time-invariant. It follows that the transmission control problem in Fig. 2 can be formulated as a stationary Markov decision process (MDP). In the following context, we drop the decision epoch notation t in A1-A5 and use the notation y and y' for the system variable y at the current and next decision epochs, respectively.

3.1 System state

Denote the system state $\mathbf{x} = (\mathbf{b}, \mathbf{g}) \in \mathcal{X}$, where $\mathbf{b} = (b_1, b_2) \in \mathcal{B}_1 \times \mathcal{B}_2$ and $\mathbf{g} = (g_1, g_2) \in \mathcal{G}_1 \times \mathcal{G}_2$, i.e., $\mathcal{X} = \mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{G}_1 \times \mathcal{G}_2$. $\times$ denotes the Cartesian product. We also use the 4-tuple notation $\mathbf{x} = (b_1, b_2, g_1, g_2)$ in the following context.

3.2 Action

Denote action $\mathbf{a} = (a_1, a_2) \in \mathcal{A}$, where $a_i \in \mathcal{A}_i = \{0, 1\}$ denotes the number of packets departed from queue i and $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 = \{0, 1\}^2$. The terminology of actions are shown in Table 1.

Table 1 Action set

a	Action itemize
(0, 0)	No transmission
(1, 0)	Forward one packet in queue 1
(0, 1)	Forward one packet in queue 2
(1, 1)	XOR two packets one in each queue, then broadcast.

3.3 State transition probabilities

The transition probability $P^{\mathbf{a}}_{\mathbf{x}\mathbf{x}'} = Pr(\mathbf{x}'|\mathbf{x}, \mathbf{a})$ denotes the probability of being in state $\mathbf{x}'$ at next decision epoch if action $\mathbf{a}$ is taken in state $\mathbf{x}$ at current decision epoch. Due to the assumptions of independent random processes in A1 and A5, the state transition probability is given by

$$P^{\mathbf{a}}_{\mathbf{x}\mathbf{x}'} = P^{\mathbf{a}}_{\mathbf{b}\mathbf{b}'} P_{\mathbf{g}\mathbf{g}'} = \prod_{i=1}^{2} P^{a_i}_{b_i b_i'} P_{g_i g_i'}, \tag{1}$$

where $P_{g_i g_i'}$ is determined by channel statistics and FSMC modeling method in A4 and $P^{a_i}_{b_i b_i'}$ is the queue state transition probability. At current decision epoch, the occupancy of queue i after decision a_i is $\min\{[b_i - a_i]^+, L_i\}$, where $[y]^+ = \max\{y, 0\}$. The occupancy at the beginning of the next decision epoch is given by

$$b_i' = \min\left\{[b_i - a_i]^+, L_i\right\} + f_i. \tag{2}$$

Therefore, the state transition probability of queue i is

$$\begin{aligned} P^{a_i}_{b_i b_i'} &= Pr\left(f_i = b_i' - \min\left\{[b_i - a_i]^+, L_i\right\}\right) \\ &= Pr\left(f_i = b_i' - [b_i - a_i]^+ + \mathcal{I}_{\{[b_i - a_i]^+ > L_i\}}\right) \\ &= \begin{cases} Pr\left(f_i = b_i' - [b_i - a_i]^+\right) & [b_i - a_i]^+ \le L_i \\ Pr\left(f_i = b_i' - L_i\right) & [b_i - a_i]^+ > L_i \end{cases}, \end{aligned} \tag{3}$$

where $\mathcal{I}_{\{\cdot\}}$ is the indicator function that returns 1 if the expression in $\{\cdot\}$ is true and 0 otherwise.

3.4 Immediate cost

$C : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+$ is the cost incurred immediately after action $\mathbf{a}$ is taken in state $\mathbf{x}$ at current decision epoch. It reflects three optimization concerns: the packet delay and queue overflow, the transmission power, and the downlink transmission error rate.

3.4.1 Holding and overflow cost

We define h_i, the holding and queue overflow cost associated with queue i, as

$$\begin{aligned} h_i(y_i) &= \lambda \min\{[y_i]^+, L_i\} + \xi_o \mathcal{I}_{\{[y_i]^+ = L_i + 1\}} \\ &= \lambda [y_i]^+ + (\xi_o - \lambda) \mathcal{I}_{\{[y_i]^+ = L_i + 1\}}. \end{aligned} \tag{4}$$

$\lambda > 0$ is the unit holding cost and $\xi_o > \lambda$ is the unit queue overflow cost, which makes $h_i(y_i)$ a nondecreasing convex function. In the case when $y_i = b_i - a_i$,

$\min\{[y_i]^+, L_i\}$ and $\mathcal{I}_{\{[y_i]^+=L_i+1\}}$ count the number of packets held in queue i and the number of packets lost due the overflow of queue i, respectively. We say that the term $\lambda \min\{[y_i]^+, L_i\}$ accounts for the packet delay because by Little's Law, the average packet delay is proportional to the average number of packets held in the queue in the long run for a given packet arrival rate [24]. We sum up h_i for $i \in \{1,2\}$ and obtain the total holding and overflow cost as

$$C_h(\mathbf{b},\mathbf{a}) = \sum_{i=1}^{2} h_i(b_i - a_i). \tag{5}$$

3.4.2 Transmission cost

Since forwarding and broadcasting one packet, either coded or non-coded, consume the same amount of energy, we have the immediate transmission cost as

$$t_r(\mathbf{a}) = \tau \mathcal{I}_{\{a_1=1 \text{ or } a_2=1\}} = \begin{cases} 0 & \mathbf{a} = (0,0) \\ \tau & \text{otherwise} \end{cases}, \tag{6}$$

where $\tau > \lambda$ is the unit transmission cost and $\mathcal{I}_{\{a_1=1 \text{ or } a_2=1\}}$ counts the number of transmissions resulting from action $\mathbf{a}$.

Note that (5) and (6) form a power-delay tradeoff. A policy that always transmits whenever there is an incoming packet without considering coding opportunities in the long run is penalized by (6), and a policy that always holds packet to wait for coding opportunities without considering the average packet delay is penalized by (5).

3.4.3 Packet error cost

Since packet errors in downlink transmissions happen only when we decide to transmit, we define the immediate packet error cost due to the action a_i as

$$\text{err}(g_{-i}, a_i) = \eta a_i P_e(g_{-i}), \tag{7}$$

where η is the unit packet error cost and $-i \in \{1,2\} \setminus \{i\}$, i.e., $-i = 2$ if $i = 1$, and $-i = 1$ if $i = 2$. The reason we have $\text{err}(g_{-i}, a_i)$ is because the packet departing queue i is transmitted through channel $-i$, e.g., the relay sends one packet in queue 1 through fading channel 2 when $a_1 = 1$. $P_e(g_i)$ is estimation of the average BER when transmitting a packet, either coded or non-coded, through channel i when the state is g_i. Since BPSK modulation is used at the relay, we define P_e as

$$P_e(g_i) = \frac{1}{2}\text{erfc}(\sqrt{\Gamma_{g_i}}). \tag{8}$$

Here, $P_e(g_i) \leq 0.5$ because $\Gamma_1 \geq 0$ in A4.

Note, the aforementioned power-delay tradeoff formed by (5) and (6) just poses the problem of whether or not to transmit if an instantaneous packet inflow is not able to form an XORing pair. However, if the scheduler considers downlink transmission error rate in addition, a policy that always broadcasts XORed packets whenever there is a coding opportunity without considering downlink channel states is penalized by (7). Therefore, (5), (6), and (7) form a power-delay-error tradeoff.

In summary, we define the immediate cost as

$$C(\mathbf{x},\mathbf{a}) = C(\mathbf{b},\mathbf{g},\mathbf{a}) = C_h(\mathbf{b},\mathbf{a}) + C_t(\mathbf{g},\mathbf{a}), \tag{9}$$

where

$$C_t(\mathbf{g},\mathbf{a}) = \sum_{i=1}^{2} \text{err}(g_{-i}, a_i) + t_r(\mathbf{a}). \tag{10}$$

Here, $C(\mathbf{x},\mathbf{a})$ is in fact a linear combination of loss functions (each quantifies an optimization concern). The unit cost λ, ξ_o, τ, and η can be considered as the weight factors that are either given or adjustable depending on the real applications. In Section 4, we will derive the sufficient conditions of the existence of a structured optimal policy mainly in terms of the chosen values of these unit costs.

3.5 Objective and dynamic programming

Let $\mathbf{x}^{(t)}$ and $\mathbf{a}^{(t)}$ denote the state and action at decision epoch t, respectively, and consider an infinite-horizon MDP modeling where the discrete decision making process is assumed to be infinitely long. We can describe the long-run objective as

$$\min \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t C\left(\mathbf{x}^{(t)}, \mathbf{a}^{(t)}\right) | \mathbf{x}^{(0)}\right], \forall \mathbf{x}^{(0)} \in \mathcal{X}, \tag{11}$$

where $\mathbf{x}^{(t+1)} \sim Pr(\cdot|\mathbf{x}^{(t)}, \mathbf{a}^{(t)})$ and $\beta \in [0,1)$ is the discounted factor that ensures the convergence of the series. It is proved in [14] that if the state space $\mathcal{X}$ is countable, the action set $\mathcal{A}$ is finite, and the MDP is stationary, there exists a deterministic stationary policy $\theta^* : \mathcal{X} \to \mathcal{A}$ that optimizes (11), and θ^* can be searched by DP

$$V^{(n)}(\mathbf{x}) = \min_{\mathbf{a}\in\mathcal{A}} Q^{(n)}(\mathbf{x},\mathbf{a}), \forall \mathbf{x} \in \mathcal{X}, \tag{12}$$

where

$$Q^{(n)}(\mathbf{x},\mathbf{a}) = C(\mathbf{x},\mathbf{a}) + \beta \sum_{\mathbf{x}'\in\mathcal{X}} P^{\mathbf{a}}_{\mathbf{x}\mathbf{x}'} V^{(n-1)}(\mathbf{x}'). \tag{13}$$

Here, n denotes the iteration index and $V^{(0)}(\mathbf{x}) = 0$ for all $\mathbf{x}$. Usually, a very small convergence threshold $\epsilon > 0$ is applied so that DP terminates when $\left|V^{(N-1)}(\mathbf{x}) - V^{(N)}(\mathbf{x})\right| \leq \epsilon$ for all $\mathbf{x}$ and $N < \infty$.[3] The optimal policy is obtained as $\theta^*(\mathbf{x}) = \arg\min_{\mathbf{a}\in\mathcal{A}} Q^{(N)}(\mathbf{x},\mathbf{a})$.

As discussed in Section 2, the problem under consideration is a cross-layer multi-objective one. When defining the immediate cost function (9), we use scalarization technique, i.e., $C(\mathbf{x},\mathbf{a})$ is a weighted sum of the holding and packet overflow costs incurred in the data link layer and the transmission power consumption and error rate incurred in the physical layer. Therefore, the optimal policy θ^* is in fact a Pareto optimal solution.[4] It should be

clear that a Pareto optimal solution is not optimal if we just consider an individual optimization metric, e.g., θ^* is not the optimal solution if we just want to minimize the power consumption in the physical layer.

4 Structured optimal policies

The time complexity in iteration n in DP is $O(|\mathcal{X}|^2|\mathcal{A}|)$. There are $|\mathcal{X}|$ minimization operations, each of which requires $|\mathcal{A}|$ calculations of $Q^{(n)}$, and each $Q^{(n)}$ value requires $|\mathcal{X}|$ multiplications over state $\mathbf{x}'$. Since $|\mathcal{X}| = |\mathcal{B}_1||\mathcal{B}_2||\mathcal{G}_1||\mathcal{G}_2|$, the complexity grows quadratically if the cardinality of any tuple in the state variable increases. If the node-to-node transmission in NC-TWRC is via multiple channels (e.g., single-user MIMO channels), the complexity grows exponentially with the number of user-to-relay channels, which may severely overload the CPU. In this section, we investigate the submodularity, $L^\natural$-convexity and multimodularity of functions $Q^{(n)}(\mathbf{x},\mathbf{a})$ and $V^{(n)}(\mathbf{x})$ in DP to establish the sufficient conditions for the existence of a monotonic optimal policy. These results serve as the prerequisites for the low complexity algorithms proposed in Section 5. We first clarify some concepts as follows.

Definition 4.1 (Monotonic policy)**.** *Let $\theta\colon \mathbb{Z}^n \to \mathbb{Z}^m$, $\theta(\mathbf{x})$ is monotonic nondecreasing if $\theta(\mathbf{x}^+) \succeq \theta(\mathbf{x}^-)$, for all $\mathbf{x}^+, \mathbf{x}^- \in \mathbb{Z}^n$ such that $\mathbf{x}^+ \succeq \mathbf{x}^-$, where $\succeq$ denotes componentwise greater than or equal to.*

Definition 4.2 (Submodularity [23, 25])**.** *Let $\mathbf{e}_i \in \mathbb{Z}^n$ be an n-tuple with all zero entries except the ith entry being one. $f\colon \mathbb{Z}^n \to \mathbb{R}_+$ is submodular if $f(\mathbf{x}+\mathbf{e}_i)+f(\mathbf{x}+\mathbf{e}_j) \geq f(\mathbf{x})+f(\mathbf{x}+\mathbf{e}_i+\mathbf{e}_j)$ for all $\mathbf{x} \in \mathbb{Z}^n$ and $1 \leq i,j \leq n$. f is strictly submodular if the inequality is strict.*

In DP, a submodular function $Q^{(n)}(\mathbf{x},\mathbf{a})$ has $Q^{(n)}(\mathbf{x},\mathbf{a}^-) - Q^{(n)}(\mathbf{x},\mathbf{a}^+)$ nondecreasing in $\mathbf{x}$ for all $\mathbf{a}^+ \succeq \mathbf{a}^-$, i.e., the preference of choosing action $\mathbf{a}^+$ over $\mathbf{a}^-$ is always nondecreasing in $\mathbf{x}$. Therefore, an increase in the state variable $\mathbf{x}$ implies an increase in the decision rule $\theta^{(n)}(\mathbf{x}) = \min_{\mathbf{a}} Q^{(n)}(\mathbf{x},\mathbf{a})$. This property is summarized in a general form in the following lemma.

Lemma 4.3. *If $g\colon \mathbb{Z}^n \to \mathbb{R}_+$ is submodular in $(\mathbf{x},\mathbf{y}) \in \mathbb{Z}^n$, then $f(\mathbf{x}) = \min_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$ is submodular in $\mathbf{x}$, and the minimizer $\mathbf{y}^*(\mathbf{x}) = \arg\min_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$ is nondecreasing in $\mathbf{x}$ [26].*

Definition 4.4 ($L^\natural$-convexity [23])**.** *$f\colon \mathbb{Z}^n \to \mathbb{R}_+$ is $L^\natural$-convex if $\psi(\mathbf{x},\zeta) = f(\mathbf{x}-\zeta\mathbf{1})$ is submodular in $(\mathbf{x},\zeta)$, where $\mathbf{1} = (1,1,\ldots,1) \in \mathbb{Z}^n$ and $\zeta \in \mathbb{Z}$.*

Definition 4.5 (multimodularity [23])**.** *$f\colon \mathbb{Z}^n \to \mathbb{R}_+$ is multimodular if $\psi(\mathbf{x},\zeta) = f(x_1-\zeta, x_2-x_1, \ldots, x_n-x_{n-1})$ is submodular in $(\mathbf{x},\zeta)$, where $\zeta \in \mathbb{Z}$.*

$L^\natural$-convexity and multimodularity are two concepts defined in discrete convex analysis [27]. $L^\natural$-convexity implies submodularity while multimodularity implies supermoduarity[5] [28]. They both contribute to a monotonic structure in the optimal policy.

Lemma 4.6. *If $g\colon \mathbb{Z}^n \to \mathbb{R}_+$ is $L^\natural$-convex/multimodular in $(\mathbf{x},\mathbf{y}) \in \mathbb{Z}^n$, then $f(\mathbf{x}) = \min_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$ is $L^\natural$-convex/multimodular in $\mathbf{x}$, and the minimizer $\mathbf{y}^*(\mathbf{x}) = \arg\min_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$ is nondecreasing/nonincreasing in $\mathbf{x}$ [28, 29].*

The unimodular coordinate transform below describes the relationship between $L^\natural$-convexity and multimodularity.

Lemma 4.7 (unimodular coordinate transform [23, 28])**.** *Let matrix $M_{n,i} = \begin{bmatrix} -U_i & 0 \\ 0 & L_{n-i} \end{bmatrix}$, where U_i and L_i are the $i \times i$ upper and lower triangular matrix with all nonzero entries being one, respectively, then*

(a) a function $f\colon \mathbb{Z}^n \to \mathbb{R}_+$ is multimodular if and only if it can be represented by $f(\mathbf{x}) = g(\pm M_{n,i}\mathbf{x})$ for some $L^\natural$-convex function g.
(b) a function $g\colon \mathbb{Z}^n \to \mathbb{R}_+$ is $L^\natural$-convex if and only if it can be represented by $g(\mathbf{x}) = f\left(\pm M_{n,i}^{-1}\mathbf{x}\right)$ for some multimodular function f.

Definition 4.8 (First order stochastic dominance [18])**.** *Let $\tilde{\rho}(x)$ be a random selection on space $\mathcal{X}$ according to a probability measure $\mu(x)$ where x conditions the random selection, then $\tilde{\rho}(x)$ is first order stochastically nondecreasing in x if $\mathbb{E}[u(\tilde{\rho}(x^+))] \geq \mathbb{E}[u(\tilde{\rho}(x^-))]$ for all nondecreasing functions u and $x^+ \geq x^-$.*

4.1 Structured properties of dynamic programming

To propose the prototypical procedure of proving the existence of a monotonic optimal policy, we first define a $\mathcal{P}^\star$ property as follows:

Definition 4.9 ($\mathcal{P}^\star$ property)**.** *$f\colon \mathbb{Z}^n \to \mathbb{R}_+$ has $\mathcal{P}^\star$ property in $(\mathbf{x},\mathbf{y}) \in \mathbb{Z}^n$ if $f^*(\mathbf{x}) = \min_{\mathbf{y}} f(\mathbf{x},\mathbf{y})$ has $\mathcal{P}^\star$ property in $\mathbf{x}$ and $\mathbf{y}^*(\mathbf{x}) = \arg\min_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$ is monotonic (nondecreasing/nonincreasing) in $\mathbf{x}$.*

Theorem 4.10. *Submodularity, $L^\natural$-convexity and multimodularity have $\mathcal{P}^\star$ property.*

Proof. It can be directly proved by Lemma 4.3 and Lemma 4.6. □

We therefore propose an approach, similar to Proposition 5 in [18], as follows:

Proposition 4.11. *Let DP converge at Nth iteration. The optimal value function* $V^*(\mathbf{x}) = V^{(N)}(\mathbf{x})$ *has* $\mathcal{P}^\star$ *property, and the optimal policy* θ^* *is monotonic in* $\mathbf{x}$*, if:*

(a) $C(\mathbf{x}, \mathbf{a})$ *has* $\mathcal{P}^\star$ *property,*
(b) $Q^{(n)}(\mathbf{x}, \mathbf{a}) = C(\mathbf{x}, \mathbf{a}) + \beta \sum_{\mathbf{x}' \in \mathcal{X}} P^{\mathbf{a}}_{\mathbf{x}\mathbf{x}'} V^{(n-1)}(\mathbf{x}')$ *has* $\mathcal{P}^\star$ *property for all* $\mathcal{P}^\star$ *property functions* $V^{(n-1)}$ *and n.*

Proof. Since DP starts from $V^{(0)}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$, $Q^{(1)} = C(\mathbf{x}, \mathbf{a})$ has $\mathcal{P}^\star$ property. So $V^{(1)}(\mathbf{x}) = \min_{\mathbf{a} \in \mathcal{A}} Q^{(1)}(\mathbf{x}, \mathbf{a})$ has $\mathcal{P}^\star$ property. By induction, assume $V^{(n-1)}(\mathbf{x}, \mathbf{a})$ has $\mathcal{P}^\star$ property. Then $Q^{(n)}$ and $V^{(n)}(\mathbf{x}) = \min_{\mathbf{a} \in \mathcal{A}} Q^{(n)}(\mathbf{x}, \mathbf{a})$ have $\mathcal{P}^\star$ property. Therefore, $Q^{(N)}(\mathbf{x}, \mathbf{a})$ and $V^*(\mathbf{x}) = V^{(N)}(\mathbf{x})$ must also possess $\mathcal{P}^\star$ property, and $\theta^*(\mathbf{x}) = \arg\min_{\mathbf{a} \in \mathcal{A}} Q^{(N)}(\mathbf{x}, \mathbf{a})$ is monotonic in $\mathbf{x}$. □

4.2 Monotonic policies in queues states

4.2.1 Nondecreasing a_i^* in b_i

Let the optimal action be $\mathbf{a}^* = \theta^*(\mathbf{x}) = (\theta_1^*(\mathbf{x}), \theta_2^*(\mathbf{x}))$. $a_i^* = \theta_i^*(\mathbf{x})$ is the optimal action to queue i determined by θ^*. The following theorem shows that the optimal action a_i^* is monotonic in b_i, the state of queue being controlled by a_i if the unit costs satisfy a certain condition.

Theorem 4.12. *If* $\xi_o \geq 2\lambda + \eta + \tau$,[6] *then for all* $i \in \{1, 2\}$ $C(\mathbf{x}, \mathbf{a})$ *and* $Q^{(n)}(\mathbf{x}, \mathbf{a})$ *are nondecreasing in* b_i *and* $L^\natural$*-convex in* (b_i, a_i)*,* $V^*(\mathbf{x})$ *is nondecreasing and* $L^\natural$*-convex in* b_i*, and the optimal action* a_i^* *is nondecreasing in* b_i*.*

Proof. We define two functions

$$\tilde{C}(\mathbf{y}, \mathbf{g}, \mathbf{a}) = \tilde{C}_h(\mathbf{y}) + C_t(\mathbf{g}, \mathbf{a}), \tag{14}$$

where $\tilde{C}_h(\mathbf{y}) = \sum_{i=1}^2 h_i(y_i)$ and

$$\tilde{Q}^{(n)}(\mathbf{y}, \mathbf{g}, \mathbf{a}) = \tilde{C}(\mathbf{y}, \mathbf{g}, \mathbf{a}) + \beta \mathbb{E}_{\mathbf{g}'}\left[V_{\mathbf{f}}^{(n-1)}(\mathbf{y}, \mathbf{g}') \middle| \mathbf{g} \right]. \tag{15}$$

Here,

$$V_{\mathbf{f}}^{(n-1)}(\mathbf{y}, \mathbf{g}') = \mathbb{E}_{\mathbf{f}}\left[V^{(n-1)}(\min\{[y_1]^+, L_1\} + f_1, \min\{[y_2]^+, L_2\} + f_2, \mathbf{g}') \right], \tag{16}$$

$\mathbf{y} = (y_1, y_2)$ and $\mathbf{f} = (f_1, f_2)$. It is easy to see that $C(\mathbf{b}, \mathbf{g}, \mathbf{a}) = \tilde{C}(\mathbf{b} - \mathbf{a}, \mathbf{g}, \mathbf{a})$ and $Q^{(n)}(\mathbf{b}, \mathbf{g}, \mathbf{a}) = \tilde{Q}^{(n)}(\mathbf{b} - \mathbf{a}, \mathbf{g}, \mathbf{a})$. Since

$$\begin{bmatrix} b_i - a_i \\ a_i \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_i \\ a_i \end{bmatrix} = -M_{2,2}^{-1} \begin{bmatrix} b_i \\ a_i \end{bmatrix}, \tag{17}$$

according to Lemma 4.7(b), it follows that proving the $L^\natural$-convexity of $C(\mathbf{b}, \mathbf{g}, \mathbf{a})$ and $Q^{(n)}(\mathbf{b}, \mathbf{g}, \mathbf{a})$ in (b_i, a_i) is equivalent to showing the multimodularity of $\tilde{C}(\mathbf{y}, \mathbf{g}, \mathbf{a})$ and $\tilde{Q}^{(n)}(\mathbf{y}, \mathbf{g}, \mathbf{a})$ in (y_i, a_i). It is also clear that the monotonicity of $C(\mathbf{b}, \mathbf{g}, \mathbf{a})$ and $Q^{(n)}(\mathbf{b}, \mathbf{g}, \mathbf{a})$ in b_i is equivalent to the monotonicity of $\tilde{C}(\mathbf{y}, \mathbf{g}, \mathbf{a})$ and $\tilde{Q}^{(n)}(\mathbf{y}, \mathbf{g}, \mathbf{a})$ in y_i. See Appendix C for the proof of the monotonicity and multimodularity of $\tilde{C}(\mathbf{y}, \mathbf{g}, \mathbf{a})$ and $\tilde{Q}^{(n)}(\mathbf{y}, \mathbf{g}, \mathbf{a})$ in y_i and (y_i, a_i), respectively.

According to Proposition 4.7.3 in [14], $V^*(\mathbf{x})$ is nondecreasing in b_i. By Theorem 4.10 and Proposition 4.11, $V^*(\mathbf{x})$ is $L^\natural$-convex in b_i, and a_i^* is nondecreasing in b_i. □

Note, Theorem 4.12 aligns with the existing results in the literature, e.g., the adaptive MIMO transmission control [21] and the Markov game modeled adaptive modulation of cognitive radio [19]. In fact, both of them can be explained by $L^\natural$-convexity. In [21], the monotonicity of a_i^* in b_i was shown by the multimodularity in $(b_i, -a_i)$. But,

$$\begin{bmatrix} b_i \\ -a_i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} b_i \\ a_i \end{bmatrix} = -M_{2,1}^{-1} \begin{bmatrix} b_i \\ a_i \end{bmatrix} \tag{18}$$

By Lemma 4.7(b), we know that if the a function is multimodular in $(b_i, -a_i)$, then it must be $L^\natural$-convex in (b_i, a_i). Consequently, $V^{(n)}(\mathbf{x})$ is integer convex in b_i because $L^\natural$-convexity in one dimension is exactly integer convexity[7]. In [19], the monotonicity of a_i^* was shown by the submodularity of $Q^{(n)}$ in (b_i, a_i). But, $Q^{(n)}$ is a function of $b_i - a_i$. According to Definition 4.4, the $L^\natural$-convexity of $g(x_1, x_2) = f(x_1 - x_2)$ in (x_1, x_2) is equivalent to the submodularity of $g(x_1, x_2)$ in (x_1, x_2). So $Q^{(n)}$ is also $L^\natural$-convex in (b_i, a_i).

4.2.2 Nondecreasing a_i^* in (b_1, b_2)

We formulate the optimization problem in the nth iteration of DP by a 2-player 2-strategy game, which is called one-stage game in Fig. 3. Assume that action a_1 is taken by player 1, and a_2 is taken by player 2. Obviously, it is a pure coordination game where the utility $-Q^{(n)}(\mathbf{x}, (a_1, a_2))$ is the same to player 1 and 2.

We prove, in Appendix D, that Fig. 3 is a supermodular game with utility function $-Q^{(n)}(\mathbf{x}, (a_1, a_2))$ strictly supermodular in $\mathbf{a} = (a_1, a_2)$ for all $\mathbf{x}$ and $V^{(n-1)}(\mathbf{x}')$ that is $L^\natural$-convex in $\mathbf{b}' = (b_1', b_2')$. It is proved in [30] that there exists at least one equilibrium (a_1^*, a_2^*) in the form of pure strategy in a supermodular game. Then, we have the following theorem for the monotonicity of the optimal action a_i^* in $\mathbf{b} = (b_1, b_2)$.

Theorem 4.13. *If*

(a) $\xi_o \geq 2\lambda + \eta + \tau$,

	$a_1 = 0$	$a_1 = 1$
$a_2 = 0$	$-Q^{(n)}(\mathbf{x}, (0,0))$	$-Q^{(n)}(\mathbf{x}, (1,0))$
$a_2 = 1$	$-Q^{(n)}(\mathbf{x}, (0,1))$	$-Q^{(n)}(\mathbf{x}, (1,1))$

Fig. 3 Utility matrix of one-stage pure coordination game in the nth iteration in DP. $-Q^{(n)} : \boldsymbol{\mathcal{A}}_1 \times \boldsymbol{\mathcal{A}}_2 \to \mathbb{R}_-$ is considered the utility function for a fixed $\mathbf{x}$

(b) *one-stage game (in Fig. 3) has two pure strategy equilibria* $(0,0)$ *and* $(1,1)$ *for all* $\mathbf{x} = (b_1, b_2, g_1, g_2)$ *such that* $b_i < L_i + 1$ *for all* $i \in \{1,2\}$,

then $C(\mathbf{x},\mathbf{a})$ *and* $Q^{(n)}(\mathbf{x},\mathbf{a})$ *are* $L^\natural$*-convex in* $(\mathbf{b},\mathbf{a}) = (b_1, b_2, a_1, a_2)$, *the optimal value function* $V^*(\mathbf{x})$ *is* $L^\natural$*-convex in* $\mathbf{b} = (b_1, b_2)$ *and the optimal action* $\mathbf{a}^* = (a_1^*, a_2^*)$ *is nondecreasing in* $\mathbf{b} = (b_1, b_2)$.

Proof. The proof is in Appendix E. □

Here is a corollary of Theorem 4.13.

Corollary 4.14. *If*

(a) $\xi_o \geq 2\lambda + \eta + \tau$,
(b) $p_1 = p_2 = 0.5$,
(c) $\beta \leq \frac{2(\tau-\lambda)}{\tau+\eta}$,

then Theorem 4.13 holds.

Proof. The proof is in Appendix F. □

We show examples of Theorems 4.12 and 4.13 in Figs. 4, 5, 6 and 7. The results are collected by value iteration, a DP algorithm, applied on an NC-TWRC system with Bernoulli packet arrivals, 5 queue states, and 8 channel states, i.e., $f_i^{(t)} \sim$ Bernoulli(p_i), $L_i = 3$ and $K_i = 8$ for all t and $i \in \{1,2\}$. In Fig. 4, we choose the values of unit costs to make Theorem 4.12 hold. As shown in the figure, the optimal action a_1^* and a_2^* are monotonic in b_1 and b_2, respectively, i.e., a_i^* is nondecreasing in the queue state that is being controlled by a_i. In Fig. 5, we change the value of unit cost ξ_o to breach the condition in Theorem 4.12 so that the monotonicity of a_i^* in b_i is not guaranteed. In this case, a_1^* that is not monotonic in b_1.

In Fig. 6, we choose the equiprobable packet arrival rates $p_1 = p_2 = 0.5$ and the unit costs according to Corollary 4.14 to make Theorem 4.13 hold. As shown in the figure, the optimal action a_1^* and a_2^* are both nondecreasing in (b_1, b_2). As compared to Fig. 4, in this case, a_i^* is also monotonic in b_{-i}, the queue state that is affected by the message flow and transmission control in the opposite direction, i.e., the queue state that is not controlled by a_i. In Fig. 7, we switch unit cost η from 1 to 2 so that Theorem 4.13 no longer holds. In this case, neither a_1^* nor a_2^* is monotonic in (b_1, b_2). But, the condition in Theorem 4.12 is satisfied. Therefore, a_1^* and a_2^* are still nondecreasing in b_1 and b_2, respectively.

4.3 Monotonic policies in channel states

The related research work in the existing literature considers the structure of the optimal policy in queue state only, e.g., [19, 21, 24]. This section breaks this limitation in that we extend the investigation of the monotonicity to the channel states. The main results are summarized as follows.

Theorem 4.15. *If*

(a) $\xi_o \geq 2\lambda + \eta + \tau$,
(b) $P_e(g_i) \geq P_e(g_i + 1)$,
(c) $P_{g_i g_i'}$ *is first order stochastic nondecreasing in* g_i,
(d) $\beta \leq \frac{P_e(g_i) - P_e(g_i+1)}{\sum_{g_i'} P_{g_i g_i'} (P_e(g_i') - P_e(g_i'+1))}$.

then $C(\mathbf{x},\mathbf{a})$ *and* $Q^{(n)}(\mathbf{x},\mathbf{a})$ *is submodular in* (b_i, g_{-i}, a_i), $V^*(\mathbf{x})$ *is submodular in* (b_i, g_{-i}), *and the optimal action* a_i^* *is nondecreasing in* (b_i, g_{-i}).

Proof. The proof is in Appendix G. □

In Theorem 4.15, condition (b) is straightforwardly satisfied because of the definition of P_e in (8) and assumption A4. Conditions (c) and (d) depend on the fading statistics and the FSMC modeling method. In fact, condition (c) is not hard to satisfy.

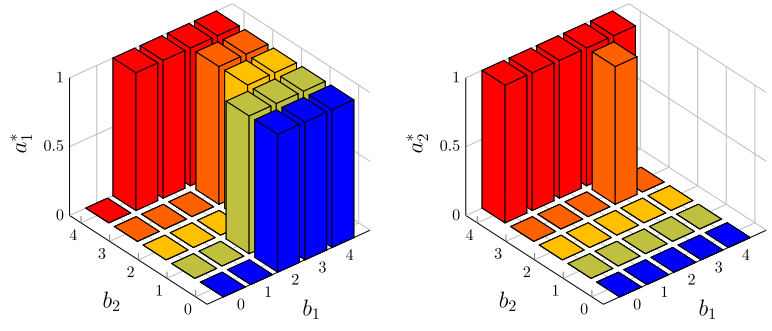

Fig. 4 The optimal action a_1^* (*left*) and a_2^* (*right*) vs. queue states b_1 and b_2 when $g_1 = 1$ and $g_2 = 2$. $p_1 = 0.1, p_2 = 0.2, \lambda = 0.05, \tau = 1, \eta = 2$, $\xi_o = 4$, and $\beta = 0.97$. In this case, $\xi_o \geq 2\lambda + \eta + \tau$. The condition in Theorem 4.12 is satisfied. Therefore, a_1^* and a_2^* are nondecreasing in b_1 and b_2, respectively

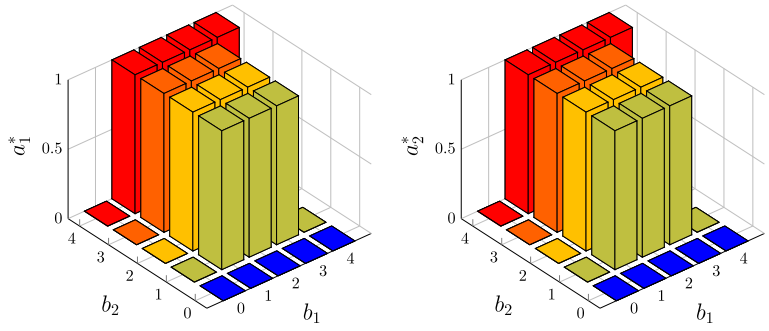

Fig. 5 The optimal action a_1^* (*left*) and a_2^* (*right*) vs. queue states b_1 and b_2 when $g_1 = 1$ and $g_2 = 2$. $p_1 = 0.1, p_2 = 0.2, \lambda = 0.05, \tau = 1, \eta = 2$, $\xi_o = 1$, and $\beta = 0.97$. In this case, $\xi_o < 2\lambda + \eta + \tau$. Theorem 4.12 no longer holds. As can be seen, a_1^* is not monotonic in b_1

Corollary 4.16. *If the FSMC of channel i adopts equiprobable partitioning (of the full range of SNR), and channel i experiences slow and flat Rayleigh fading, then condition (c) in Theorem 4.15 are satisfied.*

Proof. The proof is in Appendix H. □

We show examples of Theorem 4.15 in Figs. 8 and 9. In Fig. 8, we use the same system parameters as in Fig. 4 except that the discount factor β is switched from 0.97 to 0.95 in order to satisfy the inequality in condition (d) of Theorem 4.15. The results are obtained from an NC-TWRC system where the channels experience slow and flat Rayleigh fading with average SNR $\overline{\gamma}_1 = \overline{\gamma}_2 = 0$dB. Both FSMCs are 8-state and adopt equiprobable partition method. In this case, all the conditions in Theorem 4.15 are satisfied according to Corollary 4.16. Therefore, a_1^* is nondecreasing in (b_1, g_2), and a_2^* is nondecreasing in (b_2, g_1). In Fig. 9, we switch $\overline{\gamma}_2$ from 0dB to 3dB to breach condition (d) in Theorem 4.15. In this case, a_1^* is not monotonic in g_2. But, since Theorem 4.12 still holds, a_1^* and a_2^* are monotonic in b_1 and b_2, respectively.

Note, that the related previous studies usually placed constraints on the environments or the DP functions in order to prove the structure in the optimal policy. For example, in [19] the submodularity of the state transition probability was proved by assuming uniformly distributed traffic rates, and in [31], the strict submodularity of $Q^{(n)}$ in DP iterations was assumed to be preserved by a weight factor in the immediate cost function (however, the exact value of this factor was not given). In contrast, the basic result in this paper, Theorem 4.12, is essentially given in terms of unit costs and discount factor, the parameters in the MDP model. The practical meaning of Theorem 4.12 can be interpreted in two ways. If the unit costs and discount factor are adjustable, we can tune them to get a structured optimal policy. If they are given, we can check the sufficient conditions for the existence of a monotonic optimal policy after the MDP modeling. In addition, we also derive the results, Theorems 4.13 and 4.15 by considering the uniform traffic rates, stochastic dominance of channel transition probabilities and channel modeling, and modulation scheme in this paper. They are also applicable if the associated conditions are satisfied.

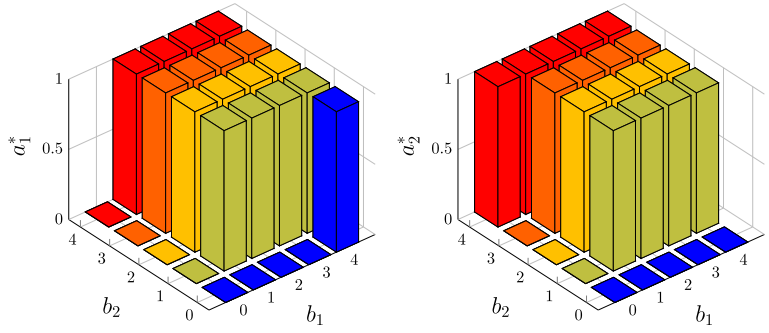

Fig. 6 The optimal action a_1^* (*left*) and a_2^* (*right*) vs. queue states b_1 and b_2 when $g_1 = 1$ and $g_2 = 5$. $p_1 = p_2 = 0.5, \lambda = 0.05, \tau = 1, \eta = 1, \xi_o = 4$, and $\beta = 0.97$. In this case, $\xi_o \geq 2\lambda + \eta + \tau$ and $\beta \leq \frac{2(\tau-\lambda)}{\tau+\eta}$. According to Corollary 4.14, Theorem 4.13 holds. Therefore, both a_1^* and a_2^* are nondecreasing in (b_1, b_2)

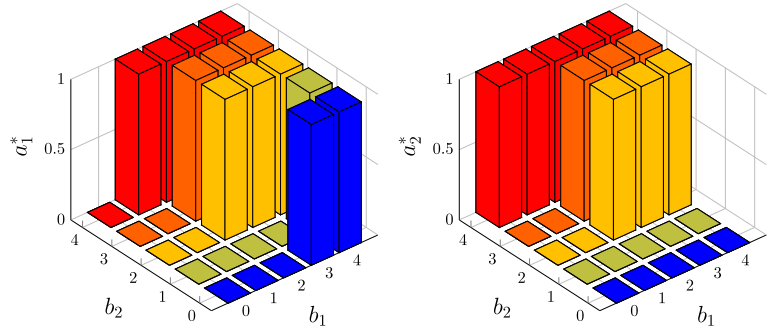

Fig. 7 The optimal action a_1^* (*left*) and a_2^* (*right*) vs. queue states b_1 and b_2 when $g_1 = 1$ and $g_2 = 5$. $p_1 = p_2 = 0.5$, $\lambda = 0.05$, $\tau = 1$, $\eta = 2$, $\xi_o = 4$, and $\beta = 0.97$. In this case, $\xi_o \geq 2\lambda + \eta + \tau$ but $\beta > \frac{2(\tau-\lambda)}{\tau+\eta}$. Theorem 4.12 holds, while Theorem 4.13 does not. As can be seen, a_1^* and a_2^* are monotonic in b_1 and b_2, respectively, but a_1^* is not monotonic in b_2

5 Low complexity algorithms

This section considers the question of how to exploit the results in Section 4 to simplify the optimization process of problem (11). For this purpose, we present MPI and DSPSA algorithms for the MDP model in Section 3.

5.1 Monotonic policy iteration

The idea of MPI is to modify (12) as

$$V^{(n)}(\mathbf{x}) = \min_{\mathbf{a}\in\mathcal{A}(\mathbf{x})} Q^{(n)}(\mathbf{x},\mathbf{a}), \forall \mathbf{x} \tag{19}$$

where $\mathcal{A}(\mathbf{x}) \subseteq \mathcal{A} = \{(0,1)\}^2 = \{(0,0),(0,1),(1,0),(1,1)\}$. Let $\theta^{(n)}(\mathbf{x}) = \arg\min_{\mathbf{a}\in\mathcal{A}(\mathbf{x})} Q^{(n)}(\mathbf{x},\mathbf{a})$. Note, $\theta^{(n)}(\mathbf{x})$ can be obtained at the same time when $V^{(n)}(\mathbf{x})$ is calculated. We express $\theta^{(n)}(\mathbf{x})$ as

$$\begin{aligned}\theta^{(n)}(\mathbf{x}) &= \theta^{(n)}(b_1,b_2,g_1,g_2)\\ &= \left(\theta_1^{(n)}(b_1,b_2,g_1,g_2), \theta_2^{(n)}(b_1,b_2,g_1,g_2)\right).\end{aligned} \tag{20}$$

Assume that Theorem 4.13 holds. We can define the action selection set $\mathcal{A}(\mathbf{x})$ as follows. Due to the $L^\natural$-convexity of $Q^{(n)}$ in (b_i,a_i), $\theta_i^{(n)}$ is always nondecreasing in b_i. Therefore, we can define $\mathcal{A}(\mathbf{x})$ as

$$\begin{aligned}\mathcal{A}(\mathbf{x}) = &\left\{a_1\in\{0,1\}: a_1 \geq \theta_1^{(n)}([b_1-1]^+,[b_2-1]^+,g_1,g_2)\right\}\\ &\times\left\{a_2\in\{0,1\}: a_2 \geq \theta_2^{(n)}([b_1-1]^+,[b_2-1]^+,g_1,g_2)\right\}\end{aligned}$$

when $b_1 \neq 0$ and $b_2 \neq 0$ and $\mathcal{A}(\mathbf{x}) = \mathcal{A}$ when $b_1 = b_2 = 0$. For example, consider the case when $g_1 = 1$ and $g_2 = 1$ at some iteration n. We need to determine the value of $\theta^{(n)}(\mathbf{x})$ for all $\mathbf{x} = (b_1,b_2,g_1,g_2)$ such that $g_1 = 1$ and $g_2 = 1$. We start with the lowest values of b_1 and b_2. For state $\mathbf{x} = (0,0,1,1)$, we have $\mathcal{A}(\mathbf{x}) = \mathcal{A} = \{0,1\}^2$. In this case, the minimization problem $\min_{\mathbf{a}\in\mathcal{A}(\mathbf{x})} Q^{(n)}(\mathbf{x},\mathbf{a})$ is equivalent to $\min_{\mathbf{a}\in\mathcal{A}} Q^{(n)}(\mathbf{x},\mathbf{a})$, i.e., we need to obtain four values of $Q^{(n)}(\mathbf{x},\mathbf{a})$ at $\mathbf{a} = (0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ to determine the minimum. If we get $\theta^{(n)}(\mathbf{x}) = (0,1)$ for $\mathbf{x} = (0,0,1,1)$, then $\mathcal{A}(\mathbf{x}) = \{0,1\}\times\{1\} = \{(0,1),(1,1)\}$ for $\mathbf{x} = (0,1,1,1)$, $(1,0,1,1)$ and $(1,1,1,1)$. It means that

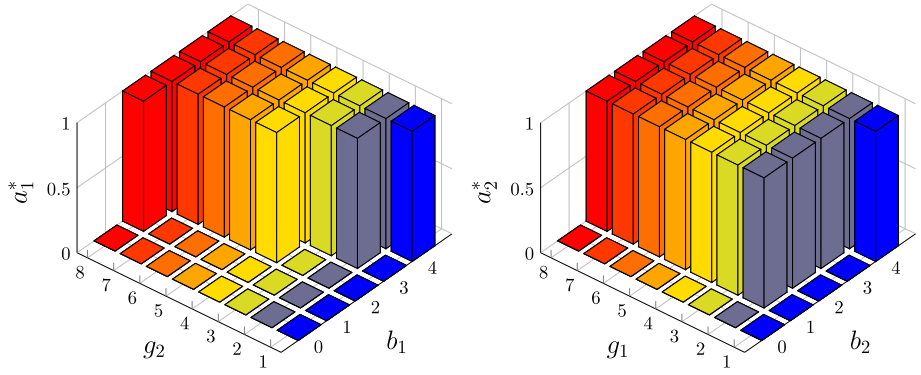

Fig. 8 The optimal action a_1^* vs. queue state b_1 and channel state g_2 when $b_2 = 3$ and $g_1 = 3$ (*left*), and a_2^* vs. b_2 and g_1 when $b_1 = 2$ and $g_2 = 1$ (*right*). $p_1 = 0.1$, $p_2 = 0.2$, $\lambda = 0.05$, $\tau = 1$, $\eta = 2$, $\xi_o = 4$ and $\beta = 0.95$. Two channels are both Rayleigh fading with $\overline{\gamma}_1 = \overline{\gamma}_2 = 0$dB and are both modeled by 8-state equiprobable FSMCs. In this case, $\beta \leq \frac{P_e(g_i)-P_e(g_i+1)}{\sum_{g_i'} P_{g_i g_i'}(P_e(g_i')-P_e(g_i'+1))}$, and according to Corollary 4.16, Theorem 4.15 holds. Therefore, a_1^* and a_2^* are nondecreasing in (b_1,g_2) and (b_2,g_1), respectively

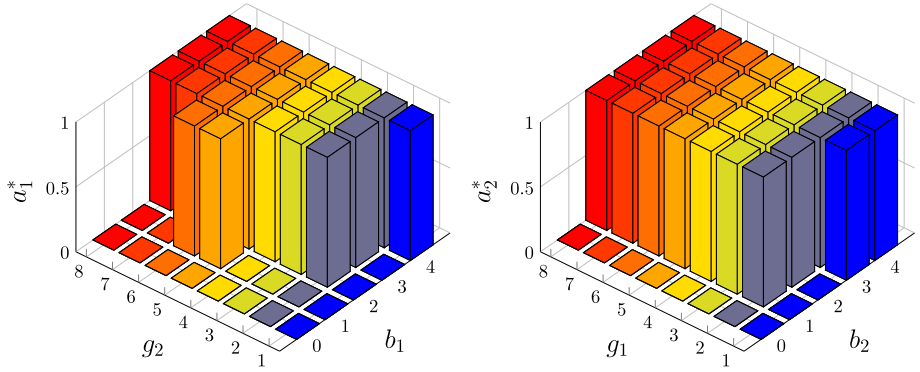

Fig. 9 The optimal action a_1^* vs. queue state b_1 and channel state g_2 when $b_2 = 3$ and $g_1 = 3$ (*left*), and a_2^* vs. b_2 and g_1 when $b_1 = 2$ and $g_2 = 1$ (*right*). $p_1 = 0.1, p_2 = 0.2, \lambda = 0.05, \tau = 1, \eta = 2, \xi_o = 4$ and $\beta = 0.95$. Two channels are both Rayleigh fading and are both modeled by 8-state equiprobable FSMCs. But, $\overline{\gamma}_1 = 0$dB and $\overline{\gamma}_2 = 3$dB. In this case, $\beta \leq \frac{P_e(g_i)-P_e(g_i+1)}{\sum_{g_i'} P_{g_i g_i'}(P_e(g_i')-P_e(g_i'+1))}$ does not hold for all g_i. We can see that a_1^* is not monotonic in g_2

only two calculations of $Q^{(n)}(\mathbf{x}, \mathbf{a})$ are required when we want to determine the value of $\min_{\mathbf{a}\in\mathcal{A}(\mathbf{x})} Q^{(n)}(\mathbf{x}, \mathbf{a})$ for these three states. In addition, if we find that $\theta^{(n)}(\mathbf{x}) = (1, 1)$ for $\mathbf{x} = (1, 1, 1, 1)$, then, for all $\mathbf{x}$ such that $b_1 > 1$, $b_2 > 1, g_1 = 1$ and $g_2 = 1$, $\mathcal{A}(\mathbf{x}) = \{(1, 1)\}$ and we can directly assign $\theta^{(n)}(\mathbf{x}) = (1, 1)$ without doing the minimization $\min_{\mathbf{a}\in\mathcal{A}(\mathbf{x})} Q^{(n)}(\mathbf{x}, \mathbf{a})$. We can find the optimal policy by repeating this process for all values of g_1 and g_2 in each iteration. From this example, it can be seen that (19) should be conducted in the increasing order of b_1 and b_2 so that the cardinality of set $\mathcal{A}(\mathbf{x})$ is progressively reducing.

5.2 Discrete simultaneous perturbation stochastic approximation

Assume Theorem 4.12 holds.[8] Due to the monotonicity of the optimal policy in queue states, the optimization problem (11) can be converted to a minimization problem over a set of queue thresholds.

For $i \in \{1, 2\}$, we define $\phi_i(b_{-i}, g_1, g_2) \in \mathcal{B}_i$ as

$$\phi_i(b_{-i}, g_1, g_2) = \min\{b_i \colon \theta_i(\mathbf{x}) = 1\}, \forall b_{-i}, g_1, g_2. \quad (21)$$

Here, $\phi_i(b_{-i}, g_1, g_2)$ is the threshold to queue i when the other user's queue state is b_{-i} and channel states are g_1 and g_2. Let $\boldsymbol{\phi}_i$ be constructed by stacking ϕ_i for all $(b_{-i}, g_1, g_2) \in \mathcal{B}_{-i} \times \mathcal{G}_1 \times \mathcal{G}_2$. The queue threshold vector is defined as $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$. In Fig. 10, we show the optimal queue threshold vector $\boldsymbol{\phi}^* = (\boldsymbol{\phi}_1^*, \boldsymbol{\phi}_2^*)$ where

$$\phi_i^*(b_{-i}, g_1, g_2) = \min\{b_i \colon \theta_i^*(\mathbf{x}) = 1\}, \forall b_{-i}, g_1, g_2 \quad (22)$$

and θ^* is the optimal policy obtained in Fig. 4. Each queue threshold vector $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ determines a deterministic policy $\theta_{\boldsymbol{\phi}}(\mathbf{x}) = (\theta_{1\boldsymbol{\phi}_1}(\mathbf{x}), \theta_{2\boldsymbol{\phi}_2}(\mathbf{x}))$ by

$$\theta_{i\boldsymbol{\phi}_i}(\mathbf{x}) = \mathcal{I}_{\{b_i \geq \phi_i(b_{-i}, g_1, g_2)\}} = \begin{cases} 1 & b_i \geq \phi_i(b_{-i}, g_1, g_2) \\ 0 & b_i < \phi_i(b_{-i}, g_1, g_2) \end{cases}. \quad (23)$$

Since θ_i^* is nondecreasing in b_i for all $i \in \{1, 2\}$ if Theorem 4.12 holds and θ^* determines $\boldsymbol{\phi}^*$ via (22), finding the optimal policy θ^* is equivalent to finding the optimal queue threshold vector $\boldsymbol{\phi}^*$. We can convert problem (11) to

$$\min_{\boldsymbol{\phi}} J(\boldsymbol{\phi}), \quad (24)$$

where

$$J(\boldsymbol{\phi}) = \sum_{\mathbf{x}^{(0)}\in\mathcal{X}} \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t C(\mathbf{x}^{(t)}, \theta_{\boldsymbol{\phi}}(\mathbf{x}^{(t)}))|\mathbf{x}^{(0)}\right]. \quad (25)$$

The advantage of formulating problem (24) is that the solutions can be approximated by the DSPSA algorithm [32] presented in Algorithm 1. The parameters/functions in this algorithm are explained as follows

- $\hat{J}(\boldsymbol{\phi})$ is an estimation of J at $\boldsymbol{\phi}$ that is obtained by simulation. The method is to simulate the state sequence $\{\mathbf{x}^{(t)}\}$ governed by the transition probability $Pr\left(\mathbf{x}^{(t+1)}|\mathbf{x}^{(t)}\right) = P^{\theta_{\boldsymbol{\phi}}(\mathbf{x}^{(t)})}_{\mathbf{x}^{(t)}\mathbf{x}^{(t+1)}}$ for all $\mathbf{x}^{(0)} \in \mathcal{X}$. $\hat{J}(\boldsymbol{\phi})$ is obtained as
$$\hat{J}(\boldsymbol{\phi}) = \sum_{\mathbf{x}^{(0)}\in\mathcal{X}} \sum_{t=0}^{T} \beta^t C\left(\mathbf{x}^{(t)}, \theta_{\boldsymbol{\phi}}(\mathbf{x}^{(t)})\right). \quad (26)$$
Each simulation stops if the increments over several successive decision epochs blow a small threshold (10^{-5}), i.e., the simulation length is finite.
- The step size parameters A, B, and α are crucial for the convergence performance of DSA algorithms. In this paper, we set as $A = 0.3$, $B = 100$, and $\alpha = 0.602$. These values are found by adopting the method suggested in [33] for practical problems

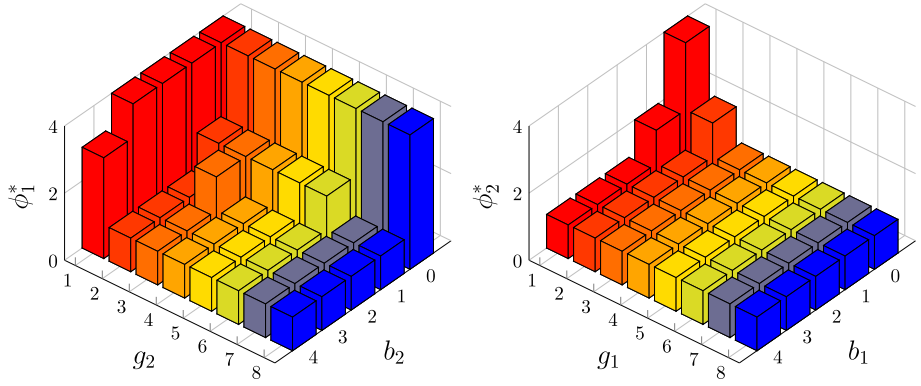

Fig. 10 The optimal threshold $\boldsymbol{\phi}_1^*$ vs. b_2 and g_2 when $g_1 = 2$ (*left*), and $\boldsymbol{\phi}_2^*$ vs. b_1 and g_1 when $g_2 = 1$ (*right*). The system parameters are the same as in Fig. 4 so that Theorem 4.12 holds

where the computation budget N, the total number of iterations, is fixed: $B = 0.095N$, $\alpha = 0.602$ and A is chosen so that $A/(B+1)^{\alpha} \| \mathbf{d}\left(\tilde{\boldsymbol{\phi}}^{(0)}\right) \| = 0.1$.

The DSPSA algorithm is a in fact a line search algorithm. It starts with any initial guess $\boldsymbol{\phi}^{(0)}$, say $\boldsymbol{\phi}^{(0)} = \mathbf{0}$, and iteratively updates the guess by the estimated descent direction $-a^{(n)}\mathbf{d}(\boldsymbol{\phi}^{(n)})$. The gradient $\mathbf{d}(\boldsymbol{\phi}^{(n)})$ in each iteration is obtained based on two values of $\hat{J}$, $\hat{J}\left(\lfloor \boldsymbol{\phi}^{(n)} \rfloor + \frac{\mathbf{1}+\Delta}{2}\right)$ and $\hat{J}\left(\lfloor \boldsymbol{\phi}^{(n)} \rfloor + \frac{\mathbf{1}-\Delta}{2}\right)$.[9] According to a study in [34], the estimation sequence $\{\boldsymbol{\phi}^{(n)}\}$ slowly converges to the optimal queue threshold vector $\boldsymbol{\phi}^*$.

Algorithm 1: DSPSA [32]

input : initial guess $\boldsymbol{\phi}^{(0)}$, total number of iterations N, step size parameters A, B and α

output: $[\boldsymbol{\phi}^{(N)}]$, the closest integer point to $\boldsymbol{\phi}^{(N)}$ by Euclidean distance.

begin

  for $n = 0$ **to** N **do**

    $a^{(n)} = \frac{A}{(B+n+1)^{\alpha}}$;

    Generate $\Delta = (\Delta_1, \ldots, \Delta_D)$ with each tuple $\Delta_d \in \{-1, 1\}$ being independent Bernoulli random variable with probability 0.5. Obtain $\mathbf{d}(\boldsymbol{\phi}^{(n)})$ with the ith entry being

$$d_i(\boldsymbol{\phi}^{(n)}) = \left(\hat{J}\left(\lfloor \boldsymbol{\phi}^{(n)} \rfloor + \frac{\mathbf{1}+\Delta}{2}\right) - \hat{J}\left(\lfloor \boldsymbol{\phi}^{(n)} \rfloor + \frac{\mathbf{1}-\Delta}{2}\right)\right)\Delta_i^{-1}, \quad (27)$$

    where $\lfloor \mathbf{x} \rfloor$ is the largest integer less than $\mathbf{x}$;

    $\boldsymbol{\phi}^{(n+1)} = \boldsymbol{\phi}^{(n)} - a^{(n)}\mathbf{d}(\boldsymbol{\phi}^{(n)})$;

  endfor

end

5.3 Complexity

MPI is in fact a modified DP algorithm that exploits $L^{\natural}$-convexity or submodularity of $Q^{(n)}$. It converges at the same rate as DP. But, since $\mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}$ and $|\mathcal{A}(\mathbf{x})|$, the cardinality of $\mathcal{A}(\mathbf{x})$, is progressively decreasing in b_1 and b_2, the complexity in each iteration in MPI is lower than that in DP. Let ρ be the average size of $\mathcal{A}(\mathbf{x})$ over all states $\mathbf{x}$. The complexity in one iteration of MPI is $O(|\mathcal{X}|^2\rho)$, where $\rho \leq |\mathcal{A}|$. The exact value of ρ varies with different systems. To show the examples of the actual complexity of MPI, we do the following experiment. We use the same system settings as in Fig. 6 and set the number of channel states of both channels to K, i.e., $K_1 = K_2 = K$. We vary K from 2 to 10. For each value of K, we run both DP and MPI and obtain the complexity as the number of calculations of $Q^{(n)}$ averaged over iterations. The results are shown in Fig. 11. It can be seen that the complexity of MPI is always less than that of DP, and MPI alleviates the

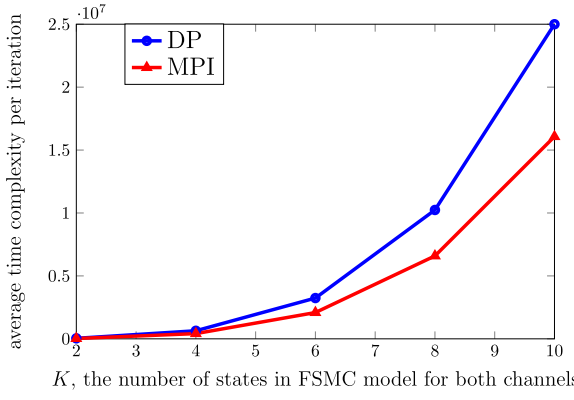

Fig. 11 The mean time complexity per iteration of DP and MPI, where $K_1 = K_2 = K$ and the value of K varies from 2 to 10. The other system parameters are the same as in Fig. 6. The complexity is obtained as the number of calculations of $Q^{(n)}$ averaged over iterations

drastically growing complexity in DP when the size of the channel state space grows large.

Consider the complexity of the DSPSA algorithm. Let ζ be the complexity of obtaining the value of $\hat{J}$ by simulation. Since we only need two values of $\hat{J}$ to calculate the gradient $\mathbf{d}$, the complexity in each iteration of DSPSA is $O(\zeta)$. But, the convergence rate depends on the parameters of the DSPSA algorithm [35], e.g., the step size parameters, and may vary with different MDP systems, i.e., DSPSA may converge slower than DP or MPI. However, we have two advantages of implementing DSPSA algorithm over DP or MPI. One is that DSPSA is a simulation-based algorithm, the runs of which do not require the full knowledge of the MDP model. Based on (26), to obtain $\hat{J}$, we only require the knowledge of the state space $\mathcal{X}$ and a simulation model that can generate a state sequence $\{\mathbf{x}^{(t)}\}$ based on a given queue threshold vector $\boldsymbol{\phi}$ and the statistics of packet arrival and channel variation processes. If the packet arrival probabilities and/or channel statistics change suddenly, the optimal policy will change accordingly, and DSPSA algorithm can adapt slowly to the new optimal policy.

The results in Fig. 12 are based on an experiment of DSPSA in an environment where the system parameters change with time. The relay is assumed to serve the first pair of users with packet arrival probabilities being $p_1 = 0.1$ and $p_2 = 0.2$ in the first 500 iterations and serve another pair of users with $p_1 = 0.8$ and $p_2 = 0.2$ in the second 500 iterations. It can be seen that DSPSA is able to adaptively track the optimum and optimizer of problem (24). In contrast, to run DP or MPI, we require the full knowledge of the MDP model. If the statistics of packet arrival and channel variation processes change, we need to determine the new MDP model by calculating all values of the state transition probability $P^{\mathbf{a}}_{\mathbf{x}\mathbf{x}'}$ before running DP or MPI. Alternatively speaking, MPI and DP are model-based algorithms while DSPSA is a model-free algorithm [36].

The other advantage of DSPSA is that it allows the scheduler to learn the optimal policy online. For example, assume that we start with any arbitrary threshold vector $\boldsymbol{\phi}^{(0)}$. We first let the scheduler adopt the policy that is determined by the queue threshold vector $\lfloor\boldsymbol{\phi}^{(0)}\rfloor + \frac{\mathbf{1}+\Delta}{2}$ (via (21)) for a while and obtain the value of $\hat{J}\left(\lfloor\boldsymbol{\phi}^{(0)}\rfloor + \frac{\mathbf{1}+\Delta}{2}\right)$ based on the actual immediate costs incurred. Then, we let the scheduler adopt the policy that is determined by $\lfloor\boldsymbol{\phi}^{(0)}\rfloor + \frac{\mathbf{1}-\Delta}{2}$ for a while and obtain the value of $\hat{J}\left(\lfloor\boldsymbol{\phi}^{(0)}\rfloor + \frac{\mathbf{1}-\Delta}{2}\right)$. By doing so, the gradient $\mathbf{d}$ can be calculated, and we update $\boldsymbol{\phi}^{(0)}$ and get a new queue vector $\boldsymbol{\phi}^{(1)}$. By repeating this process, the scheduler can slowly update the estimation $\boldsymbol{\phi}^{(n)}$ towards $\boldsymbol{\phi}^*$ and hence find the the optimal policy θ^*.

It should be noted that low complexity algorithms for searching or approximating the optimal policy θ^* are not restricted to MPI and DSPSA. With the results on monotonicity derived in Section 4, one can propose other algorithms, e.g., the random search method [37], the simulated annealing method [38], the complexity of which could be even lower than MPI and DSPSA. For example, the random search method [37] can be applied to find the solution of the multivariate minimization problem (24). In this method, the descent direction is found by random sampling in each iteration. The complexity incurred by random sampling could be lower than that incurred by simulation (as in DSPSA). But, we still need to compare the convergence rates of the random search and DSPSA algorithms. In summary, the MPI and DSPSA are two examples of low complexity algorithms that are based on the monotonicity of the optimal policy. To propose more low complexity algorithms and compare the convergence

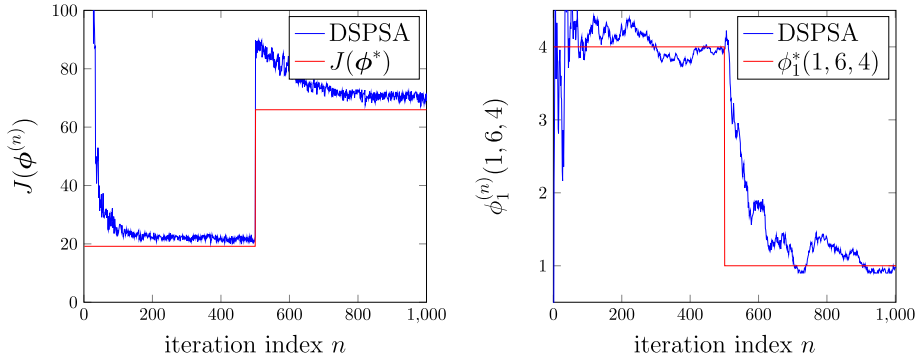

Fig. 12 Convergence performance of DSPSA: $J(\boldsymbol{\phi}^{(n)})$, the value of the objective function at the nth iteration of DSPSA (left); $\phi_1^{(n)}(1,6,4)$, the estimations of the optimal threshold to queue 1 when $b_2 = 1$, $g_1 = 6$ and $g_2 = 4$ (right). The channels are both Rayleigh fading with $\overline{\gamma}_1 = \overline{\gamma}_2 = 0$dB and modeled by 8-state FSMCs. The system parameters are set as $p_1 = 0.1, p_2 = 0.2, \lambda = 0.05, \tau = 1, \eta = 2, \xi_o = 4$ and $\beta = 0.97$ in the first 500 iterations. Then p_1 and p_2 are changed to 0.8 and 0.9, respectively, in the second 500 iterations. The optimal value of queue threshold vector $\boldsymbol{\phi}^*$ is determined via (22) by using the optimal policy θ^* found by DP

performance are out of the scope of this paper and could be one of the future directions of research.

5.4 Simulation results

We run simulations in an NC-TWRC with Rayleigh fading channels. Let DP-MDP-QC be the optimal policy searched by DP based on the MDP model in Section 3. We compare the performance of DP-MDP-QC to the following four policies:

- DSPSA-MDP-QC: This policy is searched by DSPSA based on the MDP model in Section 3 with the total number of iterations being $N = 1000$. As explained in Section 5.2, the estimation sequence produced by the DSPSA algorithm should be slowly converging to the optimal policy. Therefore, DSPSA-MDP-QC should be close to DP-MDP-QC (in Euclidian distance) and the performance of DSPSA-MDP-QC should be similar to that of DP-MDP-QC.
- MYO-QC: This policy is obtained by $\theta_{\text{MYO-QC}}(\mathbf{x}) = \arg\min_{\mathbf{a}} C(\mathbf{x}, \mathbf{a})$, where $C(\mathbf{x}, \mathbf{a})$ is the immediate cost function as defined in (9). Recall that policy DP-MDP-QC searched by DP is $\theta^*(\mathbf{x}) = \theta^N(\mathbf{x}) = \arg\min_{\mathbf{a}} C(\mathbf{x}, \mathbf{a}) + \beta \sum_{\mathbf{x}'} P^{\mathbf{a}}_{\mathbf{x}\mathbf{x}'} V^{(N-1)}(\mathbf{x})$, where N is the iteration index when DP converges. MYO-QC is the policy that neglects the aftermath $\beta \sum_{\mathbf{x}'} P^{\mathbf{a}}_{\mathbf{x}\mathbf{x}'} V^{(N-1)}(\mathbf{x})$ that is incurred by the action taken at the current decision epoch. Alternatively speaking, MYO-QC is myopic while DP-MDP-QC is far-sighted.[10] In a stochastic environment, myopic policies usually incur a higher expected long-term cost than far-sighed ones.
- AT: This policy is denoted as $\theta_{\text{AT}}(\mathbf{x}) = (\theta_{\text{AT}1}(\mathbf{x}), \theta_{\text{AT}2}(\mathbf{x}))$ where $\theta_{\text{AT}i}(\mathbf{x}) = 1$ if $b_i \neq 0$, i.e., always transmit whenever queue i is not empty. This policy minimizes the costs incurred by the packet delay and queue overflow. But, the performance of this policy should not be as good as DP-MDP-QC if the purpose is to minimize the long-term cost incurred by not only the packet delay and queue overflow but also the transmission power consumption and downlink transmission error rate.
- DP-MDP-Q: This policy is determined by DP based on an MDP model that is the same as the one in Section 3 except that the immediate cost function is defined as $C(\mathbf{x}, \mathbf{a}) = C_h(\mathbf{b}, \mathbf{a}) + t_r(\mathbf{a})$. This policy was proposed in [12], where the authors assume that the channels are lossless so that the packet error cost $\sum_{i=1}^{2} \text{err}(g_{-i}, a_i) = 0$ always. However, the wireless channels are usually not ideal in practice. If we adopt policy DP-MDP-Q, it should incur a higher downlink transmission error rate than DP-MDP-QC.

We fix $p_2 = 0.5$ and vary p_1 from 0.2 to 0.6. The other system parameters are the same as in Fig. 4. A simulation lasting for 10^5 decision epochs is run for each value of p_1. Each packet contains 100 bits, i.e., the packet length $L_P = 100$. We obtain the number of holding and overflowing packets and the number of transmissions averaged over decision epochs. The former indicates the mean packet delay and queue overflow costs, and the latter indicates the average transmission power consumption. The transmission error rate is calculated as the ratio of the number of erroneous bits received to the total number of bits sent. We also obtain the immediate cost averaged over decision epochs, which indicates the long-term cost (the minimand in (11)). The results are presented in Fig. 13. It can be seen that the average immediate cost of DSPSA-MDP-QC almost overlaps with that of DP-MDP-QC. It means that if we allow the total number of iterations in the DSPSA algorithm large enough, e.g., 1000 iterations, it is able to converge to a policy that is very close to DP-MDP-QC.

For policy MYO-QC, we can see that it always incurs a greater number of transmissions and holding and overflow packets and a higher transmission error rate than DP-MDP-QC. The average immediate cost of this policy is at least 0.23 higher than those of DP-MDP-QC, which is the worst among all policies. Therefore, a far-sighted policy outperforms a myopic one when we want to minimize the long-term cost in a stochastic system.

The number of holding and overflow packets incurred by policy AT is always zero. However, it results in the highest number of transmissions and transmission error rate. The average immediate cost incurred by AT is at least 0.09 higher than DP-MDP-QC, which justifies our expectation; AT minimizes the packet delay and queue overflow costs but incurs higher transmission power consumption and downlink transmission error rate. Therefore, the long-term cost incurred by AT is not as low as that incurred by DP-MDP-QC. For policy DP-MDP-Q, the number of transmissions is almost the same as DP-MDP-QC, and the number of holding and overflow packets is even lower than DP-MDP-QC. However, since this policy assumes that the wireless channels are ideal (but they are in fact not), the transmission error rate is about 1.3 times higher than DP-MDP-QC (almost as high as AT). Therefore, the average immediate cost is still higher than DP-MDP-QC. In summary, in a stochastic environment where the long-term loss can be incurred by multiple causes, the policy that considers all such causes simultaneously outperforms those that only consider some and neglects others.

6 Conclusion

This paper studied an MDP-modeled transmission control problem in NC-TWRC with random traffic and fading channels. The purpose was to prove the existence of a

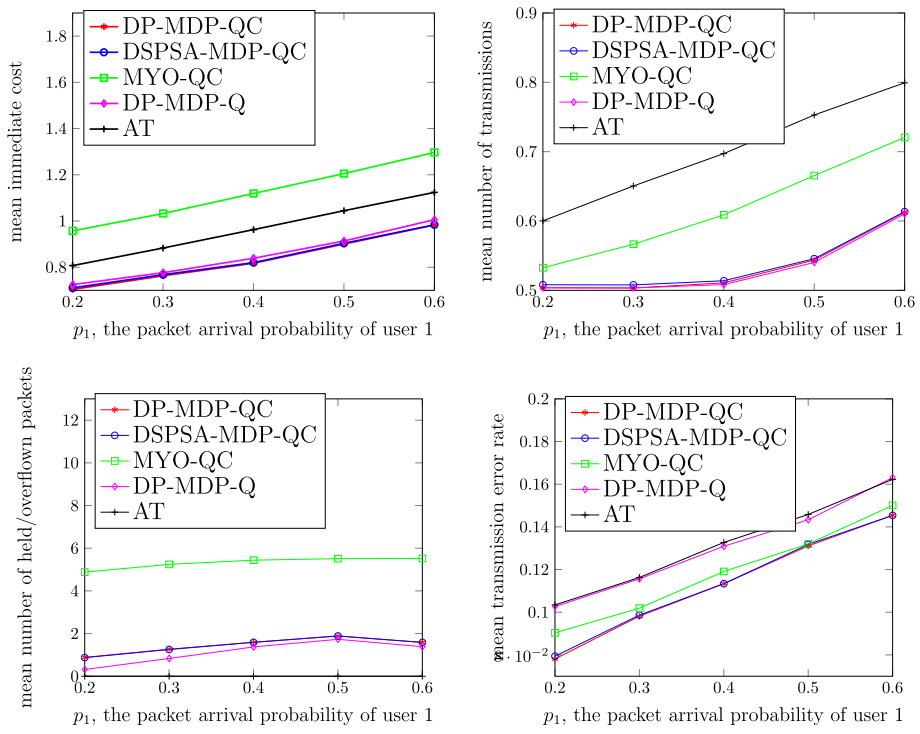

Fig. 13 Simulation results: the mean immediate cost $C(\mathbf{x}, \mathbf{a})$ (*top left*); the mean number of transmissions $\boldsymbol{\mathcal{I}}_{\{a_1=1 \text{ or } a_2=1\}}$ (*top right*); the mean number of holding packets $\min\{[y_i]^+, L_i\}$ (*bottom left*); the mean number of lost packets due to queue overflow $\boldsymbol{\mathcal{I}}_{\{[y_i]^+=L_i+1\}}$, and the mean transmission error rate (*bottom right*). These are the values averaged over 10^5 decision epochs. The channels are both Rayleigh fading with $\overline{\gamma}_1 = \overline{\gamma}_2 = 0$dB and modeled by 8-state FSMCs. The system parameters are set as $p_2 = 0.5$, $\lambda = 0.05$, $\tau = 1$, $\eta = 2$, $\xi_o = 4$ and $\beta = 0.97$. p_1 is varying from 0.2 to 0.6

monotonic optimal transmission policy that minimized packet delay, queue overflow, transmission power, and the downlink transmission error rate in the long run. We proved that the optimal policy is nondecreasing in queue and/or channel states by investigating how certain properties (submodularity, $L^\natural$-convexity and multimodularity) varied with the system parameters. Based on these properties of DP, we presented two low-complexity algorithms, MPI and DSPSA.

As a part of the conclusion, we point out two directions for the research work in the future. The structured results derived in Section 4 can be used to design model-free learning algorithms, e.g., monotonic Q-learning. Since queue-assisted transmission control is also used in cross-layer variable-rate adaptive modulation problems, it would be of interest if we can use submodularity, $L^\natural$-convexity, and multimodularity to establish the sufficient conditions for the existence of monotonic optimal transmission policies in these systems.

Endnotes

[1]The complexity of the algorithm grows drastically with the cardinality of the system variables [16].

[2]The definition of Pareto optimality is given in Appendix A. In Section 3.5, we will explain the Pareto optimality of the optimal policy of MDP.

[3]In this paper, we use $\epsilon = 10^{-5}$.

[4]See the definition of Pareto optimality and description of scalarization technique in Appendix A. The Pareto optimality of θ^* has also been discussed in [31].

[5]$f\colon \mathbb{Z}^n \to \mathbb{R}_-$ is (strictly) supermodular if $-f$ is (strictly) submodular.

[6]The interpretation of $\xi_o \geq 2\lambda + \eta + \tau$ is that the cost of overflowing a packet is greater than or equal to the sum of the cost of holding two packets, the cost when transmission error rate is increased by η and the cost of missing a coding opportunity.

[7]In [21], integer convexity was used to denote the one dimensional discrete convexity as explained in Lemma B.1(b).

[8]According to the conditions in Theorems 4.12, 4.13 and 4.15, Theorem 4.12 is straightforwardly satisfied if either Theorem 4.13 or Theorem 4.15 holds. Therefore, if DSPSA can be applied when Theorem 4.12 holds, it can be also applied when Theorem 4.13 and 4.15 hold.

[9]The gradient $\mathbf{d}$ in (27) is defined based on the discrete mid-point convexity [32].

[10]More comparisons of far-sighted and myopic policies in NC-TWRC are presented in [13].

[11]The one dimensional discrete convex function $h : \mathbb{Z} \to \mathbb{R}$ satisfies $h(x+1) + h(x-1) - 2h(x) \geq 0$ for

all $x \in \mathbb{Z}$. Moreover, by Definition 4.4 and 4.5, h is both $L^\natural$-convex and multimodular.

[12]A function $f: \mathbb{Z}^2 \to \mathbb{R}_+$ is multimodular if and only if it is (1) supermodualr: $\Delta_i \Delta_j f(\mathbf{x}) \geq 0$ and (2) superconvex: $\Delta_i f(\mathbf{x} + \mathbf{e}_i) \geq \Delta_i f(\mathbf{x} + \mathbf{e}_j)$ for all $i,j \in \{1,2\}$, where $\Delta_i f(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x} - \mathbf{e}_i)$ and $\mathbf{e}_i \in \mathbb{Z}^2$ is a 2-tuple with all zero entries except the ith entry being one.

Appendix A

In multi-objective optimization [39], there are N optimization metrics. Each of then can be quantified by a loss function $f_n: \mathbb{R}^M \to \mathbb{R}$. The problem can be expressed as

$$\min_{\mathbf{x}\in\mathbb{R}^M} (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_N(\mathbf{x})), \tag{28}$$

where $\mathbf{x}$ is the decision vector. We say $\mathbf{x}$ Pareto dominates $\mathbf{x}'$ if $f_n(\mathbf{x}) \leq f_n(\mathbf{x}')$ for all $n \in \{1,\ldots,N\}$. We call $\mathbf{x}^*$ a Pareto optimal decision vector if no $\mathbf{x} \in \mathbb{R}^M$ Pareto dominates $\mathbf{x}^*$. In a multi-objective optimization problem, we always want to seek a Pareto optimal solution. One way to solve this problem is called scalarization technique. The idea is to convert (28) to a single-objective problem

$$\min_{\mathbf{x}\in\mathbb{R}^M} w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) + \ldots + w_N f_N(\mathbf{x})), \tag{29}$$

where $w_n > 0$ is the weight. It is shown that the solution of problem (29) is a Pareto optimal solution of problem (28) in [39]. Note, based on the definition of Pareto optimality, a Pareto optimal solution is not an optimal solution if we purely consider only one optimization metric.

Appendix B

Lemma B.1. *submodularity, $L^\natural$-convexity and multimodularity has the following properties:*

(a) *If $f_i: \mathbb{Z}^n \to \mathbb{R}_+$ is submodular/$L^\natural$-convex/multimodular in $\mathbf{x} \in \mathbb{Z}^n$ and $\alpha_i \geq 0$ for all i, then $\sum_{i=1}^m \alpha_i f_i(\mathbf{x})$ is submodular/$L^\natural$-convex/multimodular in $\mathbf{x}$.*

(b) *If $h: \mathbb{Z} \to \mathbb{R}_+$ is convex[11], then $f(\mathbf{x}) = h(x_1 - x_2)$ is $L^\natural$-convex in $\mathbf{x} = (x_1, x_2)$ and $g(\mathbf{x}) = h(x_1 + x_2)$ is multimodular in $\mathbf{x} = (x_1, x_2)$.*

(c) *Let d be a random variable. If $g(\mathbf{x}, d)$ is $L^\natural$-convex/multimodular in $\mathbf{x} \in \mathbb{Z}^n$ for all d, then $\mathbb{E}_d[g(\mathbf{x}, d)]$ is $L^\natural$-convex/multimodular in $\mathbf{x}$.*

(d) *If $f: \mathbb{Z}^n \to \mathbb{R}_+$ is $L^\natural$-convex, then $\psi(\mathbf{x}, \zeta) = f(\mathbf{x} - \zeta \mathbf{1})$ is $L^\natural$-convex in $(\mathbf{x}, \zeta)$.*

Proof. The proofs of (a), (b), and (d) can be found in [26, 28, 29]. We show proof of (c). Consider function f first. Since $\psi(\mathbf{x}, \zeta) = f(\mathbf{x} - \zeta \mathbf{1}) = h(x_1 - x_2)$, according to Definition 4.4, it suffices to show the submodularity of h in (x_1, x_2). But, because of the convexity of h,

$$\begin{aligned} &h(x_1 + 1 - x_2) + h(x_1 - (x_2 + 1)) - h(x_1 - x_2) h(x_1 + 1 - (x_2 + 1)) \\ &= h(x_1 - x_2 + 1) + h(x_1 - x_2 - 1) - 2h(x_1 - x_2) \geq 0. \end{aligned} \tag{30}$$

By Definition 4.2, h is submodular in (x_1, x_2). Therefore, $f(\mathbf{x}) = h(x_1 - x_2)$ is $L^\natural$-convex in (x_1, x_2). Since, $g(\mathbf{x}) = f(-M_{2,1}\mathbf{x})$, according to Lemma 4.7(a), $g(\mathbf{x})$ is multimodular in (x_1, x_2). □

Appendix C

$\tilde{C}$ is nondecreasing in y_1 because h_1 is nondecreasing in y_1. By assuming that $V^{(n-1)}$ is nondecreasing in b_1', we have $\tilde{Q}^{(n)}$ nondecreasing in y_1 since $\min\{[y_i]^+, L_i\} + f_i$ is nondecreasing in y_i. Next, consider the multimodularity by using Proposition 1[12] in [40]. The supermodulariry and superconvexity of $\tilde{C}$ in (y_1, a_1) can be proved by the convexity of h_1. So, $\tilde{C}$ is multimodular in (y_1, a_1). Assume the monotonicity and $L^\natural$-convexity of $V^{(n-1)}$ in b_1'. $\tilde{Q}$ is supermodular and superconvex in (y_1, a_1) because

$$\begin{aligned} &\tilde{Q}^{(n)}(\mathbf{y}, \mathbf{g}, \mathbf{a}) + \tilde{Q}^{(n)}(\mathbf{y} + \mathbf{e}_1, \mathbf{g}, \mathbf{a} + \mathbf{e}_1) \\ &\qquad - \tilde{Q}^{(n)}(\mathbf{y} + \mathbf{e}_1, \mathbf{g}, \mathbf{a}) - \tilde{Q}^{(n)}(\mathbf{y}, \mathbf{g}, \mathbf{a} + \mathbf{e}_1) = 0, \end{aligned}$$

$$\begin{aligned} &\tilde{Q}^{(n)}(\mathbf{y} + \mathbf{e}_1, \mathbf{g}, \mathbf{a}) - \tilde{Q}^{(n)}(\mathbf{y}, \mathbf{g}, \mathbf{a}) - \tilde{Q}^{(n)}(\mathbf{y}, \mathbf{g}, \mathbf{a} + \mathbf{e}_1) + \tilde{Q}^{(n)}(\mathbf{y} - \mathbf{e}_1, \mathbf{g}, \mathbf{a} + \mathbf{e}_1) \\ &= \begin{cases} \lambda + \beta \mathbb{E}_{\mathbf{g}'}\left[V_{\mathbf{f}}^{(n-1)}(\mathbf{y} + \mathbf{e}_1, \mathbf{g}') - V_{\mathbf{f}}^{(n-1)}(\mathbf{y}, \mathbf{g}') \middle| \mathbf{g} \right] \geq 0 & y_1 = 0 \\ \xi_o - \lambda + \beta \mathbb{E}_{\mathbf{g}'}\left[-V_{\mathbf{f}}^{(n-1)}(\mathbf{y}, \mathbf{g}') + \right. & \\ \quad \left. V_{\mathbf{f}}^{(n-1)}(\mathbf{y} - \mathbf{e}_1, \mathbf{g}') \middle| \mathbf{g} \right] \geq 0 & y_1 = L_1 \\ \beta \mathbb{E}_{\mathbf{g}'}\left[V_{\mathbf{f}}^{(n-1)}(\mathbf{y} + \mathbf{e}_1, \mathbf{g}') - \right. & \\ \quad \left. 2V_{\mathbf{f}}^{(n-1)}(\mathbf{y}, \mathbf{g}') + V_{\mathbf{f}}^{(n-1)}(\mathbf{y} - \mathbf{e}_1, \mathbf{g}') \middle| \mathbf{g} \right] \geq 0 & \text{otherwise} \end{cases}. \end{aligned} \tag{31}$$

The second inequality in (31) (when $y_1 = L_1$) is explained as follows. Recall that we have $V^{(n-1)}(\mathbf{x}') = Q^{(n-1)}\left(\mathbf{x}', \mathbf{a}^*(\mathbf{x}')\right)$, where $Q^{(n-1)}$ is $L^\natural$-convex in (b_1', a_1') and $\mathbf{a}^*(\mathbf{x}') = \arg\min_{\mathbf{a}'} Q^{(n-1)}(\mathbf{x}', \mathbf{a}')$ is nondecreasing in b_1'. It can be shown that

$$\begin{aligned} &- V^{(n-1)}(b_1', b_2', \mathbf{g}') + V^{(n-1)}(b_1' - 1, b_2', \mathbf{g}') \\ &= -Q^{(n-1)}(b_1', b_2', \mathbf{g}', \mathbf{a}^*(b_1', b_2', \mathbf{g}')) \\ &\quad + Q^{(n-1)}(b_1' - 1, b_2', \mathbf{g}', \mathbf{a}^*(b_1' - 1, b_2', \mathbf{g}')) \\ &\geq -Q^{(n-1)}(b_1', b_2', \mathbf{g}', (1,1)) + Q^{(n-1)}(b_1' - 1, b_2', \mathbf{g}', (0,0)) \\ &\geq -C(b_1', b_2', \mathbf{g}', (1,1)) + C(b_1' - 1, b_2', \mathbf{g}', (0,0)) \\ &\geq -\lambda - \eta - \tau. \end{aligned} \tag{32}$$

Since $\xi_o \geq 2\lambda + \eta + \tau$, we have the inequality when $y_1 = L_1$ in (31). Therefore, $\tilde{Q}$ is multimodular in (y_1, a_1). The monotonicity and multimodularity of $\tilde{C}$ and $\tilde{Q}^{(n)}$ in (y_2, a_2) can be proved in the same way.

Appendix D

By knowing the $L^\natural$-convexity of $V^{(n-1)}$ in $\mathbf{b}'$, we have

$$\begin{aligned}
&Q^{(n)}(\mathbf{x},(1,0)) + Q^{(n)}(\mathbf{x},(0,1)) - Q^{(n)}(\mathbf{x},(0,0)) - Q^{(n)}(\mathbf{x},(1,1)) \\
&= \tau + \beta \mathbb{E}_{\mathbf{g}'} \Big[V_{\mathbf{f}}^{(n-1)}(\mathbf{b} - \mathbf{e}_1, \mathbf{g}') + V_{\mathbf{f}}^{(n-1)}\left(\mathbf{b} - \mathbf{e}_2, \mathbf{g}'\right. \\
&\qquad \left. - V_{\mathbf{f}}^{(n-1)}(\mathbf{b}, \mathbf{g}') - V_{\mathbf{f}}^{(n-1)}(\mathbf{b} - \mathbf{e}_1 - \mathbf{e}_2, \mathbf{g}')\Big| \mathbf{g} \right] \\
&\geq \tau > 0, \qquad (33)
\end{aligned}$$

i.e., $-Q^{(n)}$ is strictly supermodular in $\mathbf{a}$ for all $\mathbf{x}$. By definition in [30], the game is supermodular.

Appendix E

C_h is $L^\natural$-convex in $(\mathbf{b}, \mathbf{a})$ because of the convexity of h_i, and C_t is $L^\natural$-convex in $(\mathbf{b}, \mathbf{a})$ because of the submodularity of t_r in $\mathbf{a}$. By Lemma B.1(a), C is $L^\natural$-convex in $(\mathbf{b}, \mathbf{a})$. Consider the $L^\natural$-convexity of Q in $(\mathbf{b}, \mathbf{a})$. Let $BR_i(a_{-i}) = \arg\min_{a_i} Q^{(n)}(\mathbf{x}, (a_i, a_{-i}))$. Equilibria $(0,0)(1,1)$ implies $BR_i(a_{-i}) = a_{-i}$, i.e., $a_1 = a_2$. But, $Q^{(n)}(\mathbf{x}, (a_1, a_1))$ is $L^\natural$-convex in $(\mathbf{b}, a_1)$ since: When $b_i - a_1 < L_i + 1$ for all $i \in \{1,2\}$, $Q^{(n)}$ is $L^\natural$-convex in $(\mathbf{b}, a_1)$ because of the $L^\natural$-convexity of $V^{(n-1)}$ in $\mathbf{b}'$ and Lemma B.1(c) and (d); When $b_i - a_1 = L_i + 1$ for either $i = 1$ or $i = 2$, the $L^\natural$-convexity of $Q^{(n)}$ can be shown in the same way as in Appendix C under condition $\xi_o \geq 2\lambda + \tau + \eta$. By Theorem 4.10 and Proposition 4.11, $V^*(\mathbf{x})$ is $L^\natural$-convex in $\mathbf{b}$ and the optimal action $\mathbf{a}^*$ is nondecreasing in $\mathbf{b}$.

Appendix F

We just need to show that condition (b) in Theorem 4.13 is satisfied. Let $b_i - a_1 < L_i + 1$ for all $i \in \{1,2\}$. It suffices to show $BR_i(a_{-i}) = a_{-i}$ for all $i \in \{1,2\}$ in order to prove equilibria $(0,0)(1,1)$ in Theorem 4.13. Because the game has strictly supermodular utility, $BR_i(a_{-i}+1) > BR_i(a_{-i})$. So $BR_i(1) = 1$, if we can prove $BR_i(0) = 0$. By knowing that $p_1 = 0.5$, we can show that

$$\begin{aligned}
&Q^{(n)}(\mathbf{b}, \mathbf{g}, (1,0)) - Q^{(n)}(\mathbf{b}, \mathbf{g}, (0,0)) \\
&= \begin{cases} -\lambda + \tau + \eta P_e(g_1) + \\ \quad 0.5\beta\Big(V(b_1 - 1, \hat{b}_2), \mathbf{g}') - V(b_1, \hat{b}_2), \mathbf{g}')\Big) \geq 0 & 0 < b_1 < L_1 + 1 \\ -\lambda + \tau + \eta P_e(g_1) \geq 0 & \text{otherwise} \end{cases}, \qquad (34)
\end{aligned}$$

where $\hat{b}_2 = \min\{[b_2]^+, L_2\} + f_2$ and the inequality in the case when $0 < b_1 < L_1 + 1$ is because that, by a similar approach as in (32), $V^{(n-1)}(b_1' - 1, b_2', \mathbf{g}') - V^{(n-1)}(b_1' + 1, b_2', \mathbf{g}') \geq -\tau - \eta$ and $\beta \leq \frac{2(\tau - \lambda)}{\tau + \eta}$.

Similarly, we can show $Q^{(n)}(\mathbf{b}, \mathbf{g}, (0,1)) - Q^{(n)}(\mathbf{b}, \mathbf{g}, (0,0)) \geq 0$ in the case when $p_2 = 0.5$. So, $BR_i(a_{-i}) = a_{-i}$.

Appendix G

Let $i = 2$. $C(\mathbf{x}, \mathbf{a})$ is submodular in (b_2, g_1, a_2) because of the convexity of h_i and the condition $P_e(g_1) \geq P_e(g_1 + 1)$. By knowing the submodularity of $V^{(n-1)}$ in (b_2', g_1') and the $L^\natural$-convexity of $Q^{(n)}$ in (b_2, a_2) under condition $\xi_o \geq 2\lambda + \eta + \tau$, we can show the submodularity of $Q^{(n)}$ in (b_2, a_2) and (b_2, g_1). Consider the submodularity of $Q^{(n)}$ in (g_1, a_2). We can show that

$$\begin{aligned}
&Q^{(n)}(\mathbf{b}, \mathbf{g}, \mathbf{a} + \mathbf{e}_2) + Q^{(n)}(\mathbf{b}, \mathbf{g} + \mathbf{e}_1, \mathbf{a}) - Q^{(n)}(\mathbf{b}, \mathbf{g}, \mathbf{a}) \\
&\qquad - Q^{(n)}(\mathbf{b}, \mathbf{g} + \mathbf{e}_1, \mathbf{a} + \mathbf{e}_2) \\
&\qquad \geq \eta\Big(P_e(g_1) - P_e(g_1 + 1) + \beta \mathbb{E}_{g_1'}[P_e(g_1' + 1) \\
&\qquad - P_e(g_1')|g_1]\Big) \geq 0. \qquad (35)
\end{aligned}$$

The second last inequality in (35) is obtained by using a similar approach as in (32), and the last one is due to the condition $\beta \leq \frac{P_e(g_i) - P_e(g_i + 1)}{\sum_{g_i'} P_{g_i g_i'}(P_e(g_i') - P_e(g_i' + 1))}$. Therefore, $Q^{(n)}$ is submodular in (b_2, g_1, a_2). Similarly, we can show that Theorem 4.15 holds for $i = 1$.

Appendix H

In an equiprobable partitioned slow and flat Rayleigh fading channel, the channel transitions can be worked out by level crossing rate (LCR) [15] and only happens between adjacent states, i.e., $g_i' \in \{g_i - 1, g_i, g_i + 1\}$. Further, $P_{gg'} = P_{g'g}$, and $P_{gg'} \ll P_{gg}$ for all $g' \neq g$. According to Definition 4.8, for nondecreasing u, $P_{g_i g_i'}$ is first order stochastic nondecreasing in g_i because

$$\begin{aligned}
&\sum_{(g_i+1)'} P_{(g_i+1)(g_i+1)'} u\left((g_i + 1)'\right) - \sum_{g_i'} P_{g_i g_i'} u(g_i') \qquad (36) \\
&\qquad \geq (1 - 2P_{g_i g_i + 1})\left(u(g_i + 1) - u(g_i)\right) \geq 0,
\end{aligned}$$

where $1 - 2P_{g_i g_i + 1} \geq 0$ is because $P_{gg'} \ll P_{gg}$ and $\sum_{g'} P_{gg'} = 1$.

Abbreviations

NC-TWRC: network-coded two-way relay channels; MDP: Markov decision process; DP: dynamic programming; MPI: monotonic policy iteration; DSPSA: discrete simultaneous perturbation stochastic approximation; NC: network coding; FSMC: finite-state Markov chain; QoS: quality of service; SPSA: simultaneous perturbation stochastic approximation.

Competing interests

The authors declare that they have no competing interests.

Received: 9 July 2015 Accepted: 20 October 2015
Published online: 06 November 2015

References

1. R Ahlswede, N Cai, S-YR Li, RW Yeung, Network information flow. IEEE Trans. Inf. Theory. **46**(4), 1204–1216 (2000). doi:10.1109/18.850663
2. Y Wu, Information exchange in wireless networks with network coding and physical-layer broadcast. Technical Report MSR-TR-2004-78, Microsoft Research, Redmond WA (2004)
3. S Katti, H Rahul, W Hu, D Katabi, M Médard, J Crowcroft, Xors in the air: practical wireless network coding. SIGCOMM Comput. Commun. Rev. **36**(4), 243–254 (2006)
4. C Hausl, J Hagenauer, in *Proceedings of IEEE International Conference on Communications: June 2006*. Iterative network and channel decoding for the two-way relay channel (IEEE Istanbul, 2006), pp. 1568–1573. doi:10.1109/ICC.2006.255034

5. J Li, S Song, Y Guo, M Lee, Joint optimization of source and relay precoding for AF MIMO relay systems. EURASIP J. Wireless Commun. Netw. **2015**, 175 (2015)
6. M Soussi, A Zaidi, L Vandendorpe, DF-based sum-rate optimization for multicarrier multiple access relay channel. EURASIP J. Wireless Commun. Netw. **2015**, 133 (2015)
7. J Joung, AH Sayed, Multiuser two-way amplify-and-forward relay processing and power control methods for beamforming systems. IEEE Trans. Signal Process. **58**(3), 1833–1846 (2010)
8. S Katti, D Katabi, W Hu, H Rahul, M Medard, in *Proceedings of 43rd Annual Allerton Conference on Communications, Control and Computing: September 2005*. The importance of being opportunistic: Practical network coding for wireless environments (University of Illinoi Monticello, IL, 2005), pp. 756–765
9. S Peters, A Panah, K Truong, R Heath, Relay architectures for 3GPP LTE-advanced. EURASIP J. Wireless Commun. Netw. **2009**, 618787 (2009)
10. Q-T Vien, L-N Tran, E-K Hong, Network coding-based retransmission for relay aided multisource multicast networks. EURASIP J. Wireless Commun. Netw. **2011**, 643920 (2011)
11. W Chen, KB Letaief, Z Cao, in *Proceedings of IEEE International Conference on Communications: 24-28 June 2007*. Opportunistic network coding for wireless networks (IEEE Glasgow, 2007), pp. 4634–4639. doi:10.1109/ICC.2007.765
12. Y-P Hsu, N Abedini, S Ramasamy, N Gautam, A Sprintson, S Shakkottai, in *Proceedings IEEE International Symposium on Information Theory: July 31 -August 5 2011*. Opportunities for network coding: To wait or not to wait (IEEE St. Petersburg, 2011), pp. 791–795. doi:10.1109/ISIT.2011.6034243
13. N Ding, I Nevat, GW Peters, J Yuan, in *Proceedings of IEEE International Conference on Communications: 9-13 June 2013*. Opportunistic network coding for two-way relay fading channels (IEEE Budapest, 2013), pp. 5980–5985. doi:10.1109/ICC.2013.6655556
14. ML Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st edn. (Wiley, New York, 1994)
15. P Sadeghi, RA Kennedy, PB Rapajic, R Shams, Finite-state Markov modeling of fading channels: A survey of principles and applications. IEEE Signal Process. Mag. **25**(5), 57–80 (2008). doi:10.1109/MSP.2008.926683
16. RS Sutton, AG Barto, *Introduction to Reinforcement Learning*, 1st edn. (MIT Press, Cambridge, MA, 1998)
17. WB Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. (Wiley, New Jersey, 2007)
18. JE Smith, KF McCardle, Structural properties of stochastic dynamic programs. Oper. Res. **50**(5), 796–809 (2002)
19. J Huang, V Krishnamurthy, Transmission control in cognitive radio as a Markovian dynamic game: Structural result on randomized threshold policies. IEEE Trans. Commun. **58**(1), 301–310 (2010)
20. MH Ngo, V Krishnamurthy, Monotonicity of constrained optimal transmission policies in correlated fading channels with ARQ. IEEE Trans. Signal Process. **58**(1), 438–451 (2010). doi:10.1109/TSP.2009.2027735
21. DV Djonin, V Krishnamurthy, MIMO transmission control in fading channels–a constrained Markov decision process formulation with monotone randomized policies. IEEE Trans. Signal Process. **55**(10), 5069–5083 (2007)
22. DM Topkis, *Supermodularity and Complementarity*. (Princeton University Press, Princeton, 2001)
23. K Murota, Note on multimodularity and L-convexity. Math. Oper. Res. **30**(3), 658–661 (2005)
24. L Yang, YE Sagduyu, JH Li, in *Proceedings of 13th ACM International Symposium on Mobile Ad Hoc Networking and Computing: 11-14 June 2012*. Adaptive network coding for scheduling real-time traffic with hard deadlines (ACM SIGMOBILE New York, 2012), pp. 105–114
25. B Hajek, Extremal splittings of point processes. Math. Oper. Res. **10**(4), 543–556 (1985)
26. DM Topkis, Minimizing a submodular function on a lattice. Oper. Res. **26**(2), 305–321 (1978)
27. K Murota, *Discrete Convex Analysis*. (SIAM, Philadelphia, 2003)
28. QLP Yu, Multimodularity and structural properties of stochastic dynamic programs. Working Paper. School of Bus. and Manage., HongKong University of Sci. and Tech (2013)
29. P Zipkin, On the structure of lost-sales inventory models. Oper. Res. **58**(4), 937–944 (2008)
30. P Milgrom, J Roberts, Rationalizability, learning, and equilibrium in games with strategic complementarities. Econometrica J. Econ. Soc. **58**(6), 1255–1277 (1990)
31. AT Hoang, M Motani, Cross-layer adaptive transmission: Optimal strategies in fading channels. IEEE Trans. Commun. **56**(5), 799–807 (2008). doi:10.1109/TCOMM.2008.060214
32. Q Wang, JC Spall, in *Proceedings of American Control Conference: June 29-July 1 2011*. Discrete simultaneous perturbation stochastic approximation on loss function with noisy measurements (IEEE San Francisco, CA, 2011), pp. 4520–4525
33. JC Spall, Implementation of the simultaneous perturbation algorithm for stochastic optimization. IEEE Trans. Aerosp. Electron. Syst. **34**(3), 817–823 (1998). doi:10.1109/7.705889
34. N Ding, P Sadeghi, RA Kennedy, Discrete Convexity and Stochastic Approximation for Cross-layer On-off Transmission Control. Wireless Communications, IEEE Transactions on, 1536–1276 (2015). doi:10.1109/TWC.2015.2473858
35. JC Spall, Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans. Autom. Control. **37**(3), 332–341 (1992). doi:10.1109/9.119632
36. A Gosavi, *Simulation-based Optimization: Parametric Optimization Techniques and Reinforcement Learning*, vol. 55. (Springer, New York, 2014)
37. L Rastrigin, Convergence of random search method in extremal control of many-parameter system. Automat. Remote Control. **24**(11), 1337 (1964)
38. S Kirkpatrick, Optimization by simulated annealing: quantitative studies. J. Stat. Phys. **34**(5-6), 975–986 (1984)
39. M Caramia, P Dell'Olmo, *Multi-objective Management in Freight Logistics: Increasing Capacity, Service Level and Safety with Optimization Algorithms*. (Springer, Berlin, Germany, 2008)
40. W Zhuang, MZF Li, Monotone optimal control for a class of Markov decision processes. European J. Oper. Res. **217**(2), 342–350 (2012)