2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Online Learning Algorithms for Zero-Sum Games of Linear Systems over Wireless MIMO Fading Channels with Uncountable State Space

Minjie Tang and Vincent K. N. Lau

***Abstract*— In this paper, we consider an online learning framework for a zero-sum game of an unstable linear dynamic system in the presence of wireless MIMO fading channels between the remote controllers and the actuator of the dynamic plant. We first formulate the stochastic zero-sum game as ergodic optimization problems, and propose a pair of equivalent reduced-state Bellman optimality equations to address the "curse of dimensionality" for Nash equilibrium of the game. Based on the reduced-state Bellman optimality equations, we analyze the structural properties of the Nash equilibrium and propose a novel low-complexity online stochastic-approximation(SA)-based algorithm to solve the reduced-state Bellman optimality equations. Numerical results are analyzed for the proposed learning scheme and several state-of-the-art learning approaches in terms of the computational complexity, the convergence performance as well as the robustness performance. We show that a significant performance gain can be achieved by the proposed scheme compared to the baseline approaches.**

I. Introduction

Serving as a powerful mathematical approach for robust control of dynamic systems, the zero-sum games attract great interest over the past decade [1], [2]. The approach considers finding the optimal control policies to stabilize unstable control systems in the presence of the negative effects caused by disturbances. Basically, the controller and the disturbance in a dynamic system can be viewed as a stabilizing and a destabilizing controller that struggles to minimize their control performance under the zero-sum constraint by a competitive approach, respectively [3]. We consider a typical linear system for a non-cooperative zero-sum game that is comprised of a potentially unstable *dynamic plant*, a *stabilizing controller*, a *destabilizing controller* as well as an *actuator* collocated with the dynamic plant, as shown in Fig. 1. Specifically, the controllers generate real time control actions based on the instantaneous state feedback from the dynamic plant, and deliver the control commands to the actuator over the wireless network to neutralize the instability of the dynamic plant. The presence of the wireless network between the remote controllers and the actuator of the dynamic plant will induce several impairments such as fading and dropout, which seriously jeopardize the stability of the dynamic system.

The zero-sum games for linear systems over static channels have been widely considered in the existing literatures [4]–[9], and the optimal solutions, or the Nash equilibrium of the games are iteratively obtained via numerical approaches such as value iteration, the policy iteration and Q-learning method. Note that in all the above works, the static channels were assumed between the remote controllers and the actuators. Existing approaches [4]–[9] cannot learn the Nash equilibrium of the zero-sum games over wireless networks because the learned control policies via these approaches are static while the optimal control policies should adapt to both the wireless fading realizations between the remote controllers and the actuators (to capture good transmission opportunities induced by the fading channels) as well as the dynamic plant state realizations (to capture the dynamic urgency of the control).

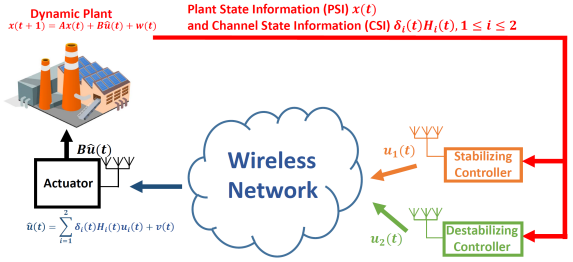

Fig. 1: The architecture of the zero-sum games of linear systems over the wireless network.

Recently, the zero-sum games for linear systems over wireless communication channels have been investigated in [10]–[15]. Specifically, in [13], [15], [16], the authors considered the on-off channels between the remote controllers and the actuators. The structural form of the optimal control solutions are parameterized by the on-off channel state information (CSI) and learned via the value-iteration-based approaches. However, the modeling with 0-1 process in [13], [15], [16] may oversimplify practical wireless communication channels. In [10], [11], the authors considered the zero-sum game for a linear system over the wireless fading channels with finite channel state information (CSI) state space. The optimal control policies are obtained by means of the policy-iteration-based and parallel-learning-based algorithms. However, the assumptions of finite CSI state space in [10], [11] may not hold in practice and brute-force applications of these approaches to the general fading channels with uncountable CSI state space requires discretization of the CSI, which may induce the large deviations of the learned control policies from the equilibrium strategies of the games due to the discretization errors. Moreover, due to large discretized CSI state space, solving the stochastic games in a brute-force manner may lead to the "curse of dimensionality" w.r.t. the

The authors are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology (HKUST), Clear Water Bay, Hong Kong (e-mail: mtangad@connect.ust.hk; eeknlau@ust.hk).

state space.

In this paper, we propose a novel online learning framework for a stochastic zero-sum game of a linear system over wireless MIMO fading channels with uncountable CSI state space. We formulate the problem as a stochastic ergodic game and derive a pair of equivalent "*reduced-state Bellman optimality equations*". A low-complexity online learning solution is derived to solve the optimality equations. The contributions of the paper can be summarized into following three folds: 1) *Reduced-State Bellman Optimality Equations over Uncountable CSI State Space*. We establish equivalent reduced-state Bellman optimality equations for finding the Nash equilibrium of the stochastic zero-sum ergodic game. The optimality equations can be applied to uncountable CSI state space without "curse of dimensionality"; 2) *Closed-Form Structure of Nash Equilibrium*. Based on the reduced-state Bellman optimality equations, we derive the closed-form structure of the Nash equilibrium of the stochastic zero-sum ergodic game; 3) *Online Learning Algorithm Design for the Nash Equilibrium.* Based on the reduced-state Bellman optimality equations, we develop a novel stochastic-approximation(SA)-based online learning algorithm for finding the Nash equilibrium of the stochastic zero-sum ergodic game.

Notation: Uppercase and lowercase boldface denotes matrices and vectors, respectively. The operators $(\cdot)^T$, and $\mathrm{Tr}(\cdot)$ is the transpose and trace of a matrix, respectively. $\mathbf{0}_{m\times n}$ and $\mathbf{0}_m$ denotes an $m\times n$ and $m\times m$ dimensional matrix with all the elements being 0, respectively. $\mathbf{1}_S$ denotes the $S\times S$ dimensional identity matrix. $\mathbb{R}^{m\times n}$, $\mathbb{S}_+^m$, $\mathbb{S}^m$, $\mathbb{Z}_+$ and $\mathbb{R}_+$ denotes the set of $m\times n$ dimensional real matrices, the set of $m\times m$ dimensional positive definite matrices, the set of $m\times m$ dimensional positive semi-definite matrix, the set of positive integers and the set of positive real numbers, respectively. $\|\mathbf{A}\|$ is the spectral norm of a matrix $\mathbf{A}$.

II. System Model

A. Unstable Dynamic Plant

We consider a discrete-time system with S-dimensional plant state $\mathbf{x}(t)\in\mathbb{R}^{S\times 1}$. The remote controllers are equipped with N_t transmission antennas and the actuator is equipped with N_r receiving antennas. The plant system is characterized by first-order coupled linear difference equations, given by

$$\mathbf{x}(t+1)=\mathbf{A}\mathbf{x}(t)+\mathbf{B}\widehat{\mathbf{u}}(t)+\mathbf{w}(t),\ \ t=1,2,..., \qquad (1)$$

where $\widehat{\mathbf{u}}(t)\in\mathbb{R}^{N_r\times 1}$ is the received noisy control signal from the remote controllers at the actuator. $\mathbf{A}\in\mathbb{R}^{S\times S}$ is the unstable drift matrix that characterizes the internal evolution of the dynamic plant and satisfies $\|\mathbf{A}\|>1$. $\mathbf{B}\in\mathbb{R}^{S\times N_r}$ is the actuator matrix. $\mathbf{w}(t)\in\mathcal{N}(0,\mathbf{W})$ is additive plant noise with zero mean and finite noise covariance $\mathbf{W}\in\mathbb{S}_+^S$. We assume the plant system $(\mathbf{A},\mathbf{B})$ is controllable.

B. Wireless MIMO Fading Channel Model

We model the communication channels between the remote controllers and the actuator as wireless MIMO fading channels. The active controllers transmit control signals $\mathbf{u}_i(t)\in\mathbb{R}^{N_t\times 1}, i\in\{1,2\}$, to the actuator through wireless communication channels. The received signal $\widehat{\mathbf{u}}(t)\in\mathbb{R}^{N_r\times 1}$ at the actuator is given by:

$$\widehat{\mathbf{u}}(t)=\delta_1(t)\mathbf{H}_1(t)\mathbf{u}_1(t)+\delta_2(t)\mathbf{H}_2(t)\mathbf{u}_2(t)+\mathbf{v}(t), \qquad (2)$$

where $\mathbf{H}_i(t)\in\mathbb{R}^{N_r\times N_t}$ is the wireless MIMO fading realization between the actuator of the dynamic plant and i-th remote controller, where $i\in\{1,2\}$. $\mathbf{v}(t)\in\mathcal{N}(0,\mathbf{1}_{N_r})$ is the additive channel noise at the actuator. $\delta_i(t)\in\{0,1\}$ is used to model the random access activity of the i-th remote controller. Moreover, $\delta_i(t)$ is i.i.d. distributed across timeslots and remote controllers satisfying $\Pr(\delta_1(t)=1)=\Pr(\delta_2(t)=1)=p\in[0,1]$. We have the following assumption on $\mathbf{H}_i(t)$.

Assumption 1: (Wireless MIMO Fading Channel Model [17]) The realization of wireless MIMO fading channels $\mathbf{H}_i(t)$ between i-th controller and the actuator remains quasi-static within each timeslot and each controller, and is i.i.d. over remote controllers and the timeslots. Moreover, $\mathbf{H}_i(t)$ follows a Gaussian distribution with zero mean and unit variance. ■

Note that the wireless MIMO fading channel model in Assumption 1 is commonly applied in practice. This assumption implies that the symbol period in a timeslot is approximately equal to the coherence time of the wireless channels between the dynamic plant and the remote controllers, and the positions of the controllers remain fixed over time.

C. Problem Formulation for the Stochastic Zero-sum Game over Wireless MIMO Fading Channels

When wireless channels between the remote controllers and the actuator are random, the system is a linear and time-varying system, where the equivalent plant dynamics can be obtained by substituting (2) into (1), given by

$$\mathbf{x}(t+1)=\mathbf{A}\mathbf{x}(t)+\sum_{i=1}^{2}\delta_i(t)\mathbf{B}\mathbf{H}_i(t)\mathbf{u}_i(t)+\mathbf{B}\mathbf{v}(t)+\mathbf{w}(t). \qquad (3)$$

The zero-sum game for linear time-varying system (3) can be modelled over the aggregated state space $\mathcal{S}=\{\mathbf{S}(1),\mathbf{S}(2),...\}$, where $\mathbf{S}(t)=\{\mathbf{x}(t),\delta_1(t)\mathbf{H}_1(t),\delta_2(t)\mathbf{H}_2(t)\}$ is an aggregation of plant state information (PSI) $\mathbf{x}(t)$ and the wireless channel state information (CSI) $\{\delta_1(t)\mathbf{H}_1(t),\delta_2(t)\mathbf{H}_2(t)\}$. The control policy π_i for i-th controller is a mapping from the state space $\mathbf{S}(t)\in\mathcal{S}$ to the control action of i-th controller $\mathbf{u}_i(t)\in\mathcal{U}, t\in\mathbb{Z}_+$. The per-stage utility function for i-th controller $r_i(\mathbf{S}(t),\mathbf{u}_1(t),\mathbf{u}_2(t))$ is given by:

$$\begin{aligned}&r_1(\mathbf{S}(t),\mathbf{u}_1(t),\mathbf{u}_2(t))=\mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t)+\mathbf{u}_1^T(t)\mathbf{R}_1\mathbf{u}_1(t)-\\&\gamma^2\mathbf{u}_2(t)\mathbf{R}_2\mathbf{u}_2(t)+(\delta_1(t)\mathbf{B}\mathbf{H}_1(t)\mathbf{u}_1(t))^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t)\\&\mathbf{u}_1(t)-\gamma^2(\delta_2(t)\mathbf{B}\mathbf{H}_2(t)\mathbf{u}_2(t))^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t)\mathbf{u}_2(t),\end{aligned} \qquad (4)$$

and

$$r_2(\mathbf{S}(t),\mathbf{u}_1(t),\mathbf{u}_2(t))=-r_1(\mathbf{S}(t),\mathbf{u}_1(t),\mathbf{u}_2(t)), \qquad (5)$$

where $\mathbf{Q} \in \mathbb{S}_+^S$, $\mathbf{R}_1 \in \mathbb{S}_+^{N_t}$, $\mathbf{R}_2 \in \mathbb{S}_+^{N_t}$, $\mathbf{M}_1 \in \mathbb{S}_+^S$ and $\mathbf{M}_2 \in \mathbb{S}_+^S$ are weighting constant matrices. $\gamma > 0$ is a positive constant penalizing the non-cooperation between controllers. Notice that the utility functions for remote controllers satisfy the zero-sum constraint [3] given by $r_1(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)) + r_2(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)) = 0, \forall t \in \mathbb{Z}_+$. For $i, j \in \{1, 2\}, i \neq j$, we formally summarize the stochastic zero-sum game for a linear system over the wireless MIMO fading channels as follows.

Problem 1: (The Stochastic Zero-Sum Ergodic Game of a Linear System over Wireless MIMO Fading Channels)

$$\begin{aligned} i\text{-th Controller: } & \min_{\pi_i} \max_{\pi_j} \mathcal{J}_i^{\pi_i, \pi_j} \\ & = \min_{\pi_i} \max_{\pi_j} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}^{\pi_i, \pi_j} [r_i(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t))] \\ s.t. \quad & \mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \sum_{i=1}^{2} \delta_i(t) \mathbf{B}\mathbf{H}_i(t)\mathbf{u}_i(t) + \\ & \mathbf{B}\mathbf{v}(t) + \mathbf{w}(t). \end{aligned} \tag{6}$$

The expectation in (6) is taken w.r.t the random access variable of the remote controllers $\delta_i(t)$, the wireless fading realization $\mathbf{H}_i(t)$, the plant noise $\mathbf{w}(t)$ and the additive channel noise $\mathbf{v}(t)$, where $i \in \{1, 2\}$. ■

The optimal solution to Problem 1 is called the *Nash equilibrium* of the Problem 1 in the following sense.

Definition 1: (Nash Equilibrium [18]) The control policies for remote controllers $\{\pi_1^*, \pi_2^*\}$ are said to constitute the Nash equilibrium of the Problem 1 if

$$\mathcal{J}_i^{\pi_i^*, \pi_j} \leq \mathcal{J}_i^{\pi_i^*, \pi_j^*} \leq \mathcal{J}_i^{\pi_i, \pi_j^*}, \quad \forall \{\pi_i, \pi_j\}, i \neq j \in \{0, 1\}. \tag{7}$$

■

III. Nash Equilibrium of the Stochastic Zero-sum Game of the Linear System over Wireless MIMO Fading Channels

Conventionally, the Nash equilibrium of Problem 1 is obtained via equivalently solving a pair of Bellman optimality equations [3] as follows.

Theorem 1: (Bellman Optimality Equations for Problem 1) If the Nash equilibrium of Problem 1 exists, the Nash equilibrium of Problem 1 can be obtained by the solution of a pair of Bellman optimality equations, given by

$$\begin{aligned} & \theta_i^* + V_i^*(\mathbf{S}(t)) = \min_{\mathbf{u}_i(t)} \max_{\mathbf{u}_j(t)} [r_i(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)) + \\ & \mathbb{E}[V_i^*(\mathbf{S}(t+1)) | \mathbf{x}(t), \{\delta_i(t)\mathbf{H}_i(t), \mathbf{u}_i(t)\}]], \ i \neq j \in \{1, 2\}, \end{aligned} \tag{8}$$

where $\theta_i^* = \mathcal{J}_i^* = \mathcal{J}_i^{\pi_1^*, \pi_2^*}$ is the optimal averaged cost of Problem 1. $V_i^*(\mathbf{S}(t))$ is the optimal value function over the state space $\mathbf{S}(t) = \{\mathbf{x}(t), \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)\}$, and the Nash equilibrium $\{\pi_1^*, \pi_2^*\} = \{\mathbf{u}_1^*(t), \mathbf{u}_2^*(t), \forall t \in \mathbb{Z}_+\}$, where $\mathbf{u}_1^*(t)$ and $\mathbf{u}_2^*(t)$ corresponds to the minimizer and the maximizer of the R.H.S. of (8).

Proof: Please see Chapter 6.7 of [19]. ■

When the Nash equilibrium of Problem 1 exists, the conventional iterative approaches such as value iteration or Q-learning [4], [5], [7], [8] may be considered to solve the Bellman optimality equations (8) for the Nash equilibrium of Problem 1. However, such approaches are extremely difficult to apply due to the "curse of dimensionality" in the state space $\mathbf{S}(t)$. Specifically, the total dimensions of the state space is $S + 2 \times N_t \times N_r + 2$. Brute-force applications of the value iteration and Q-learning-based approaches to solve the optimality equations (8) require the domain knowledge of the optimal value functions $V_i^*(\mathbf{S}(t))$ or the optimal Q-function $Q_i^*(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)), i \in \{1, 2\}$, and $2 \times (S + 2 \times N_t \times N_r + 2)$ or $2 \times (S + 2 \times N_t \times N_r + 2 + 2 \times N_t)$ dimensional parameters are required to be calculated, respectively. This involves the significant computational workloads of learning algorithms for the Nash equilibrium of Problem 1 when the number of plant states S, the number of transmission antennas N_t and the number of receiving antennas N_r are large.

In order to enable the low-complexity implementation of the learning algorithm for the Nash equilibrium of Problem 1, we exploit the i.i.d. properties of CSI $\{\delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)\}$ in (8) and propose equivalent reduced-state Bellman optimality equations as follows.

Theorem 2: (Reduced-State Bellman Optimality Equations for Problem 1) If the Nash equilibrium of Problem 1 exists, the Nash equilibrium of Problem 1 can be obtained by the solution of a pair of equivalent reduced-state Bellman optimality equations, given by[1]

$$\begin{aligned} & \tilde{\theta}_i^* + \tilde{V}_i^*(\mathbf{x}(t)) = \mathbb{E}[\min_{\mathbf{u}_i(t)} \max_{\mathbf{u}_j(t)} [r_i(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)) + \\ & \mathbb{E}[\tilde{V}_i^*(\mathbf{x}(t+1)) | \mathbf{x}(t), \{\delta_i(t)\mathbf{H}_i(t), \mathbf{u}_i(t)\}]]], i \neq j \in \{1, 2\}, \end{aligned} \tag{9}$$

where $\tilde{\theta}_i^* = \theta_i^* = (-1)^{i+1} \mathrm{Tr}(\mathbf{P}\mathbf{W} + \mathbf{B}^T\mathbf{P}\mathbf{B})$ is the optimal averaged cost of Problem 1, $\tilde{V}_i^*(\mathbf{x}(t)) = (-1)^{i+1}\mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t)$ is the optimal reduced-state value function, which is parameterized by $\mathbf{P} \in \mathbb{S}_+^S$, and the Nash equilibrium $\{\pi_1^*, \pi_2^*\} = \{\mathbf{u}_1^*(t), \mathbf{u}_2^*(t), \forall t \in \mathbb{Z}_+\}$, where the solution to (9) $\mathbf{u}_1^*(t) = \mathbf{K}_1(\mathbf{P}, t)\mathbf{x}(t)$, $\mathbf{u}_2^*(t) = \mathbf{K}_2(\mathbf{P}, t)\mathbf{x}(t)$ is the optimal control solution for remote controllers, and

$$\begin{aligned} & \mathbf{K}_1(\mathbf{P}, t) = -(\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) + \delta_1(t) \\ & \mathbf{H}_1^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_1(t)\mathbf{B}\mathbf{H}_1(t))^{-1}\mathbf{H}_1^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_1(t)\mathbf{A}, \end{aligned} \tag{10}$$

$$\begin{aligned} & \mathbf{K}_2(\mathbf{P}, t) = (\gamma^2\delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t) + \gamma^2\mathbf{R}_2 - \\ & \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_2(t)\mathbf{B}\mathbf{H}_2(t))^{-1}\mathbf{H}_2^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_2(t)\mathbf{A}, \end{aligned} \tag{11}$$

$$\begin{aligned} & \widetilde{\mathbf{P}}_1(t) = (\mathbf{P}^{-1} - \gamma^{-2}\delta_2(t)\mathbf{B}\mathbf{H}_2(t)(\mathbf{R}_2 + \delta_2(t)\mathbf{H}_2^T(t) \\ & \mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t))^{-1}\mathbf{H}_2^T(t)\mathbf{B}^T)^{-1}, \end{aligned} \tag{12}$$

$$\begin{aligned} & \widetilde{\mathbf{P}}_2(t) = (\mathbf{P}^{-1} - \delta_1(t)\mathbf{B}\mathbf{H}_1(t)(\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t) \\ & \mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t))^{-1}\mathbf{H}_1^T(t)\mathbf{B}^T)^{-1}. \end{aligned} \tag{13}$$

[1]Note that the i.i.d properties of $\mathbf{H}_i(t)$ across timeslots and controllers lead to homogeneous statistics of CSI at L.H.S. and R.H.S. of (8) and enable state reduction of CSI in (8) without losing the optimality.

Proof: Please see Appendix A. ■

Compared to solving the Bellman optimality equations (8) by learning the optimal value functions $V_i^*(\mathbf{S}(t)), i \in \{1,2\}$ w.r.t. the high-dimensional aggregated state $\mathbf{S}(t) = \{\mathbf{x}(t), \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)\}$, solving the equivalent reduced-state Bellman optimality equations (9) only requires learning of the reduced-state value functions $\tilde{V}_i^*(\mathbf{x}(t)), i \in \{1,2\}$ w.r.t. the low-dimensional plant state $\mathbf{x}(t)$, and hence the implementation complexity of the learning algorithm for Nash equilibrium based on the reduced-state Bellman optimality equations (9) will be significantly smaller than that based on the Bellman optimality equations (8). In Section IV, we shall propose an online learning algorithm to learn the Nash equilibrium of Problem 1 based on the reduced-state Bellman optimality equations (9).

IV. Online Learning Algorithm for the Nash Equilibrium of the Zero-sum Game of the Linear System over Wireless MIMO Fading Channels

Using the structural form of the optimal reduced-state value functions $\tilde{V}_i^*(\mathbf{x}(t)), i \in \{1,2\}$, the optimal averaged costs $\theta_i^*, i \in \{1,2\}$, and the Nash equilibrium $\{\pi_1^*, \pi_2^*\} = \{\mathbf{u}_1^*(t), \mathbf{u}_2^*(t), \forall t \in \mathbb{Z}_+\}$ in Theorem 2, the reduced-state Bellman optimality equations (9) can be written into the coupled nonlinear matrix equation as follows.

$$\mathbf{P} = \mathbb{E}[g(\mathbf{P}, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t))], \tag{14}$$

where $g(\mathbf{P}, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t))$ is given by:

$$g(\mathbf{P}, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)) = \mathbf{Q} + \mathbf{A}^T\mathbf{P}\mathbf{A} - \mathbb{E}\left[\mathbf{A}^T \begin{bmatrix} \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P} \\ \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{P} \end{bmatrix}^T \begin{bmatrix} \mathcal{M}_{11}(t) & \mathcal{M}_{12}(t) \\ \mathcal{M}_{21}(t) & \mathcal{M}_{22}(t) \end{bmatrix}^{-1} \begin{bmatrix} \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P} \\ \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{P} \end{bmatrix} \mathbf{A}\right], \tag{15}$$

and $\mathcal{M}_{11}(t) = \mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_1(t)$, $\mathcal{M}_{12}(t) = \delta_1(t)\delta_2(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_2(t)$, $\mathcal{M}_{21}(t) = \mathcal{M}_{12}^T(t)$, $\mathcal{M}_{22}(t) = -\gamma^2\delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t) + \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_2(t) - \gamma^2\mathbf{R}_2$.

Since (14) is a fixed-point equation w.r.t. the unknown variable $\mathbf{P}$, we can utilize the SA theory [20] to construct an online learning algorithm to learn the unknown variable $\mathbf{P}$ based on (14). The learned unknown variable $\mathbf{P}$ can then be applied to obtain the optimal reduced-state value functions $\tilde{V}_i^*(\mathbf{x}(t)), i \in \{1,2\}$, and the optimal control solution $\mathbf{u}_i^*(t), i \in \{1,2\}$, for the Nash equilibrium of Problem 1 $\{\pi_1^*, \pi_2^*\}$ according to Theorem 2.

Specifically, (14) can be further written into standard form $f(\mathbf{P}) = \mathbf{0}_S$, where $f(\mathbf{P})$ is given by:

$$f(\mathbf{P}) = \mathbb{E}[g(\mathbf{P}, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t))] - \mathbf{P}. \tag{16}$$

To obtain the root of $f(\mathbf{P}) = \mathbf{0}_S$, we apply the SA algorithm as shown in Algorithm 1[2]. Specifically, the estimated root $\mathbf{P}^t$ at t-th timeslot is updated as:

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \alpha^t(g(\mathbf{P}^t, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)) - \mathbf{P}^t), \tag{17}$$

where $\{\alpha^t\}$ is the step-size sequence satisfying $\sum_{t=1}^{\infty}\alpha^t = \infty$ and $\sum_{t=1}^{\infty}(\alpha^t)^2 < \infty$. The term $g(\mathbf{P}^t, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t))$ is an unbiased estimator of the term $\mathbb{E}[g(\mathbf{P}, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t))]$ in (16).

Algorithm 1 Online Learning for the Nash Equilibrium of Problem 1

• **Step 1:** Given a feasible initial value $\mathbf{P}^1 \in \mathbb{S}_+^S$, the initial estimated optimal reduced-state value functions are given by

$$\tilde{V}_i^1(\mathbf{x}(1)) = (-1)^{i+1}\mathbf{x}^T(1)\mathbf{P}^1\mathbf{x}(1), i \in \{1,2\}, \tag{18}$$

and the estimated optimal control solution for remote controllers is given by:

$$\mathbf{u}_i(1) = \mathbf{K}_i(\mathbf{P}^1, 1)\mathbf{x}(1), \ \ i \in \{1,2\}. \tag{19}$$

• **Step 2:** Using $\mathbf{P}^t$ updated at $(t-1)$-th timeslot, the estimated optimal control solution for remote controllers at t-th timeslot is given by:

$$\mathbf{u}_i(t) = \mathbf{K}_i(\mathbf{P}^t, t)\mathbf{x}(t), \ \ i \in \{1,2\}, \tag{20}$$

and the estimated optimal reduced-state value functions $\tilde{V}_i^t(\mathbf{x}(t))$ at t-th timeslot are given by:

$$\tilde{V}_i^t(\mathbf{x}(t)) = (-1)^{i+1}\mathbf{x}^T(t)\mathbf{P}^t\mathbf{x}(t), \ i \in \{1,2\}. \tag{21}$$

• **Step 3:** $\mathbf{P}^{t+1}$ is updated based on $\mathbf{P}^t$ via (17). Then return to Step 2.

V. Numerical Results

In this section, we verify the performance advantages of the proposed online algorithm for the stochastic zero-sum game over wireless MIMO fading channels. Specifically, we compare the proposed stabilizing control scheme for $\mathbf{u}_1(t)$ in the unstable linear system (3) with various existing stabilizing control approaches in the presence of the external inference signal $\mathbf{u}_2(t)$ in (3) generated by the worst-case disturbance in (11) and (13). We summarize the baseline schemes for the stabilizing control solution $\mathbf{u}_1(t)$ as follows.

- **Baseline 1:** *(Prior-Known Nash Equilibrium [3])* The Nash equilibrium of Problem 1 is known at the stabilizing controller. The optimal stabilizing control solution $\mathbf{u}_1^*(t)$ is applied for the system.
- **Baseline 2:** *(Brute-Force Value Iteration for Equilibrium Policy without State Reduction [4])* The uncountable state space of the CSI $\{\delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)\}$ are firstly discretized into finite intervals. The value for the optimal value function of the stabilizing controller is approximated by the value of the pseudo value function $\widehat{V}_1^d(\mathbf{x}(t)) = \mathbf{x}^T(t)\mathbf{P}^d\mathbf{x}(t), \mathbf{P}^d \in \mathbb{S}_+^S, 1 \le d \le (N_t \times N_r \times 2)^2 \times L$ if $\{\delta_i(t)\mathbf{H}_i(t)\}$ belongs to d-th interval. The control policy for the stabilizing controller is learned by value iteration on pseudo value functions.
- **Baseline 3:** *(Brute-Force Value Iteration for Equilibrium Policy over Static Channels [8])* The optimal value function for the stabilizing controller is approximated by the pseudo value function $\widehat{V}_1^s(\mathbf{x}(t)) =$

[2]In the Algorithm 1, the CSI $\{\delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)\}$ will be required. This can be obtained by the standard channel estimation at the actuator based on the received pilot symbol from the remote controllers and channel feedback to the remote controllers [21].

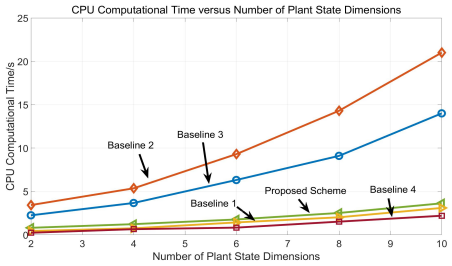

Fig. 2: CPU computational time versus the number of plant state dimensions. The system parameters are configured as follows: $\mathbf{A} \in \mathbb{R}^{S\times S}$ and $\mathbf{B} \in \mathbb{R}^{S\times N_T}$ are randomly generated with zero element in $\mathbf{A}$ and $\mathbf{B}$ generated following Gaussian distribution with zero mean and unit variance. $p = 0.8$, $N_t = N_r = 2$, $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}_2$, $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{Q} = \mathbf{I}_S$, $\gamma = 10$, $\mathbf{W} = \mathbf{I}_S$, and $L = 2$.

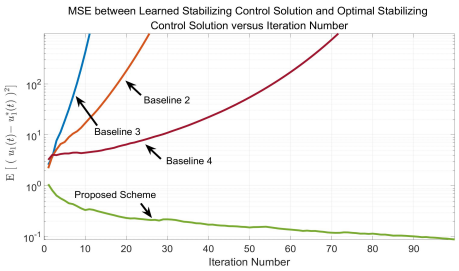

Fig. 3: MSE between the learned stabilizing control solution and the optimal stabilizing control solution versus iteration number. The system parameters are configured as follows: $\mathbf{A} = \begin{bmatrix} 1.37 & 0.44 & 0.15 \\ 0.13 & 0.82 & 0.36 \\ 0.41 & 0.57 & 0.36 \end{bmatrix}$ and control input matrix given by $\mathbf{B} = \begin{bmatrix} 1.61 & 0.67 \\ 0.74 & 0.52 \\ 1.02 & 0.56 \end{bmatrix}$. $p = 0.8$, $N_t = N_r = 2$, $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}_2$, $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{Q} = \mathbf{I}_3$, $\gamma = 10$, $\mathbf{W} = \mathbf{I}_3$, and $L = 2$.

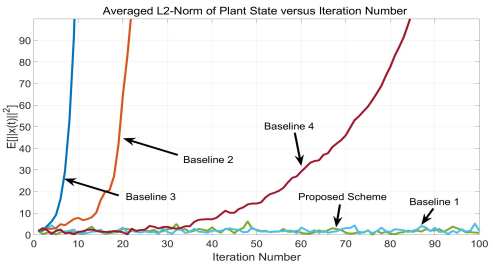

Fig. 4: Averaged L2-norm of plant state versus iteration number. The system parameters are configured as follows: $\mathbf{A} = \begin{bmatrix} 1.46 & 0.44 & 0.17 \\ 0.13 & 0.92 & 0.43 \\ 0.41 & 0.67 & 0.56 \end{bmatrix}$ and control input matrix given by $\mathbf{B} = \begin{bmatrix} 1.41 & 0.43 \\ 0.67 & 0.44 \\ 0.92 & 0.51 \end{bmatrix}$. $p = 0.8$, $N_t = N_r = 2$, $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}_2$, $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{Q} = \mathbf{I}_3$, $\gamma = 10$, $\mathbf{W} = \mathbf{I}_3$, and $L = 2$.

$\mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t), \mathbf{P} \in \mathbb{S}_+^S$. The stabilizing controller learns the control policy via brute-force value iteration using the least square.

- **Baseline 4:** *(Naive LQR Control [22])* The stabilizing controller applies the naive LQR control solution without awareness of the disturbance.

A. Complexity Analysis

We illustrate the CPU computational time with 10^6 runs versus the number of plant state dimensions S in Fig. 2. As revealed by the figure, the CPU computational time of the proposed scheme is substantially less than that of Baseline 2 and Baseline 3. This is because Baseline 2 requires update for a total number of $(N_t \times N_r \times 2)^2 \times L$ pseudo value functions with each pseudo value function containing the S-dimensional plant state. Although Baseline 3 neglects the CSI and only requires update for a pseudo value function w.r.t. the S-dimensional plant state at each timeslot, it solves the kernel value of the pseudo value function using least square with $\frac{S\times(S+1)}{2}$ neighboring plant state memory. On the contrary, the proposed scheme applies the SA update to learn a reduced-state value function w.r.t. the S-dimensional plant state without requiring plant state memory samples. The computational complexity based on the proposed scheme can be significantly reduced compared to Baseline 2 and Baseline 3. Note that Baseline 1 and Baseline 4 have prior knowledge of the stabilizing control policies, and hence the computational complexity for Baseline 1 and Baseline 4 is smaller than that of the other schemes.

B. Convergence Analysis

The MSE between the learned stabilizing control solution and the optimal stabilizing control solution versus the iteration number revealed in Fig. 3 shows the learning performance for the Nash equilibrium of Problem 1 via the proposed scheme and baseline schemes. As shown in the figure, the learned control solution deviates from the optimal solution via Baseline 2, Baseline 3 and Baseline 4, while our proposed scheme tracks the optimal solution asymptotically. As a result, the proposed scheme asymptotically track the Nash equilibrium of Problem 1. Specifically, the control solution via Baseline 2 deviates from the optimal control solution due to the discretization error when learning a tuple of pseudo value functions. The control solution via Baseline 3 also cannot track the optimal control solution since the learned control gains via Baseline 3 are static while the optimal control gains should adapt to the real-time CSI. The control solution via Baseline 4 also deviates from the optimal control solution since it applies the LQR control solution which is optimal if and only if the interference signal does not exist. The control solution via our proposed scheme, however, can asymptotically track the optimal CSI-adaptive control solution since the proposed scheme only requires learning of a reduced-state value function with a small state space efficiently without the "curse of dimensionality".

C. Robustness Analysis

The averaged L2-norm of the plant state versus iteration number illustrated in Fig. 4 indicates the robustness performance of the investigated stabilizing control schemes w.r.t. the external worst-case disturbance. As revealed by the figure, the plant state via Baseline 2, Baseline 3 and Baseline 4 grows exponentially fast and the system is unstable. On the contrary, the plant state via the proposed scheme maintains bounded over time. Specifically, Baseline 2 suffers from the "curse of dimensionality" and the learned control policies cannot stabilize the system. Baseline 3 applies the static

control policy which is independent of CSI, and hence the system via control policy in Baseline 3 is also unstable. Although Baseline 4 applies the CSI-adaptive control policy, it ignores the robustness to disturbance signals. As a result, the plant state via Baseline 4 also grows unbounded over time. On the contrary, Baseline 1 and the proposed scheme consider the optimal stabilizing control policies that are adaptive to CSI and robust to the external disturbance signals and hence, the plant state via Baseline 1 and the proposed scheme maintains bounded over time.

VI. Conclusion

In this paper, we considered a zero-sum game for an unstable linear system over wireless MIMO fading channels. We formulated the problem as a stochastic ergodic game and proposed the reduced-state Bellman optimality equations to address the "curse of dimensionality" w.r.t. the uncountable state space. Based on the equations, we analyzed the structural form of the Nash equilibrium, and further proposed a novel SA-based online learning algorithm for the Nash equilibrium. Numerical results were analyzed in terms of various aspects and we showed that the superior performance gains could be achieved by the proposed scheme compared to the baseline schemes.

Appendix

A. Proof of Theorem 2

Taking the expectation over the CSI $\{\delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)\}$ on both sides of (8), it follows the equivalent form given by

$$\theta_i^* + \widetilde{V}_i^*(\mathbf{x}(t)) = \mathbb{E}[\min_{\mathbf{u}_i(t)} \max_{\mathbf{u}_j(t)} [r_i(\mathbf{x}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)) + \mathbb{E}[\widetilde{V}_i^*(\mathbf{x}(t+1))|\mathbf{x}(t), \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t), \mathbf{u}_1(t), \mathbf{u}_2(t)]]]. \tag{22}$$

Following the similar approach as that in [22], we first assume that $\widetilde{V}_i^*(\mathbf{x}(t))$ has a quadratic form of $\mathbf{x}(t)$ and is given by $\widetilde{V}_i^*(\mathbf{x}(t)) = (-1)^{i+1}\mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t)$ with $\mathbf{P} \in \mathbb{S}_+^S$ being a constant positive definite matrix. Then, above form can be further represented as

$$\begin{aligned}
&\theta_i^* + (-1)^{i+1}\mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t) \\
&= \mathbb{E}[\min_{\mathbf{u}_i(t)} \max_{\mathbf{u}_j(t)} [(-1)^{i+1}[\mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t) + \mathbf{u}_1^T(t)\mathbf{R}_1\mathbf{u}_1(t) - \\
&\gamma^2\mathbf{u}_2^T(t)\mathbf{R}_2\mathbf{u}_2(t) + (\delta_1(t)\mathbf{B}\mathbf{H}_1(t)\mathbf{u}_1(t))^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) \\
&\mathbf{u}_1(t) - \gamma^2(\delta_2(t)\mathbf{B}\mathbf{H}_2(t)\mathbf{u}_2(t))^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t)\mathbf{u}_2(t) + (\mathbf{A} \\
&\mathbf{x}(t) + \delta_1(t)\mathbf{B}\mathbf{H}_1(t)\mathbf{u}_1(t) + \delta_2(t)\mathbf{B}\mathbf{H}_2(t)\mathbf{u}_2(t))^T\mathbf{P}(\mathbf{A} \\
&\mathbf{x}(t) + \delta_1(t)\mathbf{B}\mathbf{H}_1(t)\mathbf{u}_1(t) + \delta_2(t)\mathbf{B}\mathbf{H}_2(t)\mathbf{u}_2(t)) + \mathrm{Tr}(\mathbf{B}^T \\
&\mathbf{P}\mathbf{B} + \mathbf{B}\mathbf{W})]]].
\end{aligned} \tag{23}$$

It follows that the optimizer $\mathbf{u}_i^*(t), i \in \{1, 2\}$, is attainable at the equilibrium point of the Hamiltonian for (23). It follows that the optimizer $\mathbf{u}_1^*(t)$ and $\mathbf{u}_2^*(t)$ is given by $\mathbf{u}_1^*(t) = \mathbf{K}_1(\mathbf{P}, t)\mathbf{x}(t)$, $\mathbf{u}_2^*(t) = \mathbf{K}_2(\mathbf{P}, t)\mathbf{x}(t)$. $\theta_i^* = (-1)^{i+1}\,\mathrm{Tr}(\mathbf{P}\mathbf{W} + \mathbf{B}^T\mathbf{P}\mathbf{B})$ and $\widetilde{V}_i^*(\mathbf{x}(t)) = (-1)^{i+1}\mathbf{x}(t)^T\mathbf{P}\mathbf{x}(t) = (-1)^{i+1}\mathbf{x}(t)\mathbb{E}[g(\mathbf{P})]\mathbf{x}(t)$. This concludes the proof.

References

[1] J. Moon, T. E. Duncan, and T. Başar, "Risk-sensitive zero-sum differential games," *IEEE Trans. Autom. Control*, vol. 64, no. 4, pp. 1503–1518, 2018.

[2] L. Li and J. S. Shamma, "Efficient strategy computation in zero-sum asymmetric information repeated games," *IEEE Trans. Autom. Control*, vol. 65, no. 7, pp. 2785–2800, 2019.

[3] T. Başar and P. Bernhard, *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

[4] Q. Wei, D. Liu, Q. Lin, and R. Song, "Adaptive dynamic programming for discrete-time zero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 957–969, 2017.

[5] Y. Fu, J. Fu, and T. Chai, "Robust adaptive dynamic programming of two-player zero-sum games for continuous-time linear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3314–3319, 2015.

[6] B. Luo, Y. Yang, and D. Liu, "Policy iteration Q-learning for data-based two-player zero-sum game of linear discrete-time systems," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3630–3640, 2020.

[7] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[8] S. A. A. Rizvi and Z. Lin, "Output feedback Q-learning for discrete-time linear zero-sum games with application to the H-infinity control," *Automatica*, vol. 95, pp. 213–221, 2018.

[9] H. Li, D. Liu, and D. Wang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 3, pp. 706–714, 2014.

[10] J. Moon, "A sufficient condition for linear-quadratic stochastic zero-sum differential games for Markov jump systems," *IEEE Trans. Autom. Control*, vol. 64, no. 4, pp. 1619–1626, 2018.

[11] J. Song, S. He, Z. Ding, and F. Liu, "A new iterative algorithm for solving H-infinty control problem of continuous-time Markovian jumping linear systems based on online implementation," *Int. J. Robust Nonlinear Control*, vol. 26, no. 17, pp. 3737–3754, 2016.

[12] B. Gravell, K. Ganapathy, and T. Summers, "Policy iteration for linear quadratic games with stochastic parameters," *IEEE Contr. Syst. Lett.*, vol. 5, no. 1, pp. 307–312, 2020.

[13] H. Xu, S. Jagannathan, and F. Lewis, "Stochastic optimal design for unknown linear discrete-time system zero-sum games in input-output form under communication constraints," *Asian J. Control*, vol. 16, no. 5, pp. 1263–1276, 2014.

[14] C. Wu, X. Li, W. Pan, J. Liu, and L. Wu, "Zero-sum game-based optimal secure control under actuator attacks," *IEEE Trans. Autom. Control*, vol. 66, no. 8, pp. 3773–3780, 2020.

[15] C. Wu, W. Yao, W. Pan, G. Sun, J. Liu, and L. Wu, "Secure control for cyber-physical systems under malicious attacks," *IEEE Trans. Control Netw.*, vol. 9, no. 2, pp. 775–788, 2021.

[16] A.-G. Wu, H.-J. Sun, and Y. Zhang, "A novel iterative algorithm for solving coupled Riccati equations," *Appl. Math. Comput.*, vol. 364, p. 124645, 2020.

[17] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[18] J. F. Nash Jr, "The bargaining problem," *Econometrica: Journal of the econometric society*, pp. 155–162, 1950.

[19] D. P. Bertsekas *et al.*, "Dynamic programming and optimal control 3rd edition, volume ii," *Belmont, MA: Athena Scientific*, 2011.

[20] T. T. Doan, "Nonlinear two-time-scale stochastic approximation convergence and finite-time performance," *IEEE Trans. Autom. Control*, 2022, early access.

[21] B. Zheng, C. You, W. Mei, and R. Zhang, "A survey on channel estimation and practical passive beamforming design for intelligent reflecting surface aided wireless communications," *IEEE Commun. Surv. Tutor.*, vol. 24, no. 2, pp. 1035–1071, 2022.

[22] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.