A Multi-Agent View of Wireless Video Streaming with Delayed Client-Feedback

Nouman Khan*, Ujwal Dinesha†, Subrahmanyam Arunachalam†, Dheeraj Narasimha†,
Vijay Subramanian*, and Srinivas Shakkottai†
*University of Michigan, Ann Arbor, MI 48109–2122
†Texas A&M University, College Station, TX 77843-3127
{knouman, vgsubram}@umich.edu, {ujwald36, subrahmanyam_a, sshakkot}@tamu.edu, {dheeraj.narasimha}@gmail.com

***Abstract*—We study the optimal control of multiple video streams over a wireless downlink from a base-transceiver-station (BTS)/access point to N end-devices (EDs). The BTS sends video packets to each ED under a joint transmission energy constraint, the EDs choose when to play out the received packets, and the collective goal is to provide a high Quality-of-Experience (QoE) to the clients/end-users. All EDs send feedback about their states and actions to the BTS which reaches it after a fixed deterministic delay. We analyze this team problem with delayed feedback as a cooperative Multi-Agent Constrained Partially Observable Markov Decision Process (MA-C-POMDP).**

First, using a recently established strong duality result for MA-C-POMDPs, the original problem is decomposed into N independent unconstrained transmitter-receiver (two-agent) problems—all sharing a Lagrange multiplier (that also needs to be optimized for optimal control). Thereafter, the common information (CI) approach and the formalism of approximate information states (AISs) are used to guide the design of a neural-network based architecture for learning-based multi-agent control in a single unconstrained transmitter-receiver problem. Finally, simulations on a single transmitter-receiver pair with a stylized QoE model are performed to highlight the advantage of delay-aware two-agent coordination over the transmitter choosing both transmission and play-out actions (perceiving the delayed state of the receiver as its current state).

I. Introduction

The emergence of Open Radio Access Networks (O-RANs) is disrupting the traditional cellular RAN by disaggregating the components and making them accessible to data collection and control methods that are based on open standards. Concurrently, RAN intelligent control (RIC) is being developed for the deployment of measurement and control policies at multiple timescales over O-RAN. Not only can RIC utilize RAN-level information—channel conditions and data backlog associated with each connected device—, but it can also use application-level state and performance information shared by devices to optimize resource allocation. Thus, applications such as media streaming, mixed reality, or robot control can be enhanced via machine learning (ML) at the RIC with tighter integration of RAN-level decision making and application-level performance metrics.

While RIC promises new paradigms for supporting ML-driven policies with application inputs, two questions arise on its feasibility: (*i*) **Scalability**: As the number of connected devices at any given time might be in the thousands, and the system state is high-dimensional—spanning wireless and application information—*will the control policies be too complex?* (*ii*) **Partial Observability**: *How much will partial observability of the application-level states due to delayed (or sporadically available) information degrade system performance?*

We address the twin issues of scalability and partial-observability for application-aware RIC, focusing on video streaming as a specific use-case. Here, the RIC located at the base-transceiver-station (BTS) must decide on forwarding video packets to the connected end-devices (EDs) under a joint transmission energy constraint across all the EDs, and the video players at the EDs must decide how best to play out the buffered packets to ensure high quality of user experience (QoE). Each ED provides delayed feedback on its application states and decisions (via the uplink), and so the RIC is only partially aware of the underlying system state. Our goal is to design a simple and scalable approach for optimal control and learning by factorizing the above complex multi-agent constrained decision making problem into simpler sub-problems.

A. System Overview and Main Results

We consider a cellular downlink with N EDs that are connected to the Internet via a BTS. Each ED streams a video from a remote server which is assumed to reach the BTS in a steady manner in the form of chunks. These chunks are buffered up at the BTS and then await transmission to the appropriate ED. The BTS has a transmission constraint and may provide high quality of service to a subset of the EDs only. Each ED maintains a playback buffer and in each time-step, to maximize a QoE function, EDs with non-empty playback buffers must decide whether to play out a chunk or wait for more chunks to be buffered up. Importantly, we assume that the BTS and all EDs are interested in maximizing the long-term discounted sum utility of all the video streams (over a finite time-horizon, if there is no discounting).

Our contributions are: (*i*) a structural decomposition that allows us to simplify a complex multi-agent constrained decision making problem without loss of optimality; and (*ii*) a scalable learning-based algorithm that uses this structural decomposition. We now discuss key aspects of our results.

1) Lagrangian Decomposition: The system state is defined as a pair consisting of the states of the playback buffers (one

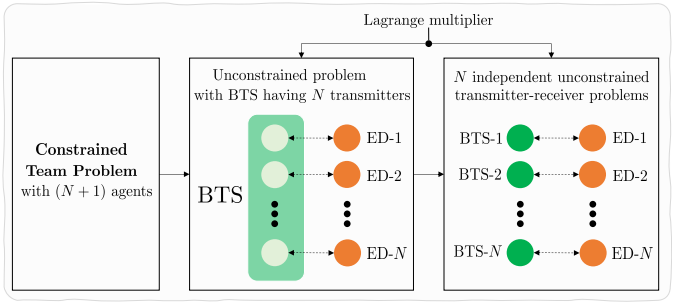

Fig. 1. The first block has the initial $(N+1)$-agent constrained problem. The second block has an unconstrained problem as a result of Lagrangian decomposition, where the BTS has N transmitters, each having the delayed feedback from all the receivers. The third block has N independent unconstrained transmitter-receiver problems, all sharing the Lagrange multiplier.

for each ED) and QoE tracking variables (one for each ED). The transmission constraint at the BTS applies jointly across all the EDs. The BTS has delayed information on states and actions of all EDs, and each ED has access to its own local information as well as the transmission actions of the BTS for it. This setting becomes an $(N+1)$-agent constrained POMDP problem. First, we use a Lagrangian decomposition (justified by a recently established strong duality result) to decompose the problem into an unconstrained one wherein the BTS has N *transmitters* inside, each sending video packets to a specific ED, now synonymous with a *receiver*. Each of the N transmitters has the (delayed) feedback of all the receivers.

2) Factorization of Transmission Policy: Having N transmitters at the BTS that share (and use) all the fed-back information is not scalable. Ideally, we desire scalable microservices, each handling a particular ED with only limited information. Our next result states that after decomposition, we may factorize the problem such that each transmitter can ignore the feedback from all receivers except the one it transmits to. We thus obtain N independent unconstrained transmitter-receiver problems, all sharing the Lagrange multiplier. We prove that this simplification does not sacrifice optimality.

Fig. 1 shows the steps of the structural simplifications.

3) Microservice-scale Learning-based Control with Approximate Information States: Each one of the N (independent and unconstrained) transmitter-receiver problems is still challenging. On one hand, the transmission-policy should be selected based on the past of the play-out policy, because it influences the conditional distribution of the unknown states and actions of the receiver. On the other hand, by the nature of dynamic programming, the optimal choice of play-out actions at a given time, depends on future costs, which are determined by the future choices made by the transmission policy. The common-information (CI) approach [1] breaks such dependency cycles by decomposing a given multi-agent POMDP problem into an equivalent single-agent POMDP problem.[1] However, in the learning context, as the wireless channel's transition-law is unknown, we cannot use the belief-based information state (IS) to be able to use an MDP-based Reinforcement Learning (RL) algorithm.[2] Hence, we use the notion of approximate information states (AISs) recently introduced for the multi-agent team setting [2], [3]. These notions help us design a multi-agent RL (MARL) approach wherein the AISs and the corresponding control policies can be learned concurrently.

4) Simulations: We empirically evaluate the performance of our proposed MARL approach on a single unconstrained transmitter-receiver problem with a stylized QoE model. Specifically, we compare the performance of the approach with a delay-oblivious BTS choosing both the transmission and play-out actions. Our simulation results show that the delay-aware MARL approach improves the system performance significantly compared to the (single-agent) delay-oblivious one. We, however, leave the implementation of a complete primal-dual setup and its extensive performance evaluation for future work.

B. Related Work

The growing popularity of video streaming applications, especially over wireless, has led to a concerted attempt to address the various challenges that arise. In particular, many works aim to improve QoE metrics in wireless environments: [4], [5] analyzed the flow level dynamics to study buffer starvation and the frequency of interruptions; [6]–[8] used the network utility maximization paradigm, and then provided factorized primal-dual congestion control-type algorithms; with no feedback delay, full information sharing, and model knowledge, together which yield an MDP, [9] proved that factorized policies are optimal for the loosely coupled constrained discounted cost MDP and developed a primal-dual algorithm, and in the same setting, [10] developed similar results for the average cost problem, again using LP duality; in same MDP setting [10], [11] used index-based policies for finite-time horizon, discounted cost, and average cost settings, all with hard constraints; and fixing the policy of one of the two agents, RL algorithms have also been proposed [12]–[14] for solving the (unknown) MDP that results. For a non-exhaustive survey, see [15].

MDPs require full observability of state—see [9]–[11] for video streaming—, but realistic mechanisms (such as delayed ED-feedback in video streaming) result in partial observability of state, so POMDPs [16], [17] apply. With the model known, principled solution approaches exist like [18]–[20], but POMDPs are known to be PSPACE complete [21]. Use of the belief-based information state (IS) [22], [23] is not robust, and is also not implementable in the learning context (model unknown case). Various other ways to abstract the large state spaces that go hand in hand with the POMDP problem have been suggested [24]. Recent methods [25], [26] using formal notions of approximate information state (AIS) discovered from data are currently most promising for this.

Besides partial observability due to the delayed ED-feedback, the video streaming problem is also a multi-agent

[1]This decomposition holds for both unconstrained and constrained settings.

[2]The update of belief-based IS requires model knowledge.

problem (with information asymmetry) when the BTS and the EDs are viewed as a single team. Viewed as such, the problem is an instance of a cooperative multi-agent constrained POMDP (MA-C-POMDP) [3], [27], wherein enabling coordination between agents is hard. This perspective has not been studied for video streaming: related work [9]–[11] has full and instantaneous information sharing. The common information (CI) approach [1], [28] is able to circumvent this difficulty (conceptually, by formulating a POMDP) when the model is known. In the learning context, we can employ the AIS framework for learning in multi-agent POMDPs [2], [3].

C. Notation

The key notations in this paper are as follows:

- Probability and expectation operators are denoted by $\mathbb{P}(\cdot)$ and $\mathbb{E}[\cdot]$ respectively. Random variables are denoted by upper-case letters and their realizations by the corresponding lower-case letters. We also use shorthand, $\mathbb{E}[\cdot|x] \triangleq \mathbb{E}[\cdot|X=x]$ and $\mathbb{P}(y|x) \triangleq \mathbb{P}(Y=y|X=x)$, for conditional quantities.
- For every conditional probability or conditional expectation, it is implicitly assumed that the conditioning event has positive probability.
- The sets of integers and positive integers are respectively denoted by $\mathbb{Z}$ and $\mathbb{N}$. For integers a and b, $[a,b]_{\mathbb{Z}}$ represents the set $\{a, a+1, \dots, b\}$ if $a \leqslant b$ and $\varnothing$ otherwise. The notations [a] and $[a,\infty]_{\mathbb{Z}}$ are used as shorthand for $[1,a]_{\mathbb{Z}}$ and $\{a, a+1, \dots\}$, respectively.
- For integers $a \leqslant b$ and $c \leqslant d$, and a quantity of interest q, $q^{a:b}$ and $q_{c:d}$ are shorthand respectively for vectors $\left(q^a, q^{a+1}, \dots, q^b\right)$ and $(q_c, q_{c+1}, \dots, q_d)$. The notation $q_{a:b}^{c:d}$ is shorthand for the vector $(q_i^j : i \in [a,b]_{\mathbb{Z}}, j \in [c,d]_{\mathbb{Z}})$. Infinite tuples $\left(q^a, q^{a+1}, \dots,\right)$ and $(q_c, q_{c+1}, \dots,)$ are respectively denoted by $q^{a:\infty}$ and $q_{c:\infty}$.
- A list of important symbols is given in Appendix A.

D. Organization

The rest of the paper is organized as follows. The multi-agent constrained video streaming (MA-C-VS) problem is formulated in Section II. Structural decomposition results leading to N independent unconstrained transmitter-receiver problems are laid out in Section III. For a single transmitter-receiver problem, the CI approach and AISs are used in Section IV to guide the design of a neural-network based architecture for learning-based multi-agent control. Simulation results on a single unconstrained transmitter-receiver pair with a stylized QoE model are presented in Section V. Finally, in Section VI, concluding remarks are presented.

II. System Model and Problem Formulation

As described earlier, the BTS transmits packets of different video files on a shared downlink to N EDs, which use their playback buffers to store and play them out. While the BTS is able to maintain steady-streams of video packets from the remote servers (and hence, always has packets to transmit to each ED), it has a fixed energy budget that applies jointly to all downlink transmissions. Fig. 2 illustrates our system model.

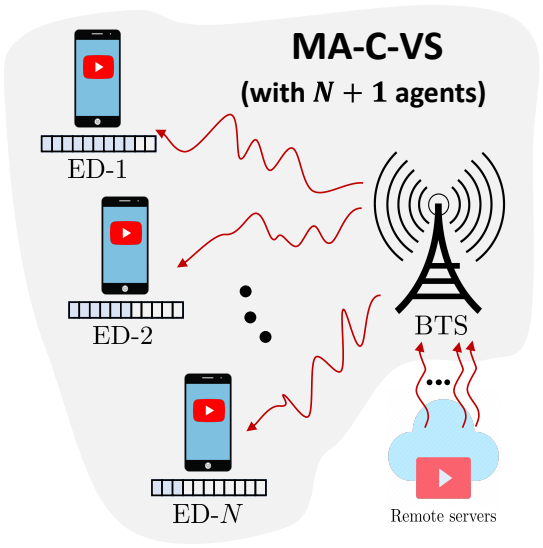

Fig. 2. Multi-Agent Constrained Video Streaming (MA-C-VS) problem with N end-devices (EDs) and 1 base-transceiver-station (BTS).

A. System Model Components

We consider a discrete-time system with $t = 1, 2, \dots$ denoting a single time-step. We now define the different elements of the system model.

1) State Space: We denote the state of ED-n by X_t^n which is assumed to have two components, namely $X_t^{1,n}$ and $X_t^{2,n}$: $X_t^{1,n}$ represents the state of the ED's playback buffer (e.g., the length of the playback buffer, the resolution of the packet at its head etc.), whereas $X_t^{2,n}$ represents a QoE tracking variable for the ED (e.g., the number of stalls experienced by the ED so far, whether the ED is already in stall and the stall's duration etc.). We assume that each X_t^n takes values from a finite set $\mathcal{X}$. Finally, the state of the system at time t is given by $X_t \triangleq X_t^{1:N}$ which is controlled jointly by the actions of the BTS and the EDs, as described next.

2) EDs' Action Spaces: At time t, each ED-n decides a *play-out action* from a finite set of possible actions denoted by $\mathcal{V}$ (e.g., whether to play the packet at the head of its playback buffer). We denote this decision of ED-n by V_t^n, and the play-out decisions of all EDs by $V_t \triangleq V_t^{1:N}$.

3) BTS's Action Space: At time t, when sending a packet to ED-n, the BTS chooses a *transmission action* from a finite set $\mathcal{U}$. For example, a transmission action could include service class (e.g., a high service class being associated with a high probability of successful transmission), the resolution of the packet being sent, etc. We denote this decision for ED-n by $U_t^n \in \mathcal{U}$ and assume that the received packet arrives after the ED's action V_t^n has already been taken. The collective transmission decision of the BTS is denoted by $U_t \triangleq U_t^{1:N}$.

4) EDs' Feedback: We assume that the BTS's transmission action for ED-n at time t, U_t^n, is available to the ED at the start of the time-step $t+1$ (for example, through a separate control channel). On the other hand, each ED's feedback on their state and play-out action is assumed to have a deterministic delay

of $d \in \mathbb{Z}_{\geqslant 0} \triangleq \{0, 1, \dots\}$ time-steps before it reaches the BTS. Importantly, the feedback by ED-n is only known to it and the BTS, and not to the rest of the EDs. Viewed collectively, the $N+1$ controllers have no common information. This inherent information asymmetry where each ED is unaware of the states and actions of other EDs, is an important feature of the problem. The BTS, on the other hand, knows and (uses in its decisions) the (delayed) information it receives from all the EDs. We also note that, whenever $d > 0$, then for $t \in \{0, -1, -2, \dots, 1-d\}$, we assume $X_t = X_1$ and $V_t^n = 0 = U_t^n$ for all $n \in [N]$. Fig. 3 shows the timeline of events in each time-step.

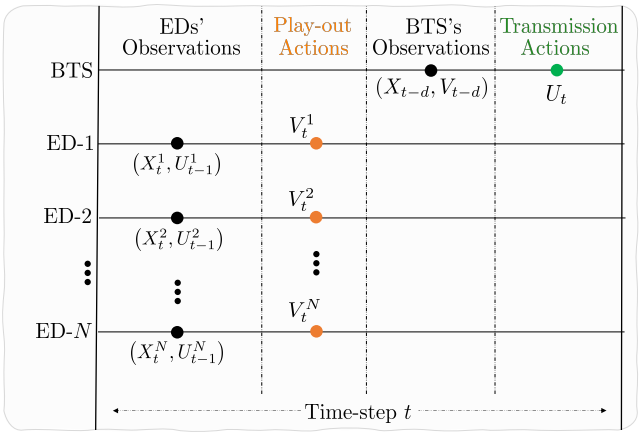

Fig. 3. Timeline of events in each time-step.

5) Transition-Law and Initial Distribution: Given the current system state and actions of the BTS and the EDs, the next state is assumed to be generated via a transition law,

$$\mathcal{P}_{tr} = \{P^n(x^n, u^n, v^n, \tilde{x}^n) : n \in [N]\}, \tag{1}$$

in a time-homogeneous and factorized manner as follows:

$$\begin{aligned}&\mathbb{P}\left(X_{t+1} = x_{t+1} \middle| X_t = x_t, U_t = u_t, V_t = v_t\right)\\ &= \prod_{n=1}^{N} \mathbb{P}\left(x_{t+1}^n \middle| x_t^n, u_t^n, v_t^n\right) = \prod_{n=1}^{N} P^n\left(x_t^n, u_t^n, v_t^n, x_{t+1}^n\right). \end{aligned} \tag{2}$$

Furthermore, we assume that the initial system state X_1 is fixed at $x_1 \triangleq x_1^{1:N}$.

B. Transmission and Play-out Policies

We denote the information available to the BTS right before it chooses its transmission actions at time t by H_t^{BTS}, i.e.,

$$H_t^{BTS} \triangleq (X_{1:t-d}, V_{1:t-d}, U_{1:t-1}). \tag{3}$$

On the other hand, the information available to ED-n before it chooses its play-out action at time t is denoted by H_t^n, i.e.,

$$H_t^n \triangleq \left(X_{1:t}^n, V_{1:t-1}^n, U_{1:t-1}^n\right). \tag{4}$$

Importantly, we note that

$$H_t^{BTS} = \oplus_{n=1}^{N} \tilde{H}_t^n, \text{ for } \tilde{H}_t^n = (X_{1:t-d}^n, V_{1:t-d}^n, U_{1:t-1}^n). \tag{5}$$

With (3), the *transmission policy* of the BTS can be described by a collection of functions $f = f_{1:\infty}$, where each f_t returns a probability distribution on $\mathcal{U}^N$ for a given realization of H_t^{BTS}. The returned distribution is used by the BTS to draw its collective transmission decision, i.e., $U_t \sim f_t(H_t^{BTS})$. Similarly, with (4), the *play-out policy* of ED-n can be described by a collection of functions $g^n = g_{1:\infty}^n$ where each g_t^n returns a probability distribution on $\mathcal{V}$ for a given realization of H_t^n. The returned distribution is used by ED-n to draw its play-out action, i.e., $V_t^n \sim g_t^n(H_t^n)$. The pair, (f, g) (here $g = g^{1:N}$) is called the *policy-profile* of the team.

We denote the set of all behavioral transmission policies of the BTS by $\mathcal{F}$, the set of all behavioral play-out policies of ED-n by $\mathcal{G}^n$, and let $\mathcal{G} = \prod_{n=1}^{N} \mathcal{G}^n$ denote the set of all EDs' play-out policies (g denoting a typical element of $\mathcal{G}$).

C. End-Device and Team Cost Functions, and Team Problem

1) Immediate Cost Functions: We assume that at time t, the team achieves a QoE metric that can be associated with an immediate objective cost $c(X_t, V_t)$ of an additively separable form, i.e.,

$$c(X_t, V_t) = \sum_{n=1}^{N} c^n(X_t^n, V_t^n). \tag{6}$$

Similarly, we assume that the total energy spent by the BTS at time t is given by

$$e(U_t) = \sum_{n=1}^{N} e^n(U_t^n). \tag{7}$$

2) Long-Term Costs: For a given policy-profile (f, g) and a discount factor $\alpha \in (0, 1)$, with the initial system state fixed at x_1, we can define the expected infinite-horizon discounted[3] objective and constraint costs respectively as follows:

$$C, E : \mathcal{F} \times \mathcal{G} \to \mathbb{R},$$

$$C(f, g) \triangleq \mathbb{E}_{x_1}^{(f,g)}\left[\sum_{t=1}^{\infty} \alpha^{t-1} c(X_t, V_t)\right], \tag{8}$$

$$E(f, g) \triangleq \mathbb{E}_{x_1}^{(f,g)}\left[\sum_{t=1}^{\infty} \alpha^{t-1} e(U_t)\right]. \tag{9}$$

As in [27], one can give $\mathcal{F} \times \mathcal{G}$ a suitable topology in which it is compact and then define a compact set of probability measures on it, here denoted by $M_1(\mathcal{F} \times \mathcal{G})$. This allows the team to work with mixtures of policy-profiles, i.e., before interacting with the environment, say at time 0, the team uses a measure μ from $M_1(\mathcal{F} \times \mathcal{G})$ to draw its policy-profile and then proceeds with it from time 1 onward. We can then extend the definitions of C and E as follows.

$$\widehat{C}, \widehat{E} : M_1(\mathcal{F} \times \mathcal{G}) \to \mathbb{R},$$

$$\widehat{C}(\mu) \triangleq \mathbb{E}^{((f,g)\sim\mu)}[C(f, g)], \tag{10}$$

$$\widehat{E}(\mu) \triangleq \mathbb{E}^{((f,g)\sim\mu)}[E(f, g)]. \tag{11}$$

[3]Due to technical challenges, the case of average costs is for future work.

3) Team Problem: Denoting the long-term discounted energy budget of the BTS by $K \in \mathbb{R}_{>0}$, we have the following multi-agent constrained video streaming (MA-C-VS) problem.

$$\text{minimize } \widehat{C}(\mu) \tag{12}$$

$$\text{s.t. } \mu \in M_1(\mathcal{F} \times \mathcal{G}) \text{ and } \widehat{E}(\mu) \leqslant K. \tag{13}$$

(*MA-C-VS*)

Note that (*MA-C-VS*) is feasible (since the BTS can choose a no-transmission policy). We denote its optimal value by $\underline{C}$.

Remark 1. *By Kuhn's theorem [29], for a single-agent, a behavioral policy is equivalent to some mixture of its pure policies with mixing done at time 0 (before interaction with the environment). Thus, the team jointly mixing over the space $M_1(\mathcal{F} \times \mathcal{G})$ at time 0 is equivalent to the team jointly mixing over pure policy-profiles at time 0. This, in turn, via the coordinator's viewpoint of the common-information approach, is equivalent to the coordinator[4] adopting a behavioral policy, necessitating common randomness across all agents. In Section IV-A, we shall see that the role of coordinator, in a single transmitter-receiver problem is taken by the BTS. Thus, the BTS can generate the random prescriptions, and then send them to each ED over the control channel.*

III. Structural Decomposition Results

A. MA-C-POMDP View, Lagrangian Relaxation, Strong Duality, and Existence of Saddle-Point

Problem (*MA-C-VS*) has the following features: each ED has a state that it fully observes; no ED observes either the state or action of any other ED; the BTS has no state of its own and observes the state and action of each ED after a fixed deterministic delay; the BTS's transmission action for a specific ED is observed by that ED (after the ED has taken its own action); the BTS needs to respect a constraint—see (13)—on the expected infinite-horizon discounted energy consumption; and all of the agents wish to minimize the expected infinite-horizon discounted objective cost—see (12). Note that there is no information that is common to all agents—EDs and BTS together. Based on this, the optimization problem (*MA-C-VS*) is an instance of the (cooperative) MA-C-POMDP problem in [27]. Importantly, (*MA-C-VS*) satisfies Slater's condition (existence of a strictly feasible policy-profile mixture).

As discussed in [27], Lagrangian relaxation can be used to solve (*MA-C-VS*). The Lagrangian function $\widehat{L} : M_1(\mathcal{F} \times \mathcal{G}) \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ is given by:

$$\widehat{L}(\mu, \lambda) = \mathbb{E}^{((f,g)\sim\mu)}\left[L((f,g),\lambda)\right], \text{ where}$$

$$L((f,g),\lambda) = \mathbb{E}_{x_1}^{(f,g)}\left[C(f,g) + \lambda(E(f,g) - K)\right]. \tag{14}$$

The following proposition, which is a restatement of [27][Theorem 1] for our simpler context, establishes that the solution to (*MA-C-VS*) can be obtained from solving the unconstrained optimization problem corresponding to (14).

[4]For details, see Section IV-A.

Proposition 1 (Strong Duality and Existence of Saddle-Point)**.** *The following statements hold:*
(a) The optimal value of (*MA-C-VS*) *satisfies*

$$\underline{C} = \inf_{\mu \in M_1(\mathcal{F}\times\mathcal{G})} \sup_{\lambda \in \mathbb{R}_{\geqslant 0}} \widehat{L}(\mu, \lambda). \tag{15}$$

(b) Strong duality holds for (*MA-C-VS*)*, i.e.,*

$$\begin{aligned}\underline{C} &= \inf_{\mu \in M_1(\mathcal{F}\times\mathcal{G})} \sup_{\lambda \in \mathbb{R}_{\geqslant 0}} \widehat{L}(\mu, \lambda) \\ &= \sup_{\lambda \in \mathbb{R}_{\geqslant 0}} \inf_{\mu \in M_1(\mathcal{F}\times\mathcal{G})} \widehat{L}(\mu, \lambda).\end{aligned} \tag{16}$$

Moreover, there exist $\mu^\star \in M_1(\mathcal{F} \times \mathcal{G})$ and $\lambda^\star \in \mathbb{R}_{\geqslant 0}$ such that the following saddle-point condition holds for all $(\mu, \lambda) \in M_1(\mathcal{F} \times \mathcal{G}) \times \mathbb{R}_{\geqslant 0}$,

$$\widehat{L}(\mu^\star, \lambda) \leqslant \widehat{L}(\mu^\star, \lambda^\star) \leqslant \widehat{L}(\mu, \lambda^\star), \tag{17}$$

that is, $\mu^\star$ minimizes $\widehat{L}(\cdot, \lambda^\star)$ and $\lambda^\star$ maximizes $\widehat{L}(\mu^\star, \cdot)$.

B. Sufficiency of Factorized Transmission Policies

Proposition 1 yields an important structural simplification. Deferring the calculation of the optimal $\lambda^\star$ using a primal-dual approach, here we restrict attention to solving the primal problem resulting from the Lagrangian decomposition. Given a candidate Lagrange multiplier λ, this decomposition asserts that instead of viewing the problem as a BTS with a single joint constraint and sending packets to N EDs, one may view it as an unconstrained problem wherein the BTS has N transmitters—each sending packets to a specific ED. Here, the BTS pays a cost of $\lambda e^n(U_t^n)$ for the n^{th} transmitter, whereas ED-n pays a cost of $c^n(X_t^n, V_t^n)$. Despite this simplification, the information structure remains unchanged; each of the N transmitters has access to the delayed feedback from all the N EDs. In general, this would require the BTS to randomize over the space of all of its transmission actions, $\mathcal{U}^N$. Next, we show that with factorized transition law and additively separable costs, the problem admits a pair-wise factorization, that is, each of the N transmitters can discard the information about all the receivers except its own. Thus, each transmitter can be viewed as an agent on its own which randomizes over its individual action space, $\mathcal{U}$. We thus obtain N independent unconstrained *transmitter-receiver* problems—all sharing the (common) Lagrange multiplier λ.

Definition 1. *A factorized transmission policy f is one whose t^{th} component f_t is given by $f_t = \prod_{n=1}^N f_t^n$ where each f_t^n returns a probability distribution on $\mathcal{U}$ for a given realization of $\widetilde{H}_t^n$. The returned distribution is then used by the BTS to draw (independently) its transmission action for ED-n, i.e., $U_t^n \sim f_t^n(\widetilde{H}_t^n)$. (See* (5) *for the definition of $\widetilde{H}_t^n$.) We denote the space of all factorized transmission policies by $^{\otimes}\mathcal{F}$.*

The following lemma states that one can restrict the search of optimal transmission policies to $^{\otimes}\mathcal{F}$.

Lemma 1. *Consider the constrained team optimization problem* (*MA-C-VS*)*. Let the EDs' play-out policies be fixed to*

$g \in \mathcal{G}$. Then, for every transmission policy $\hat{f} \in \mathcal{F}$, there exists $f \in {}^{\otimes}\mathcal{F}$ such that f performs the same as $\hat{f}$.

Proof. Fix $g \in \mathcal{G}$ and $\hat{f} \in \mathcal{F}$. The objective and constraint costs are separable (see (6) and (7)). If for two transmission policies, the joint marginals of (X_t^n, U_t^n, V_t^n) are the same for every $t \in \mathbb{N}$ and $n \in [N]$, then they incur the same long-term costs. For a fixed $n \in [N]$, we have

$$\begin{aligned}
&\mathbb{P}_{x_1}^{(\hat{f},g)}\left(x_t^n, u_t^n, v_t^n\right) \\
&\overset{(a)}{=} \sum_{\widetilde{h}_t^n} \mathbb{P}_{x_1}^{(\hat{f},g)}\left(\widetilde{h}_t^n, u_t^n\right) \mathbb{P}_{x_1}^{(\hat{f},g)}\left(x_t^n, v_t^n \middle| \widetilde{h}_t^n, u_t^n\right) \\
&\overset{(b)}{=} \sum_{\widetilde{h}_t^n} \mathbb{P}_{x_1}^{(\hat{f},g)}\left(\widetilde{h}_t^n, u_t^n\right) \mathbb{P}_{x_1^n}^{(g^n)}\left(x_t^n, v_t^n \middle| \widetilde{h}_t^n, u_t^n\right). \qquad (18)
\end{aligned}$$

Here, (a) uses the law of total probability over all realizations of $\widetilde{H}_t^n$; and (b) follows because X_t^n, V_t^n are independent of $\hat{f}, g^{-n}, X_1^{-n}$ given $\widetilde{H}_t^n$, U_t^n, and X_1^n.[5]

Based on (18), if we can define a factorized transmission policy $f \in {}^{\otimes}\mathcal{F}$ such that for all $t \in \mathbb{N}$,

$$\mathbb{P}_{x_1}^{(\hat{f},g)}\left(\widetilde{h}_t^n, u_t^n\right) = \mathbb{P}_{x_1^n}^{(f^n,g^n)}\left(\widetilde{h}_t^n, u_t^n\right), \qquad (19)$$

then our proof is complete. To this end, we define f component-wise as follows:

$$f_t^n\left(u_t^n \middle| \widetilde{h}_t^n\right) \triangleq \begin{cases} \frac{\mathbb{P}_{x_1}^{(\hat{f},g)}\left(\widetilde{h}_t^n, u_t^n\right)}{\mathbb{P}_{x_1}^{(\hat{f},g)}\left(\widetilde{h}_t^n\right)}, & \text{if } \mathbb{P}_{x_1}^{(\hat{f},g)}(\widetilde{h}_t^n) > 0, \\ \frac{1}{|\mathcal{U}|}, & \text{otherwise.} \end{cases}$$

We will now proceed with the proof using mathematical induction. The base case for (19) (when $t = 1$) is trivial. Assume that (19) is true for time $t \in \mathbb{N}$, i.e.,

$$\mathbb{P}_{x_1}^{(\hat{f},g)}\left(\widetilde{h}_t^n, u_t^n\right) = \mathbb{P}_{x_1^n}^{(f^n,g^n)}\left(\widetilde{h}_t^n, u_t^n\right).$$

Then,

$$\begin{aligned}
&\mathbb{P}_{x_1}^{(\hat{f},g)}\left(\widetilde{h}_{t+1}^n\right) \\
&\overset{(a)}{=} \mathbb{P}_{x_1}^{(\hat{f},g)}\left(\widetilde{h}_t^n, u_t^n\right) \mathbb{P}_{x_1}^{(\hat{f},g)}\left(x_{t+1-d}^n, v_{t+1-d}^n \middle| \widetilde{h}_t^n, u_t^n\right) \\
&\overset{(b)}{=} \mathbb{P}_{x_1^n}^{(f^n,g^n)}\left(\widetilde{h}_t^n, u_t^n\right) \mathbb{P}_{x_1^n}^{(g^n)}\left(x_{t+1-d}^n, v_{t+1-d}^n \middle| \widetilde{h}_t^n, u_t^n\right) \\
&\overset{(c)}{=} \mathbb{P}_{x_1^n}^{(f^n,g^n)}\left(\widetilde{h}_{t+1}^n\right).
\end{aligned}$$

Here, (a) uses $\widetilde{H}_{t+1}^n = (\widetilde{H}_t^n, X_{t+1-d}^n, V_{t+1-d}^n, U_t^n)$; and (b) uses the inductive hypothesis and conditional independence of X_{t+1-d}^n, V_{t+1-d}^n from $\hat{f}$, g^{-n} and X_1^{-n}. By definition of f, it then follows from (c) that (19) is true for time $t+1$. □

IV. Learning Based Multi-Agent Control Via Approximate Information States in a Single Unconstrained Transmitter-Receiver Problem

In light of Proposition 1 and Lemma 1, we now focus on solving the unconstrained version of a single (say n^{th}) transmitter-receiver problem wherein the immediate cost is parametrized by the Lagrange multiplier λ, namely

$$l^n\left(X_t^n, V_t^n, U_t^n; \lambda\right) = c^n\left(X_t^n, V_t^n\right) + \lambda\left(e^n\left(U_t^n\right) - K\right). \quad (20)$$

In this problem, at time t, the transmitter's information is $H_t^{BTS\text{-}n} = \widetilde{H}_t^n$, and the receiver's information is H_t^n; the information structure is nested, i.e., $H_t^n \supseteq H_t^{BTS\text{-}n}$.

A. The Coordinator's Viewpoint

To solve this cooperative two-agent POMDP problem, we can use the common-information (CI) approach (see [1] for details) using which we can transform it into a single-agent POMDP problem. This is achieved by constructing a *coordinated system* from the point of view of a fictitious *coordinator* who observes only the common observations of the agents but not the private ones. Thus, from the perspective of the coordinator, the unknown state is $(X_{t+1-d:t}^n, V_{t+1-d:t-1}^n)$.

At time t, the coordinator decides *prescriptions* for the transmitter and receiver that map their respective local information to their decisions. This choice of prescriptions is based on the realization of the common information and the prescriptions the coordinator has chosen before time t. In our setting, since the transmitter has no private information, the coordinator's prescription for the transmitter is simply a prescribed decision, U_t^n. On the other hand, the prescription for the receiver, Γ_t^n, is a mapping from $\mathcal{X}^d \times \mathcal{V}^{d-1}$ to $\mathcal{V}$ which the receiver uses to generate its action as follows:

$$V_t^n = \Gamma_t^n\left(X_{t-d+1:t}^n, V_{t-d+1:t-1}^n\right). \qquad (21)$$

Denoting the policy of the coordinator by $\psi = \left(\psi_{1:\infty}^{BTS\text{-}n}, \psi_{1:\infty}^n\right)$, we have

$$\begin{aligned}
U_t^n &= \psi_t^{BTS\text{-}n}\left(H_t^{BTS\text{-}n}, \Gamma_{1:t-1}^n\right), \\
\Gamma_t^n &= \psi_t^n\left(H_t^{BTS\text{-}n}, \Gamma_{1:t-1}^n\right).
\end{aligned} \qquad (22)$$

The system dynamics and the cost are same as in the original problem, and one can show equivalence between the original system with pure policy-profiles and (the new) coordinated system—which in light of Remark 1 is sufficient.

Remark 2. *As the common information between the transmitter and receiver is the same as the information at the transmitter, the transmitter can play the role of the coordinator.*

B. Approximate Information State Representations

A key issue in solving the coordinated system is that the domain of coordinator's information grows exponentially in time, and whereas one may compress it, without loss of optimality, to a belief-based information state (IS), the update of the belief-based IS requires knowledge of $\mathcal{P}_{tr}$ (the transition-law) which is not available in the learning context. To address this challenge, we will use the notion of approximate information states (AISs) [25] specialized for multi-agent systems [2], [3]. The aim of such approximations is to compress the common and private histories of agents into a system statistic that can be used for developing approximately optimal solutions, akin to the belief-based IS used for optimal control in cooperative

[5] Here, q^{-n} and $q_\star^{-n}$ denote $q^{1:N}\backslash q^n$ and $q_\star^{1:N}\backslash q_\star^n$ respectively.

multi-agent POMDPs. For more details, we refer the reader to [2], [3].

The compression framework in [3], applied to our two-agent problem, involves the following steps: *i*) compressing the private history of the receiver to an *approximate sufficient private state* (ASPS), denoted by $\hat{Z}_t^n$. This leads to a reduction in the space of coordinator's prescriptions. We denote the random variable representing the reduced prescription by $\hat{\Gamma}_t^n$. Finally, the coordinator's prescription-observation history (now with the reduced prescriptions) is compressed to an *approximate sufficient common state* (ASCS), denoted by $\hat{Z}_t^{BTS\text{-}n}$.

Next, we specify the ASPS and ASCS properties at a high-level, and refer the reader to [3] for the details of the mathematical formulations. We have the following properties for the ED's ASPS:

1) It evolves (in a time-homogeneous manner) to $\hat{Z}_t^n$ based on $\hat{Z}_{t-1}^n$, X_t^n, V_{t-1}^n. Optionally, one may also include X_{t-d}^n, V_{t-d}^n, U_{t-1}^n, $c^n\left(X_{t-1}^n, V_{t-1}^n\right)$, and $e^n\left(U_{t-1}^n\right)$.
2) Using it, the agents' actions, and the coordinator's history, the immediate objective and constraint costs can be well-approximated.
3) Using it, the agents' actions, and the coordinator's history, the next observations of the transmitter and the receiver can be well-approximated.

With an ASPS and reduced prescriptions in hand, the ASCS satisfies the following properties:

1) It evolves (in a time-homogeneous manner) to $\hat{Z}_t^{BTS\text{-}n}$ based on $\hat{Z}_{t-1}^{BTS\text{-}n}$, X_{t-d}^n, V_{t-d}^n, U_{t-1}^n, and $\hat{\Gamma}_{t-1}^n$. Optionally, one may include $e^n\left(U_{t-1}^n\right)$
2) Using it and the reduced prescription, the immediate objective and constraint costs can be well-approximated.
3) Using it and the reduced prescription, the next observations of the transmitter can be well-approximated.

The existence of good ASPS and ASCS representations helps achieve approximate optimality, see [2], [3]. During deployment, realizations of the ASCS are used to produce actions at the BTS and prescriptions for the ED, and the ED uses the ASPS and the suggested prescription to determine its action. Next, we use neural-networks as function-approximators to learn ASPS and ASCS representations and corresponding ASPS-ASCS based control policies.

C. Neural-Network Architecture for Learning-Based Control

The notions of ASCS and ASPS lead to a multi-agent reinforcement learning (MARL) algorithmic framework [2], [3]. Here, we present our implementation of such a framework for a single unconstrained transmitter-receiver problem. The framework is based on centralized training distributed execution (CTDE), assumes a constant Lagrange multiplier λ, and performs training in a two time-scale stochastic approximation setup. First, the ASCS/ASPS are learnt on a fast time-scale, and then the corresponding control policies are learnt on a slower time-scale. Fig. 4 shows the architectural setup. Below we give a brief outline of the different neural-networks, in line with the high-level descriptions of ASPS and ASCS

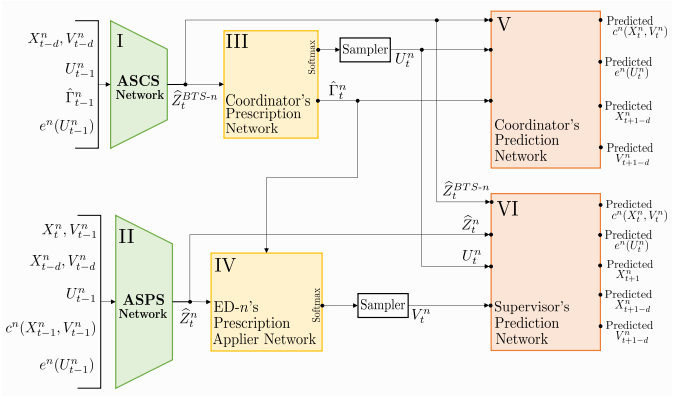

Fig. 4. Neural-network based architecture for learning-based multi-agent control in a single transmitter-receiver problem.

given earlier. Further details on the implementation including training of the neural-networks and the specific loss functions, are similar to those prescribed in [3], and are skipped due to space constraints.

1) Block I is a recurrent neural network (RNN) that tries to learn a good ASCS.
2) Block II, another RNN, tries to learn a good ASPS.
3) Block III is a feed-forward neural-network (FNN) for generating the prescriptions, namely a distribution on $\mathcal{U}$ (used by the transmitter to draw its transmission action U_t^n), and a *pseudo-prescription* $\hat{\Gamma}_t^n$[6] for the receiver.
4) Block IV, another FNN, uses the coordinator's pseudo-prescription $\hat{\Gamma}_t^n$ and the ASPS $\hat{Z}_t^n$ to generate a distribution on $\mathcal{V}$ (used by the receiver to draw its play-out action V_t^n).
5) Block V, another FNN, learns an ASCS via prediction of the objective and constraint costs, and the next observations of the transmitter.
6) Block VI, another FNN, learns an ASPS via prediction of the objective and constraint costs, and the next observations of the transmitter and the receiver.

Remark 3. *Blocks V and VI are not used in the (distributed) execution phase.*

V. Simulation Results

In this section, we perform simulations on a single (unconstrained) transmitter-receiver problem with a stylized QoE model. We will compare two alternatives: 1) the decentralized team set-up using the CI approach with a controller each at the BTS and the ED; and 2) the BTS choosing both transmission and play-out actions using the fed-back (delayed) state from the ED as the ED's current state. We construct an environment simulator with a BTS and an ED. The state of the ED is the tuple of the number of video chunks buffered ($X_t^{1,n}$) and the number of stalls incurred thus far ($X_t^{2,n}$). In a single time-step, the ED may choose to 'play' out ($V_t^n = 1$) a single

[6]The term *pseudo-prescription* is used because the original interpretation of a prescription is lost.

video chunk (if possible) or 'pause' ($V_t^n = 0$) for buffering chunks. The BTS, on the other hand, has three possible actions namely 'high' ($U_t^n = 2$), 'medium' ($U_t^n = 1$) and 'low' ($U_t^n = 0$), corresponding to transmission energies that yield different transmission success probabilities. Each successful transmission results in L new chunks in the ED's playback buffer if there is room; if not, the L chunks are discarded. It is assumed that (after the Lagrangian decomposition), the costs of high, medium, and low transmission actions are 2λ, λ, and 0, respectively. The ED gets a reward r for successful chunk playback while it suffers a cost due to stalling or pausing the video—the cost of a stall event is proportional to the number of stalls thus far. The transition-law and cost functions are as follows.

Transition-law: Let $\widetilde{X} = (X_t^{1,n} - V_t^n)^+ + LR_t^n$ and $R_t^n \in \{0, 1\}$ denote the random variable that indicates the packet's successful reception at time t. Then,

$$X_{t+1}^{1,n} = \begin{cases} \widetilde{X} & \text{if } \widetilde{X} \leqslant B, \\ \left(X_t^{1,n} - V_t^n\right)^+ & \text{otherwise.} \end{cases}$$

$$X_{t+1}^{2,n} = \min\left(M, X_t^{2,n} + \mathbb{1}[\{V_t^n = 0\} \cup \{X_t^{1,n} = 0\}]\right).$$

Here, B is the playback buffer's chunk capacity and M is the maximum possible value for the stall counter.

Immediate objective cost: Let $k > 0$. Then,

$$c^n\left(X_t^n, V_t^n\right) \triangleq -rV_t^n \mathbb{1}[X_t^{1,n} > 0] + kX_t^{2,n}\left(\mathbb{1}\left[\{X_t^{1,n} = 0\} \cup \{V_t^n = 0\}\right]\right).$$

Immediate constraint cost: $e^n\left(U_t^n\right) \triangleq U_t^n$.

Few important parameters used in the simulations are shown in Table I. Next, we describe the two aforementioned approaches towards control of this system.

Delay-oblivious BTS Choosing Both Transmission and Play-out Actions: Here, the delayed receiver's state is used to obtain both the transmission and play-out actions. If the play-out action is infeasible ('play' action chosen with an empty playback buffer), it defaults to the 'pause' action. The RL algorithm used is Reinforce [30]. In the discussions that follow, we refer to this approach as 'SA-DS' for single-agent delayed-state.

TABLE I
IMPORTANT SIMULATION PARAMETERS

Parameter	Value
Episode's time-horizon	50 time-steps
Discount factor	0.95
State space	(Buffer Length, Stall-Count)
Max. playback buffer length (B)	30
Max. stall-count (M)	30
Transmission actions & success probs.	(low: 0.0, med: 0.2, hi: 0.85)
Play-out actions	(Pause: 0, Play: 1)
Packet's chunk-size L	2
Cost function parameters	$k = 40$, $r = 10$, $\lambda = 6$
No. of iterations/gradient-steps	30,000

Delay-aware Two-agent Coordination with AISs: Here, the approximate information states (ASPS and ASCS) are computed using separate long short-term memory networks (LSTMs). The ASCS is used to generate a distribution on the transmission action and a pseudo-prescription for the ED. The psuedo-prescription is then passed through a prescription-applier network to output a distribution on the play-out action. Again, the RL algorithm used for learning of the ASPS-ASCS based control policies is Reinforce. We implement the system based on single-agent AIS codebase from [26], which we extended to the multi-agent learning system shown in Fig. 4. We will refer to this approach as 'MA-AIS' for multi-agent approximate-information-state.

Given the simulation setup, we now answer important performance analysis questions.

Can we successfully train multi-agent based video streaming policies with partial observability? Fig. 5 shows the training curves of SA-DS and MA-AIS for different delays. The x-axis shows the number of training iterations (gradient steps), and the y-axis displays the running average (over a window of 25 iterations) of episode-reward, i.e., negative of the episode's total (discounted) cost. Each episode has a finite time-horizon of 50 time-steps where a discount factor of 0.95 is used. We note that the training performance of MA-AIS is almost invariably bounded below by that of SA-DS.

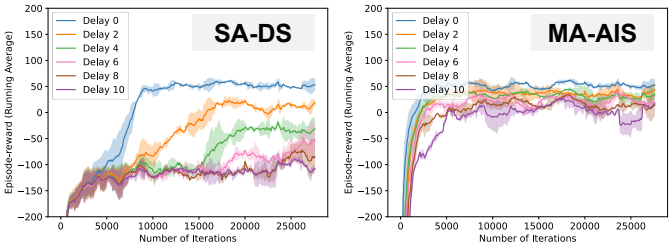

Fig. 5. Training curves of SA-DS and MA-AIS for different delays. (The plot was generated using three distinct seeds.)

What is the impact of MA-AIS on episode-rewards, stall-counts, and playback buffer lengths? We study this in Fig. 6 which was generated using 1000 test episodes on the trained SA-DS and MA-AIS models. In (a) we note that, as the training curves indicated, MA-AIS outperforms SA-DS more significantly at larger delays. It is interesting to note that for delays of 6 units or higher, only MA-AIS yields a positive episode-reward. In (b) the empirical CDF of stall-count follows on expected lines—MA-AIS has fewer stalls than SA-DS. We would also expect that MA-AIS can be more aggressive in maintaining shorter playback buffer lengths; in (c), we see that this is indeed the case.

What is the structure of the learned policies? We delve into the structure of the learned policies for a delay of 10 time-steps. Fig. 7 (generated using 1000 test episodes on the trained SA-DS and MA-AIS models) shows the fraction of times each transmission action is taken at a specific length of the playback buffer (aggregated over all stall-count values). We note that for a given length of playback buffer, SA-DS (which has lower

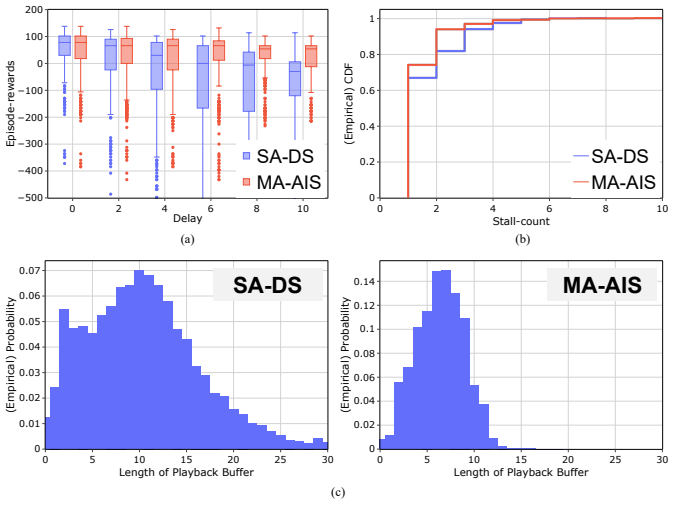

Fig. 6. Comparison of SA-DS and MA-AIS in episode-rewards, stall-counts, and playback buffer lengths.

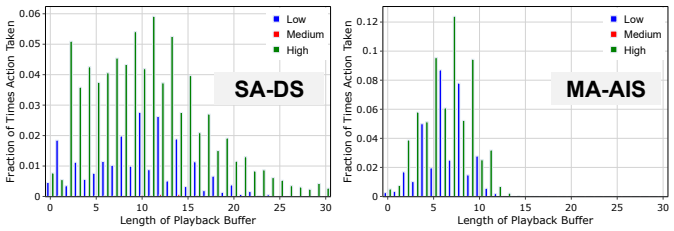

Fig. 7. Comparison of SA-DS and MA-AIS in transmission policy.

visibility into the ED's state) chooses the 'high' action even at large lengths of the playback buffer. The learnt play-out policy in the case of both SA-DS and MA-AIS is to always play out a chunk if possible.

VI. Conclusion

In this work, we explored the idea of application-aware learning agents which engage in cross-layer optimization for wireless resource allocation with imperfect information. We showed how a complex multi-dimensional problem with delayed client-feedback, in the context of video streaming, can be reduced to multiple two-agent POMDP problems (all sharing a common Lagrange multiplier). Furthermore, once decomposed, the individual POMDPs are amenable to multi-agent RL (MARL) using the notion of approximate information states (AISs) as the number of agents is two. We showed how to construct AIS-based learning agents, and illustrated the performance improvement over a vanilla RL approach in which the BTS perceives the delayed state of the ED as the ED's current state and decides both the transmission and play-out actions.

Acknowledgment

This work was funded in part by NSF grants CNS-1955777, CCF-2008130, CMMI-2240981, ECCS-2038416, ECCS-2038963, CNS-1955696, and CNS-2312978, START grant from Michigan Engineering, and ARO grant W911NF-19-1-0367. All opinions are those of the authors and do not necessarily represent those of the funding agencies.

Appendix A
List of Important Symbols

- N: Number of end-devices (EDs).
- X_t^n: State of ED-n at time t.
- $X_t^{1,n}$: State of ED-n's playback buffer at time t.
- $X_t^{2,n}$: QoE tracker for ED-n at time t.
- $\mathcal{X}$: Finite set of all possible realizations of X_t^n.
- V_t^n: Play-out action of ED-n at time t.
- $\mathcal{V}$: Finite set of all possible play-out actions for a given ED.
- U_t^n: Transmission action of BTS for ED-n at time t.
- $\mathcal{U}$: Finite set of all possible transmission actions for a given ED.
- d: Denotes delay in each ED's feedback.
- $\mathcal{P}_{tr}$, P^n: See (1) and (2).
- H_t^{BTS}: Information available to BTS right before it chooses its transmission actions at time t. See (3).
- H_t^n: Information available to ED-n right before it chooses its play-out action at time t. See (4).
- $\widetilde{H}_t^n$: See (5).
- $\mathcal{F}$: Set of all possible transmission policies.
- $\mathcal{G}^n$: Set of all possible play-out policies of ED-n.
- $\mathcal{G}$: $\prod_{n=1}^{N} \mathcal{G}^n$.
- c, c^n, e, e^n: See (6) and (7).
- α: Discount factor.
- $C, E, \widehat{C}, \widehat{E}$: See (8)-(11).
- K: Total long-term discounted energy budget of the BTS. See (*MA-C-VS*).
- $\underline{C}$: Optimal value of (*MA-C-VS*) problem.
- $\widehat{L}, L$: See (14).
- λ: Lagrange multiplier.
- $\otimes\mathcal{F}$: Set of all factorized transmission policies.
- l^n: See (20).
- $H_t^{BTS\text{-}n}$: Same as $\widetilde{H}_t^n$.
- Γ_t^n: Prescription of the coordinator for ED-n at time t. See (21).
- ψ: Coordinator's (pure) policy in the the coordinated system. See (22).
- $\hat{Z}_t^n$: ASPS in the n^{th} transmitter-receiver problem.
- $\hat{\Gamma}_t^n$: Coordinator's (reduced) prescription (based on ASPS) in the n^{th} transmitter-receiver problem.
- $\hat{Z}_t^{BTS\text{-}n}$: ASCS in the n^{th} transmitter-receiver problem.

References

[1] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized Stochastic Control with partial history sharing: A common information approach," *IEEE Trans. on Automatic Control*, vol. 58, no. 7, pp. 1644–1658, 2013.

[2] H. Kao and V. Subramanian, "Common information based approximate state representations in Multi-Agent Reinforcement Learning," in *AISTATs*, vol. 151. PMLR, 28–30 Mar 2022, pp. 6947–6967.

[3] N. Khan and V. Subramanian, "Cooperative multi-agent constrained POMDPs: Strong duality and primal-dual reinforcement learning with approximate information states," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.16536

[4] Y. Xu, S. E. Elayoubi, E. Altman, R. El-Azouzi, and Y. Yu, "Flow-level qoe of video streaming in wireless networks," *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2762–2780, 2015.

[5] S. Poojary, R. El-Azouzi, E. Altman, A. Sunny, I. Triki, M. Haddad, T. Jimenez, S. Valentin, and D. Tsilimantos, "Analysis of QoE for adaptive video streaming over wireless networks," in *WiOpt 2018*, 2018, pp. 1–8.

[6] Y. He, I. Lee, and L. Guan, "Optimized multi-path routing using dual decomposition for wireless video streaming," in *2007 IEEE International Symposium on Circuits and Systems*, 2007, pp. 977–980.

[7] L. Zhou, B. Geller, B. Zheng, A. Wei, and J. Cui, "System scheduling for multi-description video streaming over wireless multi-hop networks," *IEEE Transactions on Broadcasting*, vol. 55, no. 4, pp. 731–741, 2009.

[8] H. Seferoglu, L. Keller, B. Cici, A. Le, and A. Markopoulou, "Cooperative video streaming on smartphones," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2011, pp. 220–227.

[9] F. Fu and M. Van Der Schaar, "A systematic framework for dynamically optimizing multi-user wireless video transmission," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 3, pp. 308–320, 2010.

[10] G. Xiong, X. Qin, B. Li, R. Singh, and J. Li, "Index-aware reinforcement learning for adaptive video streaming at the wireless edge," in *ACM-MobiHoc*, New York, NY, USA, 2022, p. 81–90. [Online]. Available: https://doi.org/10.1145/3492866.3549726

[11] R. Singh and P. R. Kumar, "Dynamic adaptive streaming using index-based learning algorithms," 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1612.05864

[12] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," ser. SIGCOMM '17, 2017, p. 197–210. [Online]. Available: https://doi.org/10.1145/3098822.3098843

[13] R. Bhattacharyya, A. Bura, D. Rengarajan, M. Rumuly, S. Shakkottai, D. Kalathil, R. K. Mok, and A. Dhamdhere, "QFlow: A reinforcement learning approach to high qoe video streaming over wireless networks," in *ACM-Mobihoc*, 2019, pp. 251–260.

[14] Y. Li, Q. Zheng, Z. Zhang, H. Chen, and Z. Ma, "Improving abr performance for short video streaming using Multi-Agent Reinforcement Learning with expert guidance," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.04637

[15] A. Bentaleb, B. Taani, A. C. Begen, C. Timmerer, and R. Zimmermann, "A survey on bitrate adaptation schemes for streaming media over http," *IEEE Comms Surveys & Tutorials*, vol. 21, no. 1, pp. 562–585, 2019.

[16] K. J. Astrom, "Optimal control of Markov processes with incomplete state information," *Journal of Mathematical Analysis and Applications*, vol. 10, pp. 174–205, 1965.

[17] F. A. Oliehoek and C. Amato, "A concise introduction to decentralized POMDPs," in *SpringerBriefs in Intelligent Systems*, 2016.

[18] A. R. Cassandra, M. L. Littman, and N. L. Zhang, "Incremental pruning: A simple, fast, exact method for Partially Observable Markov Decision Processes," 2013. [Online]. Available: https://doi.org/10.48550/arXiv.1302.1525

[19] D. Kim, J. Lee, K.-E. Kim, and P. Poupart, "Point-Based Value Iteration for Constrained POMDPs," in *IJCAI'11*. AAAI Press, 2011, p. 1968–1974.

[20] J. Lee, G.-h. Kim, P. Poupart, and K.-E. Kim, "Monte-Carlo tree search for constrained POMDPs," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[21] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queueing network control," *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*, pp. 318–322, 1994. [Online]. Available: https://api.semanticscholar.org/CorpusID:11950868

[22] J. S. Dibangoye, C. Amato, O. Buffet, and F. Charpillet, "Optimally solving dec-POMDPs as continuous-state MDPs," *J. Artif. Int. Res.*, vol. 55, no. 1, p. 443–497, jan 2016.

[23] R. D. Smallwood and E. J. Sondik, "The optimal control of Partially Observable Markov Processes over a finite horizon," *Operations research*, vol. 21, no. 5, pp. 1071–1088, 1973.

[24] D. Abel, D. E. Hershkowitz, and M. L. Littman, "Near optimal behavior via approximate state abstraction," 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1701.04113

[25] J. Subramanian and A. Mahajan, "Approximate information state for Partially Observed systems," in *CDC 2019*, 2019, pp. 1629–1636.

[26] J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan, "Approximate information state for approximate planning and reinforcement learning in Partially Observed systems," *JMLR*, vol. 23, no. 12, pp. 1–83, 2022.

[27] N. Khan and V. Subramanian, "A strong duality result for constrained POMDPs with multiple cooperative agents," 2023, To appear in Proceedings of IEEE CDC 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.14932

[28] A. Nayyar, A. Mahajan, and D. Teneketzis, *The Common-Information Approach to Decentralized Stochastic Control*. Cham: Springer International Publishing, 2014, pp. 123–156.

[29] R. J. Aumann, *Mixed and behavior strategies in infinite extensive games*. Princeton University Princeton, 1961.

[30] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.