

# Queue-Aware Distributive Resource Control for Delay-Sensitive Two-Hop MIMO Cooperative Systems

Rui Wang, Vincent K. N. Lau and Ying Cui

Department of ECE, The Hong Kong University of Science and Technology

Email: wray@ust.hk, eeknlau@ust.hk, cuiying@ust.hk

## Abstract

In this paper, we consider a queue-aware distributive resource control algorithm for two-hop MIMO cooperative systems. We shall illustrate that relay buffering is an effective way to reduce the intrinsic half-duplex penalty in cooperative systems. The complex interactions of the queues at the source node and the relays are modeled as an average-cost infinite horizon Markov Decision Process (MDP). The traditional approach solving this MDP problem involves centralized control with huge complexity. To obtain a distributive and low complexity solution, we introduce a linear structure which approximates the value function of the associated Bellman equation by the sum of per-node value functions. We derive a distributive *two-stage two-winner auction-based* control policy which is a function of the local CSI and local QSI only. Furthermore, to estimate the *best fit* approximation parameter, we propose a distributive online stochastic learning algorithm using stochastic approximation theory. Finally, we establish technical conditions for almost-sure convergence and show that under heavy traffic, the proposed low complexity distributive control is global optimal.

## I. INTRODUCTION

Cooperative relay communication has been a hot research topic in both the academia [1], [2] and the industry [3], [4] because it could exploit the broadcast nature of wireless communication to achieve cooperative diversity. One potential issue of cooperative communication is the half-duplex penalty in the relay nodes. There have been some recent works to address the half-duplex issue in cooperative relay systems. For example, complex echo cancelation technique is used at the relay to cancel the coupled interference from the transmitting path [5], [6]. However, these works all focused at the physical layer signal processing. In [7], the authors exploit special topology and proposed some relay protocols to get

rid of the half-duplex penalty. Moreover, this approach depends heavily on the locations of the relays and it cannot be extended to general relay channel. In this paper, we are interested to explore a system level solution to deal with the half-duplex issue. We consider a simple MIMO cooperative relay system with a multi-antenna source node (Src),  $M$  multi-antenna relay nodes (RS) and a multi-antenna destination node (Dst). We shall illustrate that relay buffering can be utilized to significantly reduce the intrinsic half-duplex penalty. Since buffering is involved, it is important to consider not only the throughput performance but also the associated end-to-end delay performance. As a result, we shall focus on delay-optimal resource control for the two-hop protocol in MIMO cooperative relay systems.

Delay-optimal resource control in cooperative relay system is a very difficult problem. Most of the existing works have assumed infinite backlogs of information and focus on optimizing the throughput performance only. A systematic approach is to model the delay-optimal control as Markov Decision Process (MDP) [8], [9]. However, there is a well-known issue of the *curse of dimensionality* and brute force value iteration or policy iteration could not give simple implementable solutions<sup>1</sup>. For multi-hop systems, there is a unique challenge concerning the complex interactions of buffers at the source node and the  $M$  RS nodes and the existing solutions for single-hop systems cannot be extended easily to deal with this situation. There are a few recent works that considered queue dynamics in relay systems [10], [11]. However, these works have focused on the characterization of the *stability region* and throughput optimal control. The question of delay-optimal control for cooperative relay system remains to be open. In addition, another important technical challenge is the distributive implementation consideration. For instance, the entire system state could be characterized by the *global CSI* (CSI among every pair of nodes in the system) as well as the *global QSI* (QSI of every buffer in the system). Brute-force solution of the MDP will yield a control policy that is adaptive to the global CSI and global QSI. This poses a huge implementation challenges because these global system state information are distributed locally at each of the source and relay nodes.

In this paper, we shall address the above challenges as follows. We shall first formulate the delay-optimal resource control policy (such as the power control and RS selection) as an average-cost infinite horizon Markov Decision Process (MDP). To alleviate the *curse of dimensionality*, and to obtain a distributive and

<sup>1</sup>For example, for a system with maximum buffer length of 20, 3 CSI states and  $M$  RSs, the total number of system states is  $20^{M+1} \times 3^{2M}$ , which is unmanageable even for small number of RS.

low complexity solution, we first introduce a *per-node value function* to approximate the value function of the associated Bellman equation. Based on the per-node value function, we derive a distributive *two-stage two-winner auction-based* control policy, which is a function of the local CSI and local QSI. The per-node value function is obtained via a distributive online stochastic learning algorithm, which requires local CSI and local QSI only. The proposed online stochastic learning is quite different from the conventional reinforced learning [12] in mainly two ways: (1) We are dealing with *constrained MDP* (CMDP) and our online iterative solution updates both the value function and the Lagrange multipliers (LM) simultaneously; (2) The control action is determined from the per-node value function of all the nodes via a per-slot auction mechanism. Therefore, the algorithm dynamics of the per-node online learning is not a *contraction mapping* and hence, standard convergence proof using fixed point theorem cannot be applied in our case directly. Using the technique of separation of different time scales, we establish technical conditions for the almost sure convergence of the proposed distributive stochastic learning. We also show that the proposed low complexity distributive solution is asymptotically global optimal under heavy traffic loading. Finally, we demonstrate by simulation that the proposed scheme has significant performance gain over various baselines (such as conventional *CSIT-only* control and the *throughput-optimal control* (in stability sense)) with low complexity  $\mathcal{O}(M)$  and low signaling overhead.

## II. SYSTEM MODELS

### A. System Architecture and MIMO Relay Physical Layer Model

We consider a two-hop multi-antenna cooperative relay communication system with one multi-antenna source node ( $N_T$  antennas),  $M$  multi-antenna half-duplex relay stations (RS, each with  $N_R$  antennas) and one multi-antenna destination node ( $N_T$  antennas), as illustrated in Fig. 1. The source node cannot deliver packets directly to the destination node due to limited coverage and the cooperative RSs are deployed to extend the source node's coverage.

Denote the Rx-RS and the Tx-RS as the  $m$ -th RS and the  $n$ -th RS for notation simplicity<sup>2</sup>. Let  $N_{SR}$  and  $N_{RD}$  be the number of data streams transmitted in the S-R link and the R-D link respectively, where we require  $N_{RD} = \min(N_T, N_R - N_{SR})$  for simultaneous interference-free transmission. We shall illustrate the signal model of the S-R <sub>$m$</sub>  link and the R <sub>$n$</sub> -D link as follows:

<sup>2</sup>Since the RSs are half-duplex under practical consideration, we require  $m \neq n$  implicitly.

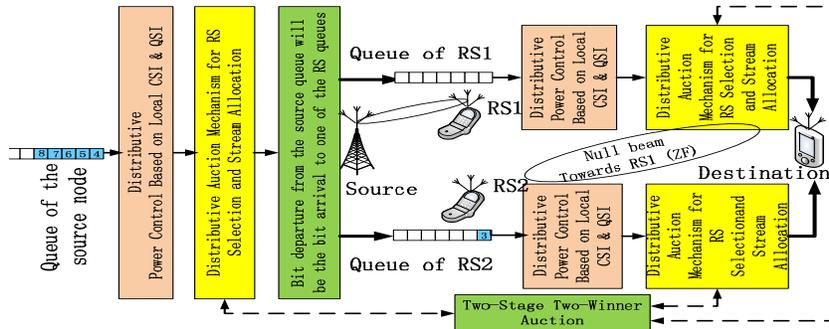


Fig. 1. Illustration of the two-hop MIMO cooperative system with a multi-antenna source node, 2 multi-antenna RS nodes and a multi-antenna destination node. By exploiting buffers at the 2 MIMO RSs, the S-R link (source node to RS1) and R-D link (RS2 to destination node) can deliver packets simultaneously without interfering with each other using signal processing techniques (with appropriate precoder and decorrelator designs). Thus, by exploiting relay buffering, we could substantially reduce the intrinsic penalty associated with half duplex relays.

- **S-R<sub>m</sub> link:** Let  $\mathbf{X}_S \in C^{N_{SR} \times 1}$  and  $\mathbf{F}_S \in C^{N_T \times N_{SR}}$  be the symbol vector and the precoder matrix of the source node respectively,  $\mathbf{G}_m \in C^{N_{SR} \times N_R}$  be the decorrelator matrix at the  $m$ -th RS node, the  $N_{SR} \times 1$  post-processing symbol vector at the  $m$ -th RS is given by  $\mathbf{Y}_m = \mathbf{G}_m \mathbf{H}_{S,m} \mathbf{F}_S \mathbf{X}_S + \mathbf{Z}_{S,m}$ , where  $\mathbf{H}_{S,m} \in C^{N_R \times N_T}$  is the zero-mean unit variance i.i.d. complex Gaussian fading matrix from the source node to the  $m$ -th RS,  $\mathbf{Z}_{S,m} \in C^{N_{SR} \times 1}$  is the zero-mean unit variance complex Gaussian channel noise.
- **R<sub>n</sub>-D link:** Let  $\mathbf{X}_n \in C^{N_{RD} \times 1}$  and  $\mathbf{F}_n \in C^{N_R \times N_{RD}}$  be the transmit symbol vector and the precoder of the  $n$ -th RS respectively, the  $N_T \times 1$  received symbol vector at the destination node is given by<sup>3</sup>  $\mathbf{Y}_D = \mathbf{H}_{n,D} \mathbf{F}_n \mathbf{X}_n + \mathbf{Z}_{n,D}$ , where  $\mathbf{H}_{n,D} \in C^{N_T \times N_R}$  is complex Gaussian fading matrix from the  $n$ -th RS to the destination node,  $\mathbf{Z}_{n,D} \in C^{N_T \times 1}$  is the complex Gaussian channel noise.

In this paper, the resource control is performed distributively on each RS and therefore, we define the local channel state information (CSI) available at each RS as follows. For the  $m$ -th RS, there are two types of *local CSI*, namely the *type-I local CSI* and *type-II local CSI* as illustrated in Fig. 2. The type-I and type-II local CSI of the  $m$ -th RS are denoted by  $\mathbf{H}_m^I = \{\mathbf{H}_{S,m}\}$  and  $\mathbf{H}_m^{II} = \{\mathbf{H}_{m,D}\} \cup \{\mathbf{H}_{m,n} | n \neq m\}$ .

<sup>3</sup>Due to the limited coverage of the source node, we assume the received signal from the source node is negligible compared with the received signal from the relay node.

$m, 1 \leq n \leq M\}$ , respectively. For notation convenience, let  $\mathbf{H}_m = \mathbf{H}_m^I \cup \mathbf{H}_m^{II}$  be the local CSI<sup>4</sup> at the  $m$ -th RS and  $\mathbf{H} = \cup_{m=1}^M \mathbf{H}_m$  be the global CSI (GCSI) of the system. Moreover, the assumption on the channel is summarized below:

*Assumption 1 (Assumption on Channel Fading):* We assume the channel fading elements in the global CSI  $\mathbf{H}$  are i.i.d.  $\mathcal{CN}(0, 1)$ . The CSI is quasi-static within a frame but i.i.d. between frames.

We assume strong channel coding is used and hence, the maximum achievable data rate is given by the instantaneous mutual information<sup>5</sup>. If the source node transmits  $R_{S,m}$  information bits to the  $m$ -th RS in the current frame, the frame will be successfully received if  $R_{S,m} \leq \tau \log_2 \det [\mathbf{I} + \mathbf{G}_m \mathbf{H}_{S,m} \mathbf{F}_S \mathbf{F}_S^\dagger \mathbf{H}_{S,m}^\dagger \mathbf{G}_m^\dagger]$ , where  $\dagger$  denotes the matrix conjugate transpose and  $\tau$  is the frame duration. Similarly, the destination node could successfully decode a frame with  $R_{n,D}$  information bits (transmitted from the  $n$ -th RS) if  $R_{n,D} \leq \tau \log_2 \det [\mathbf{I} + \mathbf{H}_{n,D} \mathbf{F}_n \mathbf{F}_n^\dagger \mathbf{H}_{n,D}^\dagger]$ .

### B. Buffered Decode and Forward

Although the RS nodes are half-duplex relays<sup>6</sup>, it is possible to reduce the system half-duplex penalty by exploiting buffers at the half-duplex RSs. Specifically, the source node could transmit a packet to the  $m$ -th RS (denoted as the Rx-RS) and at the same time, the  $n$ -th RS (denoted as the Tx-RS) transmits its *buffered packet* to the destination node without interfering the Rx-RS. This is possible by means of precoder-decorrelator designs at the source node, Rx-RS ( $m$ -th RS) and the Tx-RS ( $n$ -th RS). Let  $p_{S,m}$  and  $p_{n,D}$  denote the total transmit power at the source node for the S-R <sub>$m$</sub>  link and the Tx-RS ( $n$ -th RS) for the R <sub>$n$</sub> -D link, respectively. For any given  $N_{SR}$ ,  $p_{S,m}$  for the S-R <sub>$m$</sub>  link as well as  $N_{RD}$ ,  $p_{n,D}$  for the R <sub>$n$</sub> -D link (where  $N_{RD} = \min(N_T, N_R - N_{SR})$  implicitly), the decorrelator and precoder designs are elaborated below.

- **Precoder and Decorrelator Design of the S-R <sub>$m$</sub>  Link at the Rx-RS Node<sup>7</sup>:** The precoder at the source node ( $\mathbf{F}_S$ ) and the decorrelator at the Rx-RS node ( $\mathbf{G}_m$ ) are chosen to optimize the mutual

<sup>4</sup>Note that both the type-I and type-II local CSI at the  $m$ -th RS refers to all the outgoing links from the  $m$ -th RS and hence, they can be measured at the  $m$ -th RS using channel reciprocity and preambles. For example, there are standard signaling and channel sounding mechanisms in the WiMAX (802.16j, 802.16m) and LTE systems for the RS to acquire the local CSI.

<sup>5</sup>For example, LDPC with reasonably large block length (e.g 8kbyte) can achieve the instantaneous mutual information within 0.5dB SNR [13].

<sup>6</sup>Half-duplex relay means that the RS nodes do not have any Tx/Rx echo-cancellation capability.

<sup>7</sup>Type-I local CSI  $\mathbf{H}_m^I$  is required at the  $m$ -th Rx-RS node to compute the precoder and decorrelator of the S-R <sub>$m$</sub>  link.

information of the S-R<sub>m</sub> link subject to the transmit power constraint as follows:

$$\begin{aligned} \{\mathbf{G}_m^*(N_{SR}), \mathbf{F}_S^*(N_{SR}, p_{S,m})\} &= \arg \max_{\mathbf{F}_S, \mathbf{G}_m} \log_2 \det \left[ \mathbf{I} + \mathbf{G}_m(N_{SR}) \mathbf{H}_{S,m} \mathbf{F}_S \mathbf{F}_S^\dagger \mathbf{H}_{S,m}^\dagger \mathbf{G}_m^\dagger(N_{SR}) \right] \\ s.t. \quad &tr(\mathbf{F}_S \mathbf{F}_S^\dagger) = p_{S,m} \quad (\text{Transmit power constraint}). \end{aligned} \quad (1)$$

Let  $\mathbf{H}_{S,m} = \mathbf{U}_{S,m} \boldsymbol{\Sigma}_{S,m} \mathbf{V}_{S,m}^\dagger$  be the SVD decomposition of channel matrix  $\mathbf{H}_{S,m}$ , where the singular values in  $\boldsymbol{\Sigma}_{S,m}$  are sorted in a decreasing order along the diagonal,  $\mathbf{U}_{S,m} = [\mathbf{u}_{S,m}^1, \dots, \mathbf{u}_{S,m}^{N_R}]$  and  $\mathbf{V}_{S,m} = [\mathbf{v}_{S,m}^1, \dots, \mathbf{v}_{S,m}^{N_T}]$ . Using standard optimization techniques [14], the source precoder  $\mathbf{F}_S^*$  is given by

$$\mathbf{F}_S^*(N_{SR}, p_{S,m}) = [\mathbf{v}_{S,m}^1, \dots, \mathbf{v}_{S,m}^{N_{SR}}] \times \text{diag} \left\{ \frac{1}{\lambda_{S,m}} - \frac{1}{\eta_{S,m}^1}, \dots, \frac{1}{\lambda_{S,m}} - \frac{1}{\eta_{S,m}^{N_{SR}}} \right\}, \quad (2)$$

where  $\eta_{S,m}^1 \geq \eta_{S,m}^2 \geq \dots \geq \eta_{S,m}^{N_{SR}}$  are the first  $N_{SR}$  singular values of channel matrix  $\mathbf{H}_{S,m}$ ,  $\lambda_{S,m}$  is the Lagrange multiplier corresponding to the transmit power constraint in (1). The decorrelator  $\mathbf{G}_m^*$  is given by

$$\mathbf{G}_m^*(N_{SR}) = [\mathbf{u}_{S,m}^1, \dots, \mathbf{u}_{S,m}^{N_{SR}}]^\dagger. \quad (3)$$

- **Precoder Design of the R<sub>n</sub>-D Link at the Tx-RS Node<sup>8</sup>:** Similarly, given the decorrelator  $\mathbf{G}_m^*$  in (3), the precoder at the Tx-RS node  $\mathbf{F}_n \in C^{N_{RD} \times N_R}$  is selected to maximize R-D link mutual information subject to the transmit power constraint and the interference nulling constraint (at the Rx-RS node) as follows:

$$\begin{aligned} \mathbf{F}_n^*(N_{RD}, p_{n,m}) &= \arg \max_{\mathbf{F}_n} \log_2 \det \left[ \mathbf{I} + \mathbf{H}_{n,D} \mathbf{F}_n \mathbf{F}_n^\dagger \mathbf{H}_{n,D}^\dagger \right] \\ s.t. \quad &\mathbf{G}_m^*(N_{SR}) \mathbf{H}_{n,m} \mathbf{F}_n = 0 \quad (\text{Interference nulling constraint}) \end{aligned} \quad (4)$$

$$tr(\mathbf{F}_n \mathbf{F}_n^\dagger) = p_{n,D} \quad (\text{Transmit power constraint}) \quad (5)$$

The interference nulling constraint in (4) is to allow simultaneously active R-D and S-R links using half-duplex RSs. Let  $\mathbf{H}_{n,D} \times \text{null}(\mathbf{G}_m \mathbf{H}_{n,m}) = \mathbf{U}_{n,D} \boldsymbol{\Sigma}_{n,D} \mathbf{V}_{n,D}^\dagger$  be the SVD decomposition, where the singular values in  $\boldsymbol{\Sigma}_{n,D}$  are sorted in a decreasing order along the diagonal,  $\text{null}(\mathbf{G}_m \mathbf{H}_{n,m})$  denotes the null space of matrix  $\mathbf{G}_m \mathbf{H}_{n,m}$  and  $\mathbf{V}_{n,D} = [\mathbf{v}_{n,D}^1, \dots, \mathbf{v}_{n,D}^{N_R - N_{SR}}]$ . Using standard optimization techniques [14], the precoder at the Tx-RS ( $\mathbf{F}_n^*$ ) is given by:

$$\mathbf{F}_n^*(N_{RD}, p_{n,m}) = [\mathbf{v}_{S,n}^1, \dots, \mathbf{v}_{S,n}^{N_{RD}}] \times \text{diag} \left\{ \frac{1}{\lambda_{n,D}} - \frac{1}{\eta_{n,D}^1}, \dots, \frac{1}{\lambda_{n,D}} - \frac{1}{\eta_{n,D}^{N_{RD}}} \right\}, \quad (6)$$

<sup>8</sup>Type-II local CSI  $\mathbf{H}_n^{II}$  is required at the  $n$ -th Tx-RS node to compute the precoder of the R<sub>n</sub>-D link.

where  $\eta_{n,D}^1 \geq \eta_{n,D}^2 \dots \geq \eta_{S,m}^{N_{RD}}$  are the first  $N_{RD}$  singular values of channel matrix  $\mathbf{H}_{n,D} \times null(\mathbf{G}_m \mathbf{H}_{n,m})$ ,  $\lambda_{n,D}$  is the Lagrange multiplier corresponding to the power constraint in (5).

### C. Bursty Source Model and Queue Dynamics

There is one queue in the source node and one queue in each of the  $M$  RSs respectively for the storage of received information bits. Let  $N_Q$  be the maximum buffer size (number of bits) for the buffers in the source node and all the RSs. Let  $X(t)$  indicates the number of new information bits arrival in the  $t$ -th frame at the source node. The assumption on the bit arrival process is given below:

*Assumption 2 (Assumption on Arrival Process):* We assume  $X(t)$  is i.i.d. over frames based on a general distribution  $f_X(x)$  with  $\mathbb{E}[X(t)] = \lambda_S$  and the information bits arrive at the end of each frame.

Moreover, let  $Q_S(t)$  and  $Q_m(t)$  denote the number of information bits in the source node's queue and the  $m$ -th RS's queue ( $1 \leq m \leq M$ ) at frame  $t$ . We assume each RS has the knowledge of its own queue length and the source node's queue length. Thus, the local QSI of the  $m$ -th RS is  $(Q_S(t), Q_m(t))$ .  $\mathbf{Q}(t) = (Q_S(t), Q_1(t), \dots, Q_M(t))$  denotes the *global queue state information* (GQSI) at frame  $t$ .

The overall system queue dynamics at the source node and the RSs are summarized below.

- If the source node successfully delivers  $R_{S,m}(t)$  information bits to the  $m$ -th RS at frame  $t$ , then  $Q_S(t+1) = \min \{ \max \{ Q_S(t) - R_{S,m}(t), 0 \} + X(t), N_Q \}$  and  $Q_m(t+1) = \min \{ Q_m(t) + R_{S,m}(t), N_Q \}$ .
- If the source node fails to deliver any information bit to the RSs, then  $Q_S(t+1) = \min \{ Q_S(t) + X(t), N_Q \}$ .
- If the  $n$ -th RS successfully delivers  $R_{n,D}(t)$  information bits to the destination at frame  $t$ , then  $Q_n(t+1) = \max \{ Q_n(t) - R_{n,D}(t), 0 \}$ .

*Remark 1:* Each information bit delivered from the source node will be received by one of the RSs and different RSs may have different information bits in the buffer. When the source node is to deliver information bits to one RS, selecting different RSs with different buffer lengths may have different effects on the average packet delay of the system. Therefore, not only the CSI of all S-R links but also the QSI of all RSs should be considered in directing the source node's transmission. Such coupling on the system QSI is a unique challenge in delay-optimal control of multi-hop systems. Fig. 1 shows the top level architecture illustrating the interactions among all the queues in the two-hop cooperative system.



includes the power allocation policy of S-R link and R-D link  $\Pi_p^m$ , the first-stage and second-stage bidding policy  $\Pi_A^m$  and  $\Pi_B^m$ . Specifically,

$$\Pi_p^m(S_m) = \left\{ p_{S,m}(N_{SR}), p_{m,D}(N_{RD}) : N_{SR}, N_{RD} = 0, 1, \dots, \min(N_T, N_R) \right\} \triangleq \mathbf{P}_m \quad (7)$$

$$\Pi_A^m(S_m) = \left\{ A_m(N_{SR}) | N_{SR} = 0, 1, \dots, \min(N_T, N_R) \right\} \triangleq \mathbf{A}_m \quad (8)$$

$$\Pi_B^m(S_m, \cup_{m'=1}^M \mathbf{A}_{m'}) = \left\{ B_m, I_m, (N_{SR,m}, N_{RD,m}) | N_{RD} = \min(N_T, N_R - N_{SR}) \right\} \triangleq \mathbf{B}_m \quad (9)$$

for  $m = 1, 2, \dots, M$ , where  $p_{S,m}(N_{SR})$  is the total transmit power allocation at the source node for the S-R link with  $N_{SR}$  data streams,  $p_{m,D}(N_{RD})$  is the total transmit power allocation at the Tx-RS for the R-D link with  $N_{RD}$  data streams.

Denote the local system state of the  $m$ -th RS as  $S_m = (Q_S, Q_m, \mathbf{H}_m)$  ( $1 \leq m \leq M$ ). Therefore, the global system state is given by  $\mathbf{S} = \cup_{m=1}^M S_m = (\mathbf{Q}, \mathbf{H})$ .

*Remark 2 (Distributive Consideration of Stationary Control Policy  $\Pi$  in Definition 1):* The stationary control policy  $\Pi = \{\Pi^m | 1 \leq m \leq M\}$  is distributive in the sense that the policy  $\Pi^m$  at each RS  $m$  only depends on the local system state  $S_m$  and the broadcast bidding information available at RS  $m$ . Thus, for notation simplicity, we shall omit the bidding information when the meaning is clear, i.e. we shall use  $\Pi^m(S_m) = \{\Pi_p^m(S_m), \Pi_A^m(S_m), \Pi_B^m(S_m)\}$  in the rest of the paper.

A stationary control policy  $\Pi$  induces a joint distribution for the random process  $\{\mathbf{S}(t)\}$ . Under Assumption 1 and 2,  $\mathbf{S}(t+1)$  only depends on  $\mathbf{S}(t)$  and actions at frame  $t$ , and hence the induced random process  $\{\mathbf{S}(t)\}$  for a given control policy  $\Pi$  is Markovian with the following transition probability:

$$\Pr[\mathbf{S}(t+1) | \mathbf{S}(t), \Pi(\mathbf{S}(t))] = \Pr[\mathbf{H}(t+1)] \Pr[\mathbf{Q}(t+1) | \mathbf{S}(t), \Pi(\mathbf{S}(t))], \quad (10)$$

where the equality is because of Assumption 1 and the queue dynamics transition probability  $\Pr[\mathbf{Q}(t+1) | \mathbf{S}(t), \Pi(\mathbf{S}(t))]$  is given by

$$\Pr[\mathbf{Q}(t+1) | \mathbf{S}(t), \Pi(\mathbf{S}(t))] \quad (11)$$

$$= \begin{cases} \Pr[X(t) = Q_S(t+1) - [Q_S(t) - R_{S,m^*}(t)]^+], & \text{if } Q_m(t+1) = Q_m(t) \ (\forall m \neq m^*, n^*) \\ \quad \text{and } Q_{m^*}(t+1) = \min\{Q_{m^*}(t) + R_{S,m^*}(t), N_Q\}, \quad Q_{n^*}(t+1) = \max\{Q_{n^*}(t) - R_{n^*,D}(t), 0\} \\ 0, & \text{otherwise} \end{cases}$$

Given a unichain policy  $\Pi$ , the induced Markov chain  $\{\mathbf{S}(t)\}$  is ergodic and there exists a unique steady state distribution  $\pi_S$  [8]. Therefore, we have the average end-to-end delay of the two-hop cooperative RS

system summarized in the following lemma:

*Lemma 1 (Average End-to-End Delay):* For small average packet drop rate constraint  $D$ , the average end-to-end delay of the two-hop cooperative RS system is given by

$$\bar{T}(\Pi) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E}^{\Pi} \left[ \frac{\sum_{m=S}^M Q_m(t)}{\lambda_S} \right] = \mathbf{E}_{\pi_S} \left[ \frac{\sum_{m=S}^M Q_m}{\lambda_S} \right] \quad (12)$$

where  $m = S, 1, 2, \dots, M$  in the equation<sup>10</sup>,  $\mathbf{E}_{\pi_S}$  means taking the expectation with respect to the induced steady state distribution  $\pi_S$  (induced by the unichain control policy  $\Pi$ ) and  $\lambda_S$  is the average number of arrival bits per frame at the source node.

*Proof: Please refer to Appendix A.* ■

Similarly, the source node's average drop rate constraint<sup>11</sup>, the source node's average power constraint and each RS  $m$ 's average power constraint are given by

$$\bar{D}(\Pi) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E}^{\Pi} \left[ \mathbf{I}[Q_S(t) = N_Q] \right] = \mathbf{E}_{\pi_S}^{\Pi} \left[ \mathbf{I}[Q_S = N_Q] \right] \leq D \quad (13)$$

$$\bar{P}_S(\Pi) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E}^{\Pi} \left[ \sum_{m=1}^M \sum_{i=1}^{N_{\min}} I_{S,m}^i(t) p_{S,m}(i)(t) \right] = \mathbf{E}_{\pi_S}^{\Pi} \left[ \sum_{m=1}^M \sum_{i=1}^{N_{\min}} I_{S,m}^i p_{S,m}(i) \right] \leq P_S \quad (14)$$

$$\bar{P}_m(\Pi) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E}^{\Pi} \left[ \sum_{i=1}^{N_{\min}} I_{m,D}^i(t) p_{m,D}^i(t) \right] = \mathbf{E}_{\pi_S} \left[ \sum_{i=1}^{N_{\min}} I_{m,D}^i p_{m,D}^i \right] \leq P_R, 1 \leq m \leq M \quad (15)$$

where  $N_{\min} = \min(N_T, N_R)$ ,  $I_{S,m}^i = \mathbf{I}[m = m^*] \mathbf{I}[i = N_{SR}^*]$  and  $I_{m,D}^i = \mathbf{I}[m = n^*] \mathbf{I}[i = N_{RD}^*]$ .

### III. CONSTRAINED MARKOV DECISION PROBLEM FORMULATION

In this section, we shall formulate the delay-optimal problem as an infinite horizon average cost constrained Markov Decision Problem (CMDP) and discuss the general solution.

#### A. CMDP Formulation

The goal of the controller is to choose an optimal stationary feasible unichain policy  $\Pi^*$  that minimizes the average end-to-end transmission delay in (12). Specifically, the delay-optimal control problem is summarized below.

<sup>10</sup>This abuse will also appear in the following of this paper as long as the meaning is clear.

<sup>11</sup>Since the source node and  $M$  RSs have buffers with the same buffer size  $N_Q$ , the average drop rate at each RS node is much lower than the average drop rate at the source node. Therefore, we omit the average drop rate constraint at each RS to simplify the problem.

*Problem 1 (Delay-Optimal Control Problem for MIMO Relay System):* Find a feasible stationary unichain policy  $\Pi = (\Pi^1, \dots, \Pi^M)$  such that the average end-to-end delay is minimized subject to the average drop rate constraint at the source node and the average power constraint at the source node and each RS node<sup>12</sup>,

$$\text{i.e. } \min_{\Pi} \bar{T}(\Pi) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E}^{\Pi} \left[ \sum_{m=S}^M Q_m(t) \right] \text{ s.t. (13), (14), (15).}$$

Problem 1 is an infinite horizon average cost constrained Markov Decision Problem (CMDP) [19] with system state space  $\mathcal{S} = \{\mathbf{S}^1, \mathbf{S}^2, \dots\} = \mathcal{Q} \times \mathcal{H}$  (where  $\mathcal{Q}$  is the global QSI state space and  $\mathcal{H}$  is the global CSI state space), action space  $\mathcal{P} \times \mathcal{A} \times \mathcal{B}$  (where  $\mathcal{P} = \{\forall \mathbf{p}_m | \forall m\}$  is power allocation action space,  $\mathcal{A} = \{\forall \mathbf{A}_m | \forall m\}$  is the first-stage bidding action space and  $\mathcal{B} = \{\forall \mathbf{B}_m | \forall m\}$  is the second-stage bidding action space), transition kernel given by (10), and the per-stage cost function  $d(\mathbf{S}, \Pi(\mathbf{S})) = \sum_{m=S}^M Q_m$ .

### B. Lagrangian Approach for the CMDP

The CMDP in Problem 1 can be converted into unconstrained MDP by the Lagrange theory [14]. For any vector of Lagrange multiplier (LM)  $\gamma = [\gamma_{S,d}, \gamma_{S,p}, \gamma_{1,p}, \dots, \gamma_{M,p}]^T$ , we define the Lagrangian as  $L(\Pi, \gamma) = \lim_{T \rightarrow +\infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E}_{\mathbf{S}}^{\Pi} \left[ g(\mathbf{S}(t), \Pi(\mathbf{S}(t)), \gamma) \right]$ , where

$$g(\mathbf{S}, \Pi(\mathbf{S}), \gamma) = Q_S + \gamma_{S,p} \sum_{m=1}^M \sum_{i=1}^{N_{\min}} I_{S,m}^i p_{S,m}(i) + \gamma_{S,d} \mathbf{I}[Q_S = N_Q] + \sum_{m=1}^M \left[ Q_m + \gamma_{m,p} \sum_{i=1}^{N_{\min}} I_{m,D}^i p_{m,D}(i) \right].$$

Therefore, the corresponding unconstrained MDP for a particular vector of LMs  $\gamma$  is given by

$$G(\gamma) = \min_{\Pi} L(\Pi, \gamma) = \min_{\Pi} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E}^{\Pi} \left[ g(\mathbf{S}(t), \Pi(\mathbf{S}(t)), \gamma) \right] \quad (16)$$

where  $G(\gamma)$  gives the Lagrange dual function. The dual problem of the primal problem in Problem 1 is given by  $\max_{\gamma \succeq 0} G(\gamma)$ . It is shown in [20] that there exists a Lagrange multiplier  $\gamma \geq 0$  such that  $\Pi^*$  minimizes  $L(\Pi, \gamma)$  and the saddle point condition the saddle point condition  $L(\Pi, \gamma^*) \geq L(\Pi^*, \gamma^*) \geq L(\Pi^*, \gamma)$  holds. Using standard Lagrange theory [14],  $\Pi^*$  is the primal optimal (i.e. solving Problem 1),  $\gamma^*$  is the dual optimal (solving the dual  $\gamma$  problem) and the duality gap is zero. Thus, by solving the dual problem, we can obtain the primal optimal  $\Pi^*$ . Therefore, we shall first solve the unconstrained MDP in (16) in the following.

For a given LM vector  $\gamma$ , the optimizing unichain policy for the unconstrained MDP (16) can be obtained by solving the associated *Bellman equation* w.r.t.  $(\theta, \{J(\mathbf{S})\})$  as follows

$$\theta + J(\mathbf{S}^i) = \min_{\Pi(\mathbf{S}^i)} \left\{ g(\mathbf{S}^i, \Pi(\mathbf{S}^i), \gamma) + \sum_{\mathbf{S}^j} \Pr[\mathbf{S}^j | \mathbf{S}^i, \Pi(\mathbf{S}^i)] J(\mathbf{S}^j) \right\} \quad \forall \mathbf{S}^i \in \mathcal{S}, \quad (17)$$

<sup>12</sup>To simplify the notation, we shall normalize  $\lambda_S = 1$  in the rest of the paper.

where  $\{J(\mathbf{S})\}$  is the value function of the MDP and  $\Pr[\mathbf{S}^j|\mathbf{S}^i, \Pi(\mathbf{S}^i)]$  is the transition kernel which can be obtained from (10),  $\theta = \min_{\Pi} L(\Pi, \gamma)$  is the optimal average cost per stage and the optimizing policy is  $\Pi^*$  with  $\Pi^*(\mathbf{S}^i)$  minimizing the R.H.S. of (17) at any state  $\mathbf{S}^i$ . For any unichain policy with irreducible Markov Chain  $\{\mathbf{S}(t)\}$ , the solution to (17) is unique [19]. We restrict our policy space to be *unichain policies*<sup>13</sup> and we denote  $\Pi^*$  as the optimal unichain policy.

### C. Equivalent Bellman Equation for the CMDP

The Bellman equation in (17) is a fixed point problem over the functional space and this is very complicated to solve due to the huge cardinality of the system state space. Brute-force solution could not lead to any useful implementations. In this subsection, we shall illustrate that the Bellman equation in (17) can be simplified into a equivalent form by exploiting the i.i.d. structure of the CSI process  $\mathbf{H}(t)$ . For notation convenience, we partition the unichain policy  $\Pi$  into a collection of actions based on the QSI. Specifically, we have the following definition.

*Definition 2 (Partitioned Actions for the m-th Relay):* Given a unichain control policy  $\Pi^m$ , we define  $\Pi^m(\mathbf{Q}) = \Pi^m(Q_S, Q_m) = \{\Pi^m(Q_S, Q_m, \mathbf{H}_m) | \forall \mathbf{H}_m\}$  as the collection of actions under a given local QSI  $(Q_S, Q_m)$  for all possible local CSI  $\mathbf{H}_m$ . The complete policy  $\Pi^m$  for the m-th RS is therefore equal to the union of all the partitioned actions, i.e.  $\Pi^m = \cup_{(Q_S, Q_m)} \Pi^m(Q_S, Q_m)$ .

Therefore, we have  $\Pi = \cup_{\mathbf{Q}} \Pi(\mathbf{Q})$  and we show that the optimal policy  $\Pi^*$  of (16) can be obtained by solving an *equivalent Bellman equation* summarized in the following lemma.

*Lemma 2 (Equivalent Bellman Equation):* The control policy obtained by solving the Bellman equation in (17) is the same as that obtained by solving the *equivalent Bellman equation* defined below:

$$\theta + V(\mathbf{Q}^i) = \min_{\Pi(\mathbf{Q}^i)} \left\{ \bar{g}(\mathbf{Q}^i, \Pi(\mathbf{Q}^i), \gamma) + \sum_{\mathbf{Q}^j} \Pr[\mathbf{Q}^j | \mathbf{Q}^i, \Pi(\mathbf{Q}^i)] V(\mathbf{Q}^j) \right\}, \forall \mathbf{Q}^i \in \mathcal{Q} \quad (18)$$

where  $\theta = \min_{\Pi} L(\Pi, \gamma)$  is the original optimal average cost per stage,  $V(\mathbf{Q}^i) = \mathbf{E}_{\mathbf{H}}[J(\mathbf{Q}^i, \mathbf{H}) | \mathbf{Q}^i]$  is the conditional average value function for state  $\mathbf{Q}^i$ , and

$$\bar{g}(\mathbf{Q}^i, \Pi(\mathbf{Q}^i), \gamma) = \mathbf{E}_{\mathbf{H}} \left[ g((\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H}), \gamma) | \mathbf{Q}^i \right] \quad (19)$$

<sup>13</sup>For most of the policies we are interested, the associated Markov chain is irreducible and hence, there is virtually no loss by restricting ourselves to unichain policies.

is the conditional per-stage cost and  $\Pr[\mathbf{Q}^j|\mathbf{Q}^i, \Pi(\mathbf{Q}^i)] = \mathbf{E}_{\mathbf{H}}[\Pr[\mathbf{Q}^j|(\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H})]]$  is the conditional average transition kernel.

*Proof: Please refer to Appendix B.* ■

*Remark 3:* Note that solving the R.H.S. of (18) for each  $\mathbf{Q}^i$  will get an overall control policy which is a function of both the CSI  $\mathbf{H}$  and QSI  $\mathbf{Q}^i$ . This is illustrated by the following example.

*Example 1:* Consider a simple example with global CSI state space  $\mathcal{H} = \{\mathbf{H}^1, \mathbf{H}^2\}$  and global QSI state space  $\mathcal{Q} = \{\mathbf{Q}^1, \mathbf{Q}^2\}$ . Hence, the control variables are collectively denoted by the policy  $\Pi = \{\Pi(\mathbf{H}^1, \mathbf{Q}^1), \Pi(\mathbf{H}^2, \mathbf{Q}^1), \Pi(\mathbf{H}^1, \mathbf{Q}^2), \Pi(\mathbf{H}^2, \mathbf{Q}^2)\}$ . Using definition 2, the partitioned actions are simply regroups of variables given by  $\Pi(\mathbf{Q}^1) = \{\Pi(\mathbf{Q}^1, \mathbf{H}^1), \Pi(\mathbf{Q}^1, \mathbf{H}^2)\}$  and  $\Pi(\mathbf{Q}^2) = \{\Pi(\mathbf{Q}^2, \mathbf{H}^1), \Pi(\mathbf{Q}^2, \mathbf{H}^2)\}$ . For any QSI state  $\mathbf{Q}^i$  ( $i = 1, 2$ ), using Lemma 2, the optimal partitioned actions  $\Pi^*(\mathbf{Q}^i)$  can be obtained by solving the R.H.S. of (18) as follows

$$\begin{aligned} \Pi^*(\mathbf{Q}^i) = \arg \min_{\{\Pi(\mathbf{Q}^i, \mathbf{H}^1), \Pi(\mathbf{Q}^i, \mathbf{H}^2)\}} & \left\{ \sum_{k=1}^2 \Pr[\mathbf{H}^k] \left[ g((\mathbf{Q}^i, \mathbf{H}^k), \Pi(\mathbf{Q}^i, \mathbf{H}^k), \gamma) \right. \right. \\ & \left. \left. + \sum_{\mathbf{Q}^j} \Pr[\mathbf{Q}^j|(\mathbf{Q}^i, \mathbf{H}^k), \Pi(\mathbf{Q}^i, \mathbf{H}^k)] V(\mathbf{Q}^j) \right] \right\} \end{aligned} \quad (20)$$

Observe that the R.H.S. of (20) is a decoupled objective function w.r.t. the variables  $\{\Pi(\mathbf{Q}^i, \mathbf{H}^1), \Pi(\mathbf{Q}^i, \mathbf{H}^2)\}$ . Hence, applying standard decomposition theory,  $\forall k = 1, 2$ , we have

$$\Pi^*(\mathbf{Q}^i, \mathbf{H}^k) = \arg \min_{\Pi(\mathbf{Q}^i, \mathbf{H}^k)} \left\{ g((\mathbf{Q}^i, \mathbf{H}^k), \Pi(\mathbf{Q}^i, \mathbf{H}^k), \gamma) + \sum_{\mathbf{Q}^j} \Pr[\mathbf{Q}^j|(\mathbf{Q}^i, \mathbf{H}^k), \Pi(\mathbf{Q}^i, \mathbf{H}^k)] V(\mathbf{Q}^j) \right\}$$

Using the results in Lemma 2, the optimal control of the original problem when the QSI and CSI realizations are  $(\mathbf{Q}^1, \mathbf{H}^2)$  is  $\Pi^*(\mathbf{Q}^1, \mathbf{H}^2)$ . Hence, the solution obtained by solving (18) is adaptive to both the CSI and QSI.

#### IV. DISTRIBUTIVE ONLINE ALGORITHM BASED ON APPROXIMATED MDP

There are still two major obstacles ahead. Firstly, obtaining the value functions  $\{V(\mathbf{Q})\}$  w.r.t. (18) involves solving a system of exponential number of equations and unknowns and brute force solution has exponential complexity. Secondly, even if we could obtain the solution  $\{V(\mathbf{Q})\}$ , the derived control actions will depend on global QSI and CSI, which is highly undesirable. In this section, we shall overcome the above challenges using approximate MDP and distributive stochastic learning. The linear approximation architecture of the value function is given below [21]:

$$V(\mathbf{Q}) = \sum_{m=S}^M \sum_{q=0}^{N_Q} \tilde{V}_m(q) \mathbf{I}[Q_m = q] \quad \text{or in the vector form} \quad \mathbf{V} = \mathbf{M}\mathbf{W}, \quad (21)$$

where we shall refer  $\{\tilde{V}_m(q)\}$  as *per-node value functions*<sup>14</sup> ( $\forall m = S, 1, \dots, M$ ) and  $\{V(\mathbf{Q})\}$  as *global value function* in the rest of this paper,  $\mathbf{V} = [V(\mathbf{Q}^1), \dots, V(\mathbf{Q}^{|\mathcal{Q}|})]^T$  is the vector form of global value functions, the *parameter vector*  $\mathbf{W}$  and *mapping matrix*  $\mathbf{M}$  is given below:

$$\mathbf{W} = \left[ \tilde{V}_S(0), \dots, \tilde{V}_S(N_Q), \tilde{V}_1(0), \dots, \tilde{V}_1(N_Q), \dots, \tilde{V}_S(N_Q), \tilde{V}_M(0), \dots, \tilde{V}_M(N_Q) \right]^T$$

$$\mathbf{M} = \begin{bmatrix} \mathbf{I}[\mathbf{Q}_S^1 = 0] & \dots & \mathbf{I}[\mathbf{Q}_S^1 = N_Q] & \dots & \mathbf{I}[\mathbf{Q}_M^1 = 0] & \dots & \mathbf{I}[\mathbf{Q}_M^1 = N_Q] \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \mathbf{I}[\mathbf{Q}_S^{|\mathcal{Q}|} = 0] & \dots & \mathbf{I}[\mathbf{Q}_S^{|\mathcal{Q}|} = N_Q] & \dots & \mathbf{I}[\mathbf{Q}_M^{|\mathcal{Q}|} = 0] & \dots & \mathbf{I}[\mathbf{Q}_M^{|\mathcal{Q}|} = N_Q] \end{bmatrix},$$

where we let  $\tilde{V}_S(0) = \tilde{V}_1(0) = \dots = \tilde{V}_M(0) = 0$  and set  $\mathbf{Q}^I = (0, \dots, 0)$  (i.e. all buffer empty) as the reference state without loss of generality. Compared with the original value function in (18), the dimension of the per-node value functions is much smaller. Therefore, the per-node value function can only satisfy the Bellman equation (18) in some pre-determined system queue states. In this paper, we shall refer the pre-determined subset of system queue states as the *representative states* [21]. Without loss of generality, we define the reference states  $\mathcal{Q}_R = \{\beta_{m,q} | \forall m = S, 1, 2, \dots, M; q = 1, 2, \dots, N_Q\}$ , where  $\beta_{m,q}$  denotes the QSI with  $Q_m = q$  and  $Q_n = 0 \forall n \neq m$ . Moreover, we also define the inverse mapping matrix  $\mathbf{M}^{-1}$  as

$$\mathbf{M}^{-1} = \begin{bmatrix} 0 & \mathbf{I}[\mathbf{Q}^1 = \beta_{S,1}] & \dots & \mathbf{I}[\mathbf{Q}^1 = \beta_{S,N_Q}], & \dots & , 0 & \mathbf{I}[\mathbf{Q}^1 = \beta_{M,1}] & \dots & \mathbf{I}[\mathbf{Q}^1 = \beta_{M,N_Q}] \\ \dots & \dots \\ 0 & \mathbf{I}[\mathbf{Q}^{|\mathcal{Q}|} = \beta_{S,1}] & \dots & \mathbf{I}[\mathbf{Q}^{|\mathcal{Q}|} = \beta_{S,N_Q}], & \dots & , 0 & \mathbf{I}[\mathbf{Q}^{|\mathcal{Q}|} = \beta_{M,1}] & \dots & \mathbf{I}[\mathbf{Q}^{|\mathcal{Q}|} = \beta_{M,N_Q}] \end{bmatrix}^T.$$

Thus, we have  $\mathbf{W} = \mathbf{M}^{-1}\mathbf{V}$ . Instead of offline computing the *best fit* parameter vector  $\mathbf{W}$  (per-node value function vector) w.r.t. the global value function  $\mathbf{V}$  (which is quite complex), we shall propose an online learning algorithm to estimate the parameter vector  $\mathbf{W}$  (per-node value function) in Section IV-B.

#### A. Distributive Control Policy under Linear Value Function Approximation

Using the approximate value function in (21), we shall derive a distributive control policy which depends on the local CSI and local QSI as well as the per-node value functions  $\{\tilde{V}_m(q)\}$  at each node  $m$  ( $\forall m = S, 1, \dots, M$ ). Specifically, using the approximation in (21), the control policy in (18) can be obtained by solving the following simplified optimization problem.

<sup>14</sup>In this paper, we assume each RS (say the  $m$ -th RS) has the knowledge of the source node's queue length  $Q_S$  and its own queue length  $Q_m$ . Therefore, the per-node value function  $\tilde{V}_S$  and  $\tilde{V}_m$  is known at the  $m$ -th RS.

*Problem 2 (Optimal Control Action with Approximated Value Function):* For any given value function  $V(\mathbf{Q}^i) = \sum_{m=S}^M \sum_{q=0}^{N_Q} \tilde{V}_m(q) \mathbf{I}[Q_m^i = q]$ , the optimal control policy is given by

$$\begin{aligned} \Pi^*(\mathbf{Q}^i) &= \arg \min_{\Pi(\mathbf{Q}^i)} \left\{ \bar{g}(\mathbf{Q}^i, \Pi(\mathbf{Q}^i), \gamma) + \sum_{\mathbf{Q}^j} \Pr[\mathbf{Q}^j | \mathbf{Q}^i, \Pi(\mathbf{Q}^i)] V(\mathbf{Q}^j) \right\} \\ &= \arg \left\{ \sum_{m=S}^M Q_m^i + \gamma_{S,d} \mathbf{I}[Q_S^i = N_Q] + \sum_n f_X(n) V(\mathbf{Q}_{S,n}^i) \right. \\ &\quad \left. + \min_{\Pi(\mathbf{Q}^i)} \mathbf{E}_{\mathbf{H}} \left[ \sum_{m, N_{SR}} I_{S,m}^{N_{SR}} G_{S,m}(N_{SR}, p_{S,m}) + \sum_{n, N_{RD}} I_{n,D}^{N_{RD}} G_{n,D}(N_{RD}, p_{n,D}) \right] \right\} \\ &\Leftrightarrow \arg \min_{\Pi(\mathbf{Q}^i)} \mathbf{E}_{\mathbf{H}} \left[ \sum_{m, N_{SR}} I_{S,m}^{N_{SR}} G_{S,m}(N_{SR}, p_{S,m}) + \sum_{n, N_{RD}} I_{n,D}^{N_{RD}} G_{n,D}(N_{RD}, p_{n,D}) \right] \end{aligned} \quad (22)$$

where  $\mathbf{Q}_{S,n}^i = [Q_S^i + n, Q_1^i, Q_2^i, \dots, Q_M^i]$  and  $G_{S,m}(N_{SR}, p_{S,m}) = \gamma_{S,p} p_{S,m} + \sum_n f_X(n) (\tilde{V}_S(Q_S^i - R_{S,m}(N_{SR}, p_{S,m}) + n) - \tilde{V}_S(Q_S^i + n)) + \tilde{V}_m(Q_m^i + R_{S,m}(N_{SR}, p_{S,m})) - \tilde{V}_m(Q_m^i)$  and  $G_{n,D}(N_{RD}, p_{n,D}) = \gamma_{n,p} p_{n,D} + \tilde{V}_n(Q_n^i - R_{n,D}(N_{RD}, p_{n,D})) - \tilde{V}_n(Q_n^i)$ .

The solution of Problem 2 is summarized in Lemma 3 below.

*Lemma 3 (Distributive Control Policy):* Given the per-node value functions  $\{\tilde{V}_m(q)\}$  ( $\forall m = S, 1, \dots, M$ ) and any realization of CSI  $\mathbf{H}$  and QSI  $\mathbf{Q}^{i15}$ , the following distributive control solves the Problem 2:

- Power control for the S-R link and R-D link ( $\forall m = 1, \dots, M$ ):

$$p_{S,m}^*(N_{SR}) = \arg \min_{p_{S,m}} G_{S,m}(N_{SR}, p_{S,m}) \quad \text{and} \quad p_{m,D}^*(N_{RD}) = \arg \min_{p_{m,D}} G_{n,D}(N_{RD}, p_{n,D}) \quad (23)$$

where  $N_{SR}, N_{RD} = 0, 1, \dots, \min(N_T, N_R)$ .

- First-stage bid at RSs ( $\forall m = 1, \dots, M$ ):

$$A_m^*(N_{SR}) = G_{S,m}(N_{SR}, p_{S,m}^*(N_{SR})) \quad (24)$$

where  $N_{SR} = 0, 1, \dots, \min(N_T, N_R)$ .

- Second-stage bid at RSs ( $\forall n = 1, \dots, M$ ):

$$\begin{aligned} (I_n, N_{SR,n}) &= \arg \min_{(m, N_{SR})} \left\{ A_m^*(N_{SR}) + G_{n,D}(N_{RD}, p_{n,D}^*(N_{RD})) \right\} \\ B_n^* &= G_{S,I_n}(N_{SR,n}, p_{S,m}^*(N_{SR,n})) + G_{n,D}(N_{RD,n}, p_{n,D}^*(N_{RD,n})) \end{aligned} \quad (25)$$

where  $N_{RD} = \min(N_T, N_R - N_{SR})$ .

<sup>15</sup>Note that the following expressions are all functions of the systems state. We omit the system state for notation simplicity when the meaning is clear.

In addition, for sufficiently large source arrival rate  $\lambda_S, \frac{N_Q}{\lambda_S}$  and the average transmit power constraints  $\{P_S, P_R\}$ , the power control policy in (23) has the following closed-form expression:

$$p_{S,m}^*(N_{SR}) = \frac{N_{SR} [\tilde{V}'_S(Q_S^i) - \tilde{V}'_m(Q_m^i)]}{\gamma_{S,p} \ln 2} - \sum_{j=1}^{N_{SR}} \frac{1}{\eta_{S,m}^j} \quad (26)$$

$$p_{m,D}^*(N_{RD}) = \frac{N_{RD} \tilde{V}'_m(Q_m^i)}{\gamma_{m,p} \ln 2} - \sum_{j=1}^{N_{RD}} \frac{1}{\eta_{m,D}^j}, \quad (27)$$

where  $\tilde{V}'_S(Q_S^i) = \frac{\tilde{V}_S(Q_S^i+1) - \tilde{V}_S(Q_S^i-1)}{2}$  and  $\tilde{V}'_m(Q_m^i) = \frac{\tilde{V}_m(Q_m^i+1) - \tilde{V}_m(Q_m^i-1)}{2}$ .

*Proof:* Please refer to Appendix C. ■

*Remark 4 (Multi-level Water-Filling Structure of the Control Policy):* The power control policy (26) and (27) as well as the RS selection and data stream allocation control policy in (24) and (25) are functions of both the CSI and QSI where they depend on the QSI indirectly via the per-node value functions  $\{\tilde{V}_m(q)\}$  ( $\forall m = S, 1, \dots, M$ ). The power control solution has the form of multi-level water-filling where the power is allocated according to the CSI while the water-level is adaptive to the QSI.

*B. Online Distributive Stochastic Learning Algorithm to Estimate the Per-node Value Functions  $\{\tilde{V}_m(q)\}$  and the LMs  $\{\gamma_{S,d}, \gamma_{S,p}, \gamma_{m,p}\}$*

In Lemma 3, the control actions are functions of per-node value functions  $\{\tilde{V}_m(q)\}$  and the LMs  $\{\gamma_{S,d}, \gamma_{S,p}, \gamma_{m,p}\}$ . In this section, we propose an online learning algorithm to determine the per-node value functions and the LMs realtime. The almost-sure convergence proof of this algorithm is provided in the next section. The system procedure of the proposed distributive online learning is given below.

- **Step 1 [Initialization]:** Each RS  $m$  initiates its per-node value functions and LMs, denoted as  $\{\tilde{V}_m^0(q)\}$  and  $\gamma_{m,p}^0$ , as well as the per-node value functions and LMs for the source node, denoted as  $\{\tilde{V}_S^0(q)\}$  and  $\{\gamma_{S,p}^0, \gamma_{S,d}^0\}$ . The initialization of  $\tilde{V}_S^0$  and  $\{\gamma_{S,p}^0, \gamma_{S,d}^0\}$  at each RS should be the same.
- **Step 2 [Determination of control actions]:** At the beginning of the  $t$ -th frame, the source node broadcasts its QSI  $Q_S(t)$  to the RS nodes. Based on the local system information  $(Q_S(t), Q_m(t), \mathbf{H}_m(t))$  and the per-node value functions  $\{\tilde{V}_m^t(q)\}$  and  $\{\tilde{V}_S^t(q)\}$ , each RS  $m$  determines the distributive control actions including the S-R and R-D power allocation  $p_{S,m}^*(N_{SR}, t), p_{m,D}^*(N_{RD}, t)$  the first-stage bid  $A_m^*(N_{SR}, t)$  ( $N_{SR} = 1, \dots, N_{\min}$ ) as well as the second-stage bid  $B_m^*(t), I_n(t), N_{SR,n}(t)$  according to Lemma 3. Based on the contention resolution protocol described in Section II-D, the Rx-RS and the Tx-RS pair is given by  $(m^*(t), n^*(t))$  (where  $n^*(t) = \arg \min_n B_n^*(t)$  and

$m^*(t) = I_{n^*(t)}(t)$  and the corresponding number of data streams pair is given by  $(N_{SR}^*(t), N_{RD}^*(t))$  (where  $N_{SR}^*(t) = N_{SR, n^*(t)}(t)$  and  $N_{RD}^*(t) = N_{RD, n^*(t)}(t)$ ).

- **Step 3 [Per-node value functions and LMs update]:** Each RS  $m$  updates the per-node value function  $\{\tilde{V}_S^{t+1}(q)\}, \{\tilde{V}_m^{t+1}(q)\}$  as well as the LMs  $\{\gamma_{S,d}^{t+1}, \gamma_{S,p}^{t+1}, \gamma_{m,p}^{t+1}\}$  according to Algorithm 1. Finally, let  $t = t + 1$  and go to Step 2.

*Algorithm 1 (Online distributive learning algorithm for per-node value functions and LMs):*

$$\tilde{V}_m^{t+1}(q) = \tilde{V}_m^t(q) + \epsilon_v^t \left[ \gamma_{S,d} \mathbf{I}[Q_S(t) = N_Q] + q + B_{n^*(t)}^*(t) - \tilde{V}_m^t(q) \right] \mathbf{I}[\mathbf{Q}(t) = \beta_{m,q}], m = S, 1, \dots, M \quad (28)$$

$$\gamma_{S,d}^{t+1} = \left( \gamma_{S,d}^t + \epsilon_d^t (\mathbf{I}[Q_S(t) = N_Q] - D) \right)^+ \quad (29)$$

$$\gamma_{S,p}^{t+1} = \left( \gamma_{S,p}^t + \epsilon_p^t \left( \sum_{N_{SR}=1}^{N_{\min}} I_{S,m}^{N_{SR}}(t) p_{S,m}(N_{SR}, t) - P_S \right) \right)^+ \quad (30)$$

$$\gamma_{m,p}^{t+1} = \left( \gamma_{m,p}^t + \epsilon_p^t \left( \sum_{N_{SR}=1}^{N_{\min}} I_{m,D}^{N_{RD}}(t) p_{m,D}(N_{RD}, t) - P_R \right) \right)^+, m = 1, 2, \dots, M \quad (31)$$

where  $I_{S,m}^{N_{SR}}(t) = \mathbf{I}[m = m^*(t)] \mathbf{I}[N_{SR} = N_{SR}^*(t)]$ ,  $I_{m,D}^{N_{RD}}(t) = \mathbf{I}[m = n^*(t)] \mathbf{I}[N_{RD} = N_{RD}^*(t)]$ , and  $\{\epsilon_v^t > 0\}, \{\epsilon_d^t > 0\}, \{\epsilon_p^t > 0\}$  are the step size sequences satisfying

$$\sum_{t=0}^{\infty} \epsilon_v^t = \infty, \sum_{t=0}^{\infty} \epsilon_p^t = \infty, \sum_{t=0}^{\infty} \epsilon_d^t = \infty, \sum_{t=0}^{\infty} \left[ (\epsilon_v^t)^2 + (\epsilon_p^t)^2 + (\epsilon_d^t)^2 \right] < \infty, \lim_{t \rightarrow +\infty} \frac{\epsilon_p^t}{\epsilon_v^t} = 0, \lim_{t \rightarrow +\infty} \frac{\epsilon_d^t}{\epsilon_v^t} = 0.$$

### C. Almost-Sure Convergence of Distributive Stochastic Learning

In this section, we shall establish technical conditions for the almost-sure convergence of the online distributive learning algorithm. Since  $\{\epsilon_v^t\}, \{\epsilon_p^t\}, \{\epsilon_d^t\}$  satisfy  $\epsilon_p^t = \mathbf{o}(\epsilon_v^t)$ ,  $\epsilon_d^t = \mathbf{o}(\epsilon_p^t)$ , the LMs update and the per-node potential functions update are done simultaneously but over two different time scales. During the per-node potential functions update (timescale I), we have  $\gamma_p^{t+1} - \gamma_p^{t+1} = \mathcal{O}(\epsilon_p^t) = \mathbf{o}(\epsilon_v^t)$  and  $\gamma_{S,d}^{t+1} - \gamma_{S,d}^{t+1} = \mathcal{O}(\epsilon_d^t) = \mathbf{o}(\epsilon_v^t)$ . Therefore, the LMs appear to be quasi-static [22] during the per-node value function update in (28). For the notation convenience, define the sequences of matrices  $\{\mathbf{A}^t\}$  and  $\{\mathbf{B}^t\}$  as  $\mathbf{A}^{t-1} = (1 - \epsilon_v^{t-1}) \mathbf{I} + \mathbf{M}^{-1} \mathbf{P}(\Pi^t) \mathbf{M} \epsilon_v^{t-1}$  and  $\mathbf{B}^{t-1} = (1 - \epsilon_v^{t-1}) \mathbf{I} + \mathbf{M}^{-1} \mathbf{P}(\Pi^{t-1}) \mathbf{M} \epsilon_v^{t-1}$ , where  $\Pi^t$  is a unichain system control policy at the  $t$ -th frame,  $\mathbf{P}(\Pi^t)$  is the transition matrix of system states given the unichain system control policy  $\Pi^t$ ,  $\mathbf{I}$  is identity matrix. The convergence property of the per-node value function update is given below:

*Lemma 4 (Convergence of Per-Node Value Function Learning over Timescale I):* Assume for all the feasible policy in the policy space, there exists some positive integer  $\beta$  and  $\tau^\beta > 0$  such that

$$[\mathbf{A}^{\beta-1} \dots \mathbf{A}^1]_{(a,I)} \geq \tau^\beta, \quad [\mathbf{B}^{\beta-1} \dots \mathbf{B}^1]_{(a,I)} \geq \tau^\beta \quad \forall a, \quad (32)$$

where  $[\cdot]_{(a,I)}$  denotes the element in  $a$ -th row and  $I$ -th column (where  $I$  corresponds to the reference state  $\mathbf{Q}^I$ ) and  $\tau^t = \mathcal{O}(\epsilon_v^t)$  ( $\forall t$ ). The following statements are true:

- The update of the parameter vector (or per-node potential vector) will converge almost surely for any given initial parameter vector  $\mathbf{W}^0$  and LMs  $\gamma$ , i.e.  $\lim_{t \rightarrow \infty} \mathbf{W}^t(\gamma) = \mathbf{W}^\infty(\gamma)$ .
- The steady state parameter vector  $\mathbf{W}^\infty$  satisfies:

$$\theta \mathbf{e} + \mathbf{W}^\infty(\gamma) = \mathbf{M}^{-1} \mathbf{T}(\gamma, \mathbf{M} \mathbf{W}^\infty(\gamma)) \quad (33)$$

where  $\theta$  is a constant,  $\mathbf{W}^\infty$  is given by

$$\mathbf{W}^\infty = [\tilde{V}_S^\infty(0), \dots, \tilde{V}_S^\infty(N_Q), \tilde{V}_1^\infty(0), \dots, \tilde{V}_1^\infty(N_Q), \dots, \tilde{V}_S^\infty(N_Q), \tilde{V}_M^\infty(0), \dots, \tilde{V}_M^\infty(N_Q)]^T,$$

and the mapping  $\mathbf{T}$  is defined as  $\mathbf{T}(\gamma, \mathbf{V}) = \min_{\Pi} [\bar{\mathbf{g}}(\gamma, \Pi) + \mathbf{P}(\Pi) \mathbf{V}]$ .

*Proof: Please refer to Appendix D.* ■

*Remark 5 (Interpretation of the Sufficient Conditions in Lemma 4):* Note that  $A^t$  and  $B^t$  are related to the transition probability of the reference states. Condition (32) simply means that there is one reference state accessible from all the other reference states after some finite number of transition steps. This is a very mild condition and will be satisfied in most of the cases in practice.

Note that (33) is equivalent to the following Bellman equation on the representative states  $\mathcal{S}_R$ :

$$\theta + \tilde{V}_m^\infty(q) = \min_{\Pi(\beta_{m,q})} \left\{ \bar{g}(\beta_{m,q}, \Pi(\beta_{m,q}), \gamma_k) + \sum_{\mathbf{Q}^j} \Pr[\mathbf{Q}^j | \beta_{m,q}, \Pi(\beta_{m,q})] \sum_{m=S}^M \tilde{V}_m^\infty(Q_m^j) \right\}, \quad \forall \beta_{m,q} \in \mathcal{S}_R.$$

Hence, Lemma 4 basically guarantees the proposed online learning algorithm will converge to the *best fit* parameter vector (per-node potential) satisfying (21). On the other hand, since the ratio of step sizes satisfies  $\frac{\epsilon_p^t}{\epsilon_v^t}, \frac{\epsilon_d^t}{\epsilon_v^t} \rightarrow 0$  during the LM update (timescale II), the per-node value function will be updated much faster than the Lagrange multipliers. Hence, the update of Lagrange multipliers in timescale II will trigger another update process of the per-node value function in timescale I. By the Corollary 2.1 of [23], we have  $\lim_{t \rightarrow \infty} \|\tilde{\mathbf{V}}_m^t - \tilde{\mathbf{V}}_m^\infty(\gamma^t)\| = 0$  w.p.1. Hence, during the LM updates in (31), (30) and (29), the per-node value function update in (28) is seen as almost equilibrated. Moreover, convergence of the LMs is summarized below.

*Lemma 5 (Convergence of the LMs over Timescale II):* The iteration on the vector of LMs  $\gamma = [\gamma_{S,d}, \gamma_{S,p}, \gamma_{1,p}, \dots, \gamma_{M,p}]^T$  converges almost surely to  $\gamma^* = [\gamma_{S,d}^*, \gamma_{S,p}^*, \gamma_{1,p}^*, \dots, \gamma_{M,p}^*]^T$ , which satisfies the power and packet drop rate constraints in (14),(15) and (13).

*Proof: Please refer to Appendix E.* ■

Based on the above lemmas, we summarized the convergence performance of the online per-node value functions and LMs learning algorithm in the following theorem.

*Theorem 1 (Convergence of Online Learning Algorithm 1):* For the same conditions as in Lemma 4, we have  $(\gamma^t, \mathbf{W}^t) \rightarrow (\gamma^*, \mathbf{W}^\infty(\gamma^*))$  w.p.1., where  $(\gamma^*, \mathbf{W}^\infty(\gamma^*))$  satisfies  $\theta \mathbf{e} + \mathbf{W}^\infty(\gamma^*) = \mathbf{M}^{-1} \mathbf{T}(\gamma^*, \mathbf{M} \mathbf{W}^\infty(\gamma^*))$  and the average power constraint (14,15) as well as the average packet drop rate constraint (13), where  $\mathbf{e}$  is a  $(M+1)(N_Q+1) \times 1$  vector with all elements equal to 1.

#### D. Asymptotic Optimality

Finally, we shall show that the performance of the distributive algorithm is asymptotically global optimal for high traffic loading.

*Theorem 2 (Asymptotically Global Optimal at High Traffic Loading):* For sufficiently large  $N_Q$  and high traffic loading such that the optimization problem in Problem 1 is feasible, the performance of the proposed distributive control algorithm is asymptotically global optimal.

*Proof: Please refer to Appendix F.* ■

## V. SIMULATIONS AND DISCUSSIONS

In this section, we shall compare our proposed online per-node value function learning algorithm to five reference baselines. Baseline 1 and 4 refer to the proposed *buffered decode and forward* (BDF) protocol with *throughput optimal policy* (in stability sense), namely the *dynamic backpressure* algorithm [24], where we utilize full-duplex RSs in Baseline 1 and half-duplex RSs in Baseline 4. Baseline 2 and 5 refer to the regular decode-and-forward protocol (DF) with the *CSIT only scheduling* (the link selection and power allocation are adaptive to the CSIT only so as to optimize the end-to-end throughput). We utilize full-duplex RSs in Baseline 2 and half-duplex RSs in Baseline 5. Moreover, Baseline 3 refers to the proposed BDF protocol with CSIT only scheduling and half-duplex RSs. In the simulations, we assume the total bandwidth is 1 MHz, the packet arrival at the source node is Poisson with average arrival rate  $\lambda_S = 200\text{pck/s}$  and deterministic packet size  $N_b$  bits. The number of antennas at the source node and the

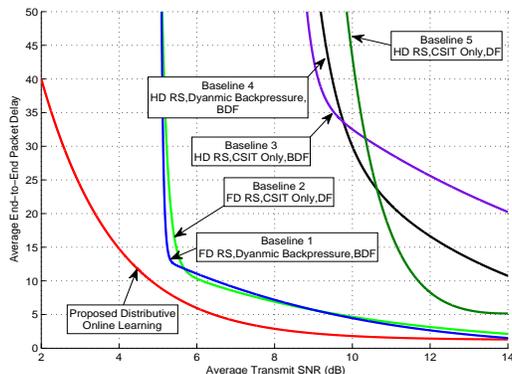


Fig. 3. Average end-to-end delay versus average transmit SNR. Baseline 1 refers to the dynamic backpressure algorithm with BDF protocol and full-duplex relays. Baseline 2 refers to the CSIT only scheduling with traditional DF protocol and full-duplex relays. Baseline 3 refers to the CSIT only scheduling with BDF protocol and half-duplex relays. Baseline 4 refers to the dynamic backpressure algorithm with BDF protocol and half-duplex relays. Baseline 5 refers to the CSIT only scheduling with traditional DF protocol and half-duplex relays. The deterministic packet size is  $N_b = 25K$  bits and the number of antennas at each RS is  $N_R = 4$ . The packet drop rates of the Baselines 1-5 and the proposed distributive online learning are 0.2% 0.2% 13%, 3%, 24% and 0.2% respectively.

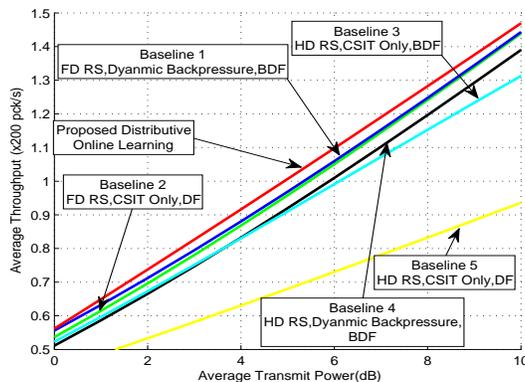


Fig. 4. Average throughput versus average transmit SNR. The deterministic packet size is  $N_b = 30K$  bits and the number of antennas at each RS is  $N_R = 4$ . The packet drop rates of the Baselines 1-5 and the proposed distributive online learning are all 10%.

destination node is  $N_T = 2$ . Moreover, the maximum buffer size of each node (source node and RSs) is  $N_Q = 10$ .

Figure 3 and Figure 4 illustrate the average end-to-end delay and average throughput versus average transmit SNR per node with  $N_R = 4$  antennas at each RS, respectively. It can be observed that the proposed

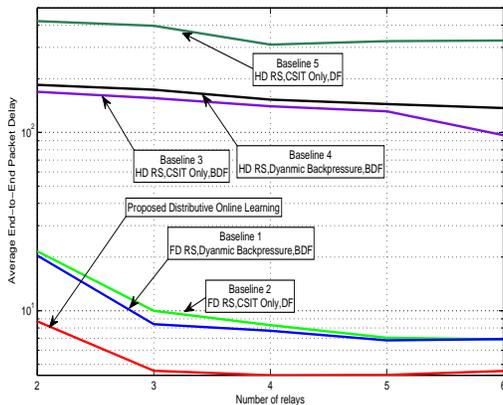


Fig. 5. Average end-to-end delay versus the number of relays with transmit SNR = 5.5dB. The deterministic packet size is  $N_b = 25K$  bits and the number of antennas at each RS is  $N_R = 4$ . The packet drop rates of the Baseline 1-5 and the proposed distributive online learning are 23%, 23%, 86%, 82%, 96% and 0.5% respectively.

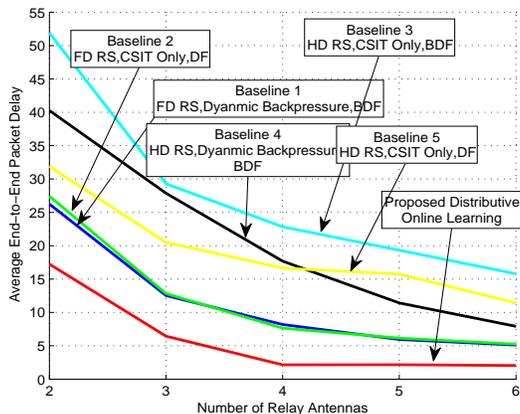


Fig. 6. Average end-to-end delay versus the number of relay antennas with transmit SNR = 5dB. The deterministic packet size is  $N_b = 20K$  bits and the number of antennas at each RS is  $N_R = 4$ . The packet drop rates of the Baseline 1-5 and the proposed distributive online learning are 3%, 4%, 9%, 5%, 20% and 0.1%, respectively.

distributive algorithm with half-duplex RS could achieve significant performance gain in both average delay and average throughput over all baselines with full-duplex RSs, and even more significant gain over the baselines with half-duplex RSs. This illustrates the advantages of the proposed BDF algorithm with distributive delay-optimal control policy, which could effectively reduce the intrinsic half-duplex penalty in the cooperative communication systems.

Figure 5 and Figure 6 illustrate the average end-to-end delay versus the number of relays and the number of relay antennas with  $N_R = 4$  antennas at each RS, respectively. It can be observed that the average delay of all the schemes decreases as the number of relays or the number of relay antennas increases. Furthermore, the proposed BDF algorithm with distributive delay-optimal control policy has significant gain in delay over all the baselines.

Figure 7 illustrates the convergence property of the proposed distributive online learning algorithm. We plot the per-node value function of the first relay versus scheduling slot index at a transmit SNR= 10dB. The average delay at the 200-th scheduling slot is already very close to the steady-state value, which is much better than all the baselines. Furthermore, unlike the iterations in deterministic NUM problems, the proposed algorithm is online, meaning that normal payload is delivered during the iteration steps.

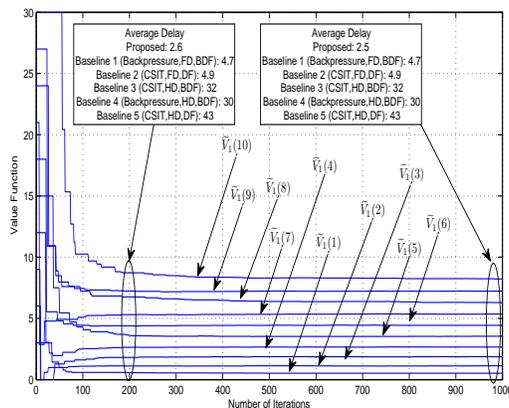


Fig. 7. Illustration of the convergence of the proposed online learning algorithm. The instantaneous per-node value function is plotted versus time slot index for a cooperative MIMO system with a source node (with 2 antennas) and 2 RS nodes (each with 4 antennas). The transmit SNR of the source and the RS nodes are 10 dB and the target packet drop rate is 0.2%. Unlike the iterations in deterministic NUM problems, the proposed algorithm is online, meaning that normal payload is delivered during the iteration steps.

## VI. SUMMARY

In this paper, we consider queue-aware resource control for two-hop cooperative MIMO systems. We show that by exploiting buffering in each MIMO relay, we could substantially reduce the intrinsic half-duplex loss in cooperative systems. The delay-optimal resource control policy is formulated as an average-cost infinite horizon Markov Decision Process (MDP). To obtain a low complexity solution, we approximate

the value function by a linear combination of per-node value functions. The per-node value function is obtained using a distributive stochastic learning algorithm. We also established technical conditions for almost-sure convergence and show that in heavy traffic limit, the proposed low complexity distributive algorithm converges to global optimal solution.

#### APPENDIX A: PROOF OF LEMMA 1

The average number of bits received by the source node is given by  $\lambda_S(1 - D)$ , which is also the average number of information bits received by the relay clusters as the source node and the relay cluster are cascade. Let  $W$ ,  $W_S$  and  $W_R$  be the average time (with the unit of frames) one information bit staying in the system, the source node's queue and some relay's queue respectively,  $N_S$  and  $N_R$  be the average number of information bits in the source node's queue and the relays' queues respectively, we have  $N_S = (1 - D)\lambda_S W_S$  and  $N_R = (1 - D)\lambda_S W_R$  by Little's Law. Notice that  $W = W_S + W_R$ , we have  $W = \frac{N_S + N_R}{\lambda_S(1 - D)}$ . Since the change of system queue state forms a Markov chain, we have  $W = \mathbf{E}_{\pi_\kappa} \left[ \frac{Q_S + \sum_{m=1}^M Q_m}{\lambda_S(1 - D)} \right]$ , where  $\pi_\kappa$  is the steady state distribution. For sufficiently small packet drop rate requirement  $1 - D \approx 1$ , the end to end average delay becomes  $W = \mathbf{E}_{\pi_\kappa} \left[ \frac{Q_S + \sum_{m=1}^M Q_m}{\lambda_S} \right]$ .

#### APPENDIX B: PROOF OF LEMMA 2

From the Bellman equation of the original state space (18), we have

$$\begin{aligned} \theta + V(\mathbf{Q}^i, \mathbf{H}) &= \min_{\Pi(\mathbf{Q}^i, \mathbf{H})} \left\{ g((\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H}), \gamma) + \sum_{(\mathbf{Q}^j, \mathbf{H}')} \Pr [(\mathbf{Q}^j, \mathbf{H}') | (\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H})] J(\mathbf{Q}^j, \mathbf{H}') \right\} \\ &\stackrel{(a)}{=} \min_{\Pi(\mathbf{Q}^i, \mathbf{H})} \left\{ g((\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H}), \gamma) + \sum_{\mathbf{Q}^j} \Pr [\mathbf{Q}^j | (\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H})] V(\mathbf{Q}^j) \right\}, \end{aligned} \quad (34)$$

where (a) is due to the definition  $V(\mathbf{Q}^j) = \mathbf{E}_{\mathbf{H}'} [V(\mathbf{Q}^j, \mathbf{H}') | \mathbf{Q}^j]$ , and the optimal control actions are given by  $\Pi^*(\mathbf{Q}^i, \mathbf{H}) = \arg \min_{\Pi(\mathbf{Q}^i, \mathbf{H})} \left\{ g((\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H}), \gamma) + \sum_{\mathbf{Q}^j} \Pr [\mathbf{Q}^j | (\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H})] V(\mathbf{Q}^j) \right\}$ . Thus, by the partitioning of the optimal control actions in Definition 1, i.e.  $\Pi^*(\mathbf{Q}^i) = \{\Pi^*(\mathbf{Q}^i, \mathbf{H}) | \forall \mathbf{H}\}$ ,

$$\Pi^*(\mathbf{Q}^i) = \arg \min_{\Pi(\mathbf{Q}^i)} \sum_{\mathbf{H}} \Pr(\mathbf{H}) \left\{ g((\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H}), \gamma) + \sum_{\mathbf{Q}^j} \Pr [\mathbf{Q}^j | (\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H})] V(\mathbf{Q}^j) \right\} \quad (35)$$

From (34) and (35), we have  $\theta + \Pr(\mathbf{H})V(\mathbf{Q}^i, \mathbf{H}) = \min_{\Pi(\mathbf{Q}^i)} \sum_{\mathbf{H}} \Pr(\mathbf{H}) \left\{ g((\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H}), \gamma) + \sum_{\mathbf{Q}^j} \Pr [\mathbf{Q}^j | (\mathbf{Q}^i, \mathbf{H}), \Pi(\mathbf{Q}^i, \mathbf{H})] V(\mathbf{Q}^j) \right\} \stackrel{(b)}{=} \min_{\Pi(\mathbf{Q}^i)} \left\{ \bar{g}(\mathbf{Q}^i, \Pi(\mathbf{Q}^i), \gamma) + \sum_{\mathbf{Q}^j} \Pr [\mathbf{Q}^j | \mathbf{Q}^i, \Pi(\mathbf{Q}^i)] V(\mathbf{Q}^j) \right\}$ , where the equality (b) is due to the definition of  $\bar{g}$  in (19). As a result, the control policy obtained by solving (18) is the same as that obtained by solving (17) and this completes the proof.

## APPENDIX C: PROOF OF LEMMA 3

We shall prove the general control policy first, followed by the closed-form power control derivation.

According to (22), given  $N_{SR}$  and  $N_{RD}$ , the optimal power control is given by:

$$\begin{aligned} & \min_{\{\Pi_p^m\}} \mathbf{E}_{\mathbf{H}} \left[ \sum_{m, N_{SR}} I_{S,m}^{N_{SR}} G_{S,m}(N_{SR}, p_{S,m}) + \sum_{n, N_{RD}} I_{n,D}^{N_{RD}} G_{n,D}(N_{RD}, p_{n,D}) \right] \\ & = \mathbf{E}_{\mathbf{H}} \left[ \sum_{m, N_{SR}} I_{S,m}^{N_{SR}} \min_{p_{S,m}} G_{S,m}(N_{SR}, p_{S,m}) + \sum_{n, N_{RD}} I_{n,D}^{N_{RD}} \min_{p_{n,D}} G_{n,D}(N_{RD}, p_{n,D}) \right] \end{aligned}$$

Therefore,  $p_{S,m}^*(N_{SR}) = \arg \min_{p_{S,m}} G_{S,m}(N_{SR}, p_{S,m})$  and  $p_{n,D}^*(N_{RD}) = \arg \min_{p_{n,D}} G_{n,D}(N_{RD}, p_{n,D})$ .

To determine the optimal Rx-RS, Tx-RS and stream allocation, the bidding is divided into two stages:

- First Biding: Each RS (say the  $m$ -th RS) broadcasts one bid for each possible  $N_{SR}$  indicating that if itself is selected as Rx-RS and the number of S-R streams is  $N_{SR}$ , what would be the corresponding  $G_{S,m}(N_{SR}, p_{S,m}^*)$ .
- Second Biding: After receiving the bids in the first round, each RS (say the  $n$ -th RS) should calculate that if itself is selected as the Tx-RS, which RS else is the best Rx-RS (say the  $m$ -th RS is the best Rx-RS), what's the best  $N_{SR}$  and  $N_{RD}$  and what's the corresponding  $B_n^* = G_{S,m}(N_{SR}, p_{S,m}^*) + G_{n,D}(N_{RD}, p_{n,D}^*)$ . Then, broadcast the calculation results  $B_n^*$  as the second bid.
- After comparing the  $B_n^*$ , the optimal Rx-RS, Tx-RS and stream allocation can be determined.

Therefore, the first-stage bidding and the second-stage bidding is straight-forward.

When  $\lambda_S$  and  $\frac{N_Q}{\lambda}$  ( $m = S, 1, 2, \dots, M$ ) are sufficiently large, it with large probability that  $\frac{Q_m}{\lambda}$  ( $m = S, 1, 2, \dots, M$ ) is sufficiently large. Hence, following a similar approach in [25], it can be proved that the value function  $\tilde{V}_m$  ( $m = S, 1, 2, \dots, M$ ) is increasing polynomially in  $Q = [Q_S, Q_1, \dots, Q_M]^T$ . The optimization on  $p_{S,m}$  is given by

$$\begin{aligned} & p_{S,m}^*(N_{SR}) = \arg \min_{p_{S,m}} G_{S,m}(N_{SR}, p_{S,m}) \\ & = \arg \min_{p_{S,m}} \left\{ \gamma_{S,p} p_{S,m} + \sum_n f_X(n) \left( \tilde{V}_S(Q_S^i - R_{S,m}(N_{SR}, p_{S,m}) + n) - \tilde{V}_S(Q_S^i + n) \right) \right. \\ & \quad \left. + \tilde{V}_m(Q_m^i + R_{S,m}(N_{SR}, p_{S,m})) - \tilde{V}_m(Q_m^i) \right\}. \end{aligned} \quad (36)$$

Similar to [25], we can do Taylor expansion as follows:

$$\tilde{V}_S(Q_S^i - R_{S,m}(N_{SR}, p_{S,m}) + n) = \tilde{V}_S(Q_S^i) + \left( n - R_{S,m}(N_{SR}, p_{S,m}) \right) \tilde{V}_S'(Q_S^i), \quad (37)$$

$$\tilde{V}_S(Q_S^i + n) = \tilde{V}_S(Q_S^i) + n \tilde{V}_S'(Q_S^i) \quad (38)$$

where  $\tilde{V}'_S$  is the first order derivative on  $\tilde{V}_S$  and the higher order is neglectable. Same approach can be used to expand  $\tilde{V}_m(Q_m^i + R_{S,m}(N_{SR}, p_{S,m}))$  as  $\tilde{V}_m(Q_m^i + R_{S,m}(N_{SR}, p_{S,m})) = \tilde{V}_m(Q_m^i) + R_{S,m}(N_{SR}, p_{S,m})\tilde{V}'_m(Q_m^i)$ . At high SNR region, we have

$$\frac{\partial R_{S,m}(N_{SR}, p_{S,m})}{\partial p_{S,m}} = \frac{N}{\ln 2} \frac{1}{p_{S,m} + \sum_{j=1}^{N_{SR}} \frac{1}{\eta_{S,m}^j}}. \quad (39)$$

According to (37,38,39), taking derivative on the RHS of (36) and letting it be zero, we can get the closed-form expression for power allocation in (26). Moreover, (27) can be proved in the same way. Finally, when  $Q_m$  and  $Q_S$  are sufficiently large, according to the definition of derivative, we have

$$\tilde{V}'_S(Q_S^i) = \frac{\tilde{V}_S(Q_S^i + 1) - \tilde{V}_S(Q_S^i - 1)}{2} \quad \tilde{V}'_m(Q_m^i) = \frac{\tilde{V}_m(Q_m^i + 1) - \tilde{V}_m(Q_m^i - 1)}{2}.$$

#### APPENDIX D: PROOF OF LEMMA 4

From [26], the convergence property of the asynchronous update and synchronous update is the same. Therefore, we consider the convergence of related synchronous version without loss of generality.

Let  $c \in R$  be a constant, we have  $T_I(c\tilde{V}_S^l) = cT_I(\tilde{V}_S^l)$ , where  $T_I$  is one element of mapping  $\mathbf{T}$  corresponding to the state with all buffers empty. Similar to [27], the per-node value function  $\{\tilde{\mathbf{V}}_m\}$  is bounded almost surely during the iterations of algorithm. According to the construction of parameter vector  $\mathbf{W}$ , the update on  $\tilde{\mathbf{V}}_m$  is equivalent to the update on  $\mathbf{W}$  and proving the convergence of Lemma 4 is equivalent to proving the convergence of update on  $\mathbf{W}$ . In the following, we first introduce and prove the following lemma on the convergence of learning noise.

*Lemma 6:* Define  $\mathbf{q}^l = \mathbf{M}^\dagger \left[ \bar{\mathbf{g}}(\Pi_l) + \mathbf{P}(\Pi_l)\mathbf{M}\mathbf{W}^l - \mathbf{M}\mathbf{W}^l - T_I(\mathbf{M}\mathbf{W}^l)\mathbf{e} \right]$ , when the number of iterations  $l \geq j \rightarrow \infty$ , the procedure of update can be written as follows with probability 1:  $\mathbf{W}^{l+1} = \mathbf{W}^j + \sum_{i=j}^l \epsilon_v^i \mathbf{q}_m^i$ .

The proof of above lemma follows the standard approach of stochastic approximation with Martingale noise [22]. Moreover, the following lemma is about the limit of sequence  $\{\mathbf{q}_m^l\}$ .

*Lemma 7:* Suppose the following two inequalities are true for  $l = a, a + 1, \dots, a + b$

$$\bar{\mathbf{g}}(\Pi^l) + \mathbf{P}(\Pi^l)\mathbf{M}\mathbf{W}^l \leq \bar{\mathbf{g}}(\Pi^{l-1}) + \mathbf{P}(\Pi^{l-1})\mathbf{M}\mathbf{W}^l \quad (40)$$

$$\bar{\mathbf{g}}(\Pi^{l-1}) + \mathbf{P}(\Pi^{l-1})\mathbf{M}\mathbf{W}^{l-1} \leq \bar{\mathbf{g}}(\Pi^l) + \mathbf{P}(\Pi^l)\mathbf{M}\mathbf{W}^{l-1}, \quad (41)$$

then we have

$$|q_i^{a+b}| \leq C_1 \prod_{i=0}^{\lfloor \frac{b}{\beta} \rfloor - 1} (1 - \tau^{a+i\beta}) \quad \forall i, \quad (42)$$

where  $q_i^{a+b}$  denotes the  $i$ th element of the vector  $\mathbf{q}^{a+b}$ ,  $C_1$  is some constant.

*Proof:* From (40) and (41), we have

$$\begin{aligned}\mathbf{q}^l &= \mathbf{M}^\dagger [\bar{\mathbf{g}}(\Pi^l) + \mathbf{P}(\Pi^l)\mathbf{M}\mathbf{W}^l - \mathbf{M}\mathbf{W}^l - w_l\mathbf{e}] \leq \mathbf{M}^\dagger [\bar{\mathbf{g}}(\Pi^{l-1}) + \mathbf{P}(\Pi^{l-1})\mathbf{M}\mathbf{W}^l - \mathbf{M}\mathbf{W}^l - w_l\mathbf{e}] \\ \mathbf{q}^{l-1} &= \mathbf{M}^\dagger [\bar{\mathbf{g}}(\Pi^{l-1}) + \mathbf{P}(\Pi^{l-1})\mathbf{M}\mathbf{W}^{l-1} - \mathbf{M}\mathbf{W}^{l-1} - w_{l-1}\mathbf{e}] \\ &\leq \mathbf{M}^\dagger [\bar{\mathbf{g}}(\Pi^l) + \mathbf{P}(\Pi^l)\mathbf{M}\mathbf{W}^{l-1} - \mathbf{M}\mathbf{W}^{l-1} - w_{l-1}\mathbf{e}]\end{aligned}$$

where  $w_l = T_I(\mathbf{M}\mathbf{W}^l) = T_I(\mathbf{M}\mathbf{W}^l)$ . According to Lemma 6, we have  $\mathbf{W}^l = \mathbf{W}^{l-1} + \epsilon_v^{l-1}\mathbf{q}^{l-1} \Rightarrow \mathbf{W}^l = \mathbf{W}^{l-1} + \epsilon_v^{l-1}\mathbf{q}^{l-1}$ . Therefore,

$$\begin{aligned}\mathbf{q}^l &\leq [(1 - \epsilon_v^{l-1})\mathbf{I} + \mathbf{M}^\dagger\mathbf{P}(\Pi^{l-1})\mathbf{M}\epsilon_v^{l-1}]\mathbf{q}^{l-1} + w_{l-1}\mathbf{e} - w_l\mathbf{e} = \mathbf{B}^{l-1}\mathbf{q}^{l-1} + w_{l-1}\mathbf{e} - w_l\mathbf{e} \\ \mathbf{q}^l &\geq [(1 - \epsilon_v^{l-1})\mathbf{I} + \mathbf{M}^\dagger\mathbf{P}(\Pi^l)\mathbf{M}\epsilon_v^{l-1}]\mathbf{q}^{l-1} + w_{l-1}\mathbf{e} - w_l\mathbf{e} = \mathbf{A}^{l-1}\mathbf{q}^{l-1} + w_{l-1}\mathbf{e} - w_l\mathbf{e}.\end{aligned}$$

Notice that  $\mathbf{A}^{l-1}\mathbf{e} = \mathbf{B}^{l-1}\mathbf{e}$ , we have  $\mathbf{A}^{l-1}\dots\mathbf{A}^{l-\beta}\mathbf{q}^{l-\beta} - C_1\mathbf{e} \leq \mathbf{q}^l \leq \mathbf{B}^{l-1}\dots\mathbf{B}^{l-\beta}\mathbf{q}^{l-\beta} - C_1\mathbf{e}$

$$\begin{aligned}\Rightarrow (1 - \tau^l)[\min \mathbf{q}^{l-\beta}] \leq \mathbf{q}^l + C_1\mathbf{e} \leq (1 - \tau^l)[\max \mathbf{q}^{l-\beta}] &\Rightarrow \begin{cases} \max \mathbf{q}^l + C_1 \leq (1 - \tau^l) \max \mathbf{q}^{l-\beta} \\ \min \mathbf{q}^l + C_1 \geq (1 - \tau^l) \min \mathbf{q}^{l-\beta} \end{cases} \\ \Rightarrow \max \mathbf{q}^l - \min \mathbf{q}^l \leq (1 - \tau^l)[\max \mathbf{q}^{l-\beta} - \min \mathbf{q}^{l-\beta}] &\Rightarrow |q_i^l| \leq \max \mathbf{q}^l - \min \mathbf{q}^l \leq C_2(1 - \tau^l) \quad \forall i,\end{aligned}$$

where the first step is due to conditions of Lemma 4 on matrix sequence  $\{\mathbf{A}^l\}$  and  $\{\mathbf{B}^l\}$ ,  $\max \mathbf{q}^l$  and  $\min \mathbf{q}^l$  denote the maximum and minimum elements in  $\mathbf{q}^l$  respectively,  $C_1$  and  $C_2$  are all constants, the first inequality of the last step is because  $\min \mathbf{q}^l \leq 0$ . This completes the proof of Lemma 7.  $\blacksquare$

Therefore, the proof of Lemma 4 can be divided into the following steps: (1) From the property of sequence  $\{\epsilon_v^l\}$ , we have  $\prod_{i=0}^{\lfloor \frac{l}{\beta} \rfloor - 1} (1 - \epsilon_v^{i\beta}) \rightarrow 0$  ( $l \rightarrow \infty$ ). (2) According to the first step, note that  $\tau^l = \mathcal{O}(\epsilon_v^l)$ , from (42), we have  $\mathbf{q}^l \rightarrow 0$  ( $l \rightarrow \infty$ ). (3) Therefore, the update on  $\{\mathbf{W}^l\}$  will converge, and the fixed point of the convergence  $\mathbf{W}^\infty$  satisfies  $T_I(\mathbf{M}\mathbf{W}^l)\mathbf{e} + \mathbf{W}^\infty = \mathbf{M}^\dagger\mathbf{T}(\mathbf{M}\mathbf{W}^\infty)$ .

#### APPENDIX E: PROOF OF LEMMA 5

Due to the page limit, we only provide the sketch of the proof. The convergence proof of the LMs  $\{\gamma_{S,p}, \gamma_{1,p}, \dots, \gamma_{M,p}\}$  for a given  $\gamma_{S,d}$  is as follows:

- For the notation convenience, we first define the average transmit power of each node as follows:

$$\tilde{\mathcal{P}}_S(\gamma) = \mathbf{E}^\Pi \left[ \sum_{m=1}^M \sum_{i=1}^{\min(N_T, N_R)} \eta_{S,m}^i p_{S,m}^i \right] \quad \text{and} \quad \tilde{\mathcal{P}}_m(\gamma) = \mathbf{E}^\Pi \left[ \sum_{m=1}^M \sum_{i=1}^{\min(N_T, N_R)} \eta_{m,D}^i p_{m,D}^i \right]$$

( $m = 1, 2, \dots, M$ ), where  $\mathbf{E}^\Pi[\cdot]$  denotes the expectation w.r.t. the policy  $\Pi(\gamma)$ . Using standard stochastic approximation theory, the dynamics of the LMs update equation  $\{\gamma_{S,p}, \gamma_{1,p}, \dots, \gamma_{M,p}\}$  can be represented by the following ODE:

$$[\dot{\gamma}_{S,p}(t), \dots, \dot{\gamma}_{M,p}(t)]^T = [\tilde{\mathcal{P}}_S(\gamma) - P_S, \tilde{\mathcal{P}}_1(\gamma) - P_R, \dots, \tilde{\mathcal{P}}_M(\gamma) - P_R]^T. \quad (43)$$

- Using perturbation analysis in [28], we have  $\frac{\partial \tilde{\mathcal{P}}_m(\gamma)}{\partial \gamma_{m,p}} < 0$  ( $m = S, 1, 2, \dots, M$ ) and  $\left| \frac{\partial \tilde{\mathcal{P}}_m(\gamma)}{\partial \gamma_{m,p}} \right| \gg \left| \frac{\partial \tilde{\mathcal{P}}_m(\gamma)}{\partial \gamma_{n,p}} \right|$  ( $m = S, 1, 2, \dots, M, n \neq m$ ). Thus, the update of  $\gamma_{m,p}$  ( $m = S, 1, \dots, M$ ) in ODE (43) will drive  $\tilde{\mathcal{P}}_m - P_R$  (or  $\tilde{\mathcal{P}}_S - P_S$ ) to 0 whenever  $\tilde{\mathcal{P}}_m - P_R$  (or  $\tilde{\mathcal{P}}_S - P_S$ ) is non-zero. Therefore, the ODE (43) will converge. The converged LMs  $\{\gamma_{S,p}^*(\gamma_{S,d}), \gamma_{1,p}^*(\gamma_{S,d}), \dots, \gamma_{M,p}^*(\gamma_{S,d})\}$  can be characterized by the equilibrium point of the ODE (43), which is given by the RHS of (43)  $\rightarrow 0$ .

Suppose for a given  $\gamma_{S,d}$ ,  $\{\gamma_{S,p}, \gamma_{1,p}, \dots, \gamma_{M,p}\}$  converge to  $\{\gamma_{S,p}^*(\gamma_{S,d}), \gamma_{1,p}^*(\gamma_{S,d}), \dots, \gamma_{M,p}^*(\gamma_{S,d})\}$ . Since  $\frac{\partial (\mathbf{E}_{\gamma_{1,p}^*, \dots, \gamma_{M,p}^*, \gamma_{S,p}^*}^\Pi [Q_{S=N_Q}])}{\partial \gamma_{S,d}} < 0$ , the update on  $\gamma_{S,d}$  will converge as well for a similar reason as in the convergence of  $\{\gamma_{S,p}, \gamma_{1,p}, \dots, \gamma_{M,p}\}$ . Similarly, the converged  $\gamma_{S,d}^*$  can be characterized by the equilibrium point of the ODE  $\dot{\gamma}_{S,d}(t) = \mathbf{E}_{\gamma_{1,p}^*, \dots, \gamma_{M,p}^*, \gamma_{S,p}^*}^\Pi [Q_{S=N_Q}] - D$ , which is given by the RHS  $\rightarrow 0$ .

#### APPENDIX F: PROOF OF THEOREM 2

Without loss of generality, we shall consider the approximate value function  $V(\mathbf{Q}) = \sum_{m=S}^M \sum_{q=1}^{N_Q} \tilde{V}_m(q) \mathbf{I}[Q_m = q]$  on the following redefined set of representative states  $\mathcal{Q}_R = \{\delta_{m,q} | m = S, 1, 2, \dots, M; q = 0, 1, \dots, q_I - 1, q_I + 1, \dots, N_Q\}$ , where the state  $\delta_{m,q}$  is given by  $\delta_{m,q} = [Q_S = q_I, Q_1 = q_I, \dots, Q_m = q, \dots, Q_M = q_I]^T$  and  $q_I < N_Q$  is sufficiently large. Correspondingly,  $\mathbf{M}^{-1}$  should also be redefined such that the per-node value function  $\{\tilde{\mathbf{V}}_m\}$  is updated on the representative states  $\mathcal{Q}_R$  [21].

First of all, following the similar approach in the proof of Lemma 4, the per-node value function (under the new reference states) would also converge almost surely to  $\{\tilde{\mathbf{V}}_m^\infty(\gamma)\}$  for any given LMs  $\gamma$ .

Next, when the conditions of Theorem 2 are satisfied, given any  $\epsilon > 0$ , there is one integer  $Q_0(\epsilon)$  such that for all  $q > Q_0(\epsilon)$  and  $q_I = Q_0(\epsilon)$ , we have (from the proof of Lemma 3):

$$\tilde{V}_m^\infty(q-r) - \tilde{V}_m^\infty(q) = \tilde{V}_m^\infty(q_I-r) - \tilde{V}_m^\infty(q_I) + \mathcal{O}(\epsilon). \quad (44)$$

Moreover, since  $\{\tilde{V}_m^\infty(q)\}$  are all monotonically increasing functions with respect to  $q$  and  $\{\tilde{V}_m^\infty(N_Q)\}$

are all bounded<sup>16</sup>, we have  $\tilde{V}_m(Q_0(\epsilon)) = \mathcal{O}(\epsilon)$  for sufficiently large arrivals. Therefore, (44) holds for all  $q \in [0, N_Q]$  for sufficiently large  $N_Q$  and input arrivals. Similarly, we have

$$\tilde{V}_S^\infty(q+n-r) - \tilde{V}_S^\infty(q+n) = \tilde{V}_S^\infty(q_I+n-r) - \tilde{V}_S^\infty(q_I+n) + \mathcal{O}(\epsilon) \quad (45)$$

$$\tilde{V}_m^\infty(q+r) - \tilde{V}_m^\infty(q) = \tilde{V}_m^\infty(q_I+r) - \tilde{V}_m^\infty(q_I) + \mathcal{O}(\epsilon). \quad (46)$$

Hence, with the above equations and substituting the converged per-node value function  $\{\tilde{\mathbf{V}}_m^\infty(\gamma)\}$  into (18) for the reference states, we get

$$\begin{aligned} \tilde{V}_S^\infty(q) = & q + \gamma_{S,d} \mathbf{I}[q = N_Q] + \sum_n f_X(n) \left( \tilde{V}_S^\infty(q+n) - \tilde{V}_S^\infty(n) \right) + \min_{\Pi} \mathbf{E}_{\mathbf{H}} \left\{ \sum_{m, N_{SR}} \eta_{S,m}^{N_{SR}} \left[ \gamma_{S,p} \sum_k p_{S,m}^{N_{SR}} \right. \right. \\ & \left. \left. + \sum_n f_X(n) \left( \tilde{V}_S^\infty(q+n-r_{S,m}^{N_{SR}}) - \tilde{V}_S^\infty(q+n) \right) + \tilde{V}_m^\infty(q_I+r_{S,m}^{N_{SR}}) - \tilde{V}_m^\infty(q_I) \right] \right\} \end{aligned} \quad (47)$$

$$\tilde{V}_m^\infty(q) = q + \tilde{V}_m^\infty(q) + \min_u \mathbf{E}_{\mathbf{H}} \left\{ \sum_{N_{RD}} \eta_{m,D}^{N_{RD}} \left[ \gamma_{m,p} \sum_k p_{m,D}^{N_{RD}} + \tilde{V}_m^\infty(q-r_{m,D}^{N_{RD}}) - \tilde{V}_m^\infty(q) \right] \right\}, \quad (48)$$

where  $m = 1, 2, \dots, M$ .

Finally, for any system state  $\mathbf{Q}^i = [Q_S^i, \dots, Q_M^i]^T$ , substitute the above equations into the RHS of the original Bellman equation in (18), we get RHS of (18)  $\stackrel{\text{a}}{=} \sum_{m=S}^M Q_m^i + \gamma_{S,d} \mathbf{I}[Q_S^i = N_Q] + \sum_n f_X(n) \tilde{V}_S^\infty(Q_S^i + n) + \sum_{m=1}^M \tilde{V}_m^\infty(Q_m^i) + \min_{\Pi(\mathbf{Q}^i)} \mathbf{E}_{\mathbf{H}} \left\{ \sum_{m, N_{SR}} \eta_{S,m}^{N_{SR}} \left[ \gamma_{S,p} p_{S,m}^{N_{SR}} + \sum_n f_X(n) \left( \tilde{V}_S^\infty(Q_S^i + n - r_{S,m}^{N_{SR}}) - \tilde{V}_S^\infty(Q_S^i + n) \right) + \tilde{V}_m^\infty(q_I + r_{S,m}^{N_{SR}}) - \tilde{V}_m^\infty(q_I) \right] + \sum_{m, N_{RD}} \eta_{m,D}^{N_{RD}} \left[ \gamma_{m,p} p_{m,D}^{N_{RD}} + \tilde{V}_m^\infty(Q_m^i - r_{m,D}^{N_{RD}}) - \tilde{V}_m^\infty(Q_m^i) \right] \right\} + \mathcal{O}(\epsilon) \stackrel{\text{b}}{=} \sum_{m=S}^M \tilde{V}_m^\infty(Q_m^i) + \sum_n f_I(n) \tilde{V}_S^\infty(n) + \mathcal{O}(\epsilon) = V(\mathbf{Q}^i) + \sum_n f_I(n) \tilde{V}_S^\infty(n) + \mathcal{O}(\epsilon)$ , where equality (a) is due to (46), equality (b) is due to (47) and (48). Since  $\sum_n f_X(n) \tilde{V}_S^\infty(n)$  is a constant independent of  $\mathbf{Q}^i$  and  $\epsilon$  is chosen arbitrarily, we have shown that the approximate value function  $V(\mathbf{Q}) = \sum_{m=S}^M \sum_{q=1}^{N_Q} \tilde{V}_m^\infty(q) \mathbf{I}[Q_m = q]$  can satisfy the original Bellman equation (18) asymptotically (when  $N_Q \rightarrow +\infty$ ). As a result, the proposed distributive update algorithm converges to the global optimal solution and this completes the proof.

## REFERENCES

- [1] E. van der Meulen, "Transmission of information in a t-terminal discrete memoryless channel," Ph.D. dissertation, Dep. of Statistics, University of California, Berkeley, 1968.

<sup>16</sup> $\tilde{V}_m^\infty(q)$  measures the contribution of cost  $q + \gamma_{m,p} \sum_k p_{m,D,k}$  if the system starts at  $Q = q$ . For finite  $N_Q$ ,  $\tilde{V}_m^\infty(q)$  is always bounded. On the other hand, since the system is stable, the queue length is bounded with probability 1 for arbitrarily large  $N_Q$  and hence,  $\tilde{V}_m^\infty(N_Q)$  must be bounded almost surely for arbitrarily large  $N_Q$ .

- 
- [2] T. Cover and A. Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, Sep 1979.
- [3] IEEE 802.16's Relay Task Group. [Online]. Available: <http://www.ieee802.org/16/relay/index.html>.
- [4] WINNER- Wireless World Initiative New Radio. [Online]. Available: <http://www.ist-winner.org/>.
- [5] P. Yip and D. Etter, "An adaptive technique for multiple echo cancelation in telephone networks," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, vol. 12, Apr 1987, pp. 2133–2136.
- [6] L. Vega, H. Rey, J. Benesty, and S. Tressens, "A new robust variable step-size nlms algorithm," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1878–1893, May 2008.
- [7] E. Lo and K. Letaief, "Optimizing downlink throughput with user cooperation and scheduling in adaptive cellular networks," in *Wireless Communications and Networking Conference, 2007.WCNC 2007. IEEE*, March 2007, pp. 4342–4347.
- [8] D. P. Bertsekas, *Dynamic Programming - Deterministic and Stochastic Models*. Prentice Hall, NJ, USA, 1987.
- [9] X. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. Springer, 2008.
- [10] E. Yeh and R. Berry, "Throughput optimal control of cooperative relay networks," in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, Sept. 2005, pp. 1206–1210.
- [11] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.
- [12] J. Abounadi, D. Bertsekas, and V. S. Borkar, "Learning algorithms for markov decision processes with average cost," *SIAM Journal on Control and Optimization*, vol. 40, pp. 681–698, 1998.
- [13] T. Richardson and R. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 599–618, Feb 2001.
- [14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [15] Z. Han, Z. Ji, and K. Liu, "Non-cooperative resource competition game by virtual referee in multi-cell ofdma networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 6, pp. 1079–1090, August 2007.
- [16] D. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," *IEEE Transactions on Automatic Control*, vol. 52, no. 12, pp. 2254–2269, Dec. 2007.
- [17] J. Huang, R. Berry, and M. L. Honig, "Auction-based spectrum sharing," *ACM Mobile Networks and Applications Journal (MONET)*, vol. 11, pp. 405–418, June 2006.
- [18] C. Curescu and S. Nadjm-Tehrani, "A bidding algorithm for optimized utility-based resource allocation in ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 12, pp. 1397–1414, Dec. 2008.
- [19] D. Bertsekas, *Dynamic Programming and Optimal Control, Vol. 2*. Athena Scientific, 2007.
- [20] V.S.Borkar, "An actor-critic algorithm for constrained markov decision processes," in *Systems Control Lett. 54*, 2005, pp. 207–213.
- [21] J. N. Tsitsiklis and B. van Roy, "Feature-based methods for large scale dynamic programming," *Machine Learning*, vol. 22, pp. 59–94, March 1996.
- [22] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.
- [23] —, "Stochastic approximation with two time scales," *Systems Control Lett. 29*, pp. 291–294, 1997.

- 
- [24] L. Georgiadis, M. Neely, and L. Tassiulas, *Resource Allocation and Cross Layer Control in Wireless Networks*. Now Publishers Inc, 2006.
- [25] I. Bettesh and S. Shamai, “Optimal power and rate control for minimal average delay: The single-user case,” *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4115–4141, Sept. 2006.
- [26] V. S. Borkar, “Asynchronous stochastic approximation,” *SIAM J. Control and Optim.*, vol. 36, pp. 840–851, 1998.
- [27] V. S. Borkar and S. P. Meyn, “The ode method for convergence of stochastic approximation and reinforcement learning algorithms,” *SIAM J. on Control and Optimization* 38, pp. 447–469, 2000.
- [28] X. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach*. Springer, 2007.