Online Non-Cooperative Zero-Sum Games for Linear Systems over Wireless MIMO Fading Channels with Uncountable State Space

Minjie Tang, *Member, IEEE*, and Vincent K. N. Lau, *Fellow, IEEE*

***Abstract*—In this paper, we explore a non-cooperative zero-sum game for a linear dynamic system, operating over wireless Multiple-Input Multiple-Output (MIMO) fading channels that feature an uncountable channel state space between the remote controllers and the actuator of the dynamic plant. We initiate our work by framing the stochastic zero-sum game as an ergodic optimization problem, from which we derive the structural properties of the coupled optimality equations. To overcome the "curse of dimensionality," we develop equivalent structured reduced-state optimality equations. Leveraging these equations, we derive the necessary and sufficient conditions for the existence of a Nash equilibrium in the stochastic game. Furthermore, we propose an online structured learning algorithm based on Stochastic Approximation (SA) to compute the Nash equilibrium. Employing Ordinary Differential Equation (ODE) techniques and Lyapunov stability analysis, we demonstrate the asymptotic optimality of our learning algorithm. Numerical results show that the proposed method outperforms existing state-of-the-art baselines.**

***Index Terms*—Zero-sum games, stochastic games, optimal control, structured online learning, stochastic approximation, Lyapunov stability analysis.**

I. Introduction

A. Background

WIRELESS control has become increasingly important in modern networked systems due to its advantages in scalability, flexibility, and cost efficiency. It enables remote actuation and coordination among distributed sensors and controllers, with wide applications in industrial automation, autonomous systems, and smart infrastructure [1], [2]. However, wireless communication channels are inherently unreliable. They suffer from impairments such as *fading*, referring to random fluctuations in signal strength caused by multipath propagation or mobility, and *packet dropouts*, which arise from interference, congestion, or noise [3]. These stochastic phenomena pose significant challenges to maintaining the stability of closed-loop control systems [4]. Moreover, wireless links are vulnerable to adversarial interference and attacks, which can further degrade control performance [5].

In such environments, conventional control methods such as Proportional-Integral-Derivative (PID) control [6] and Linear Quadratic Regulation (LQR) [7] are often suboptimal, as they are typically developed for cooperative control agents and rely on assumptions of ideal communication channels, static system dynamics, and the absence of adversarial disturbances. In the presence of time-varying wireless impairments or malicious disruptions, such static control policies are incapable of maintaining system stability.

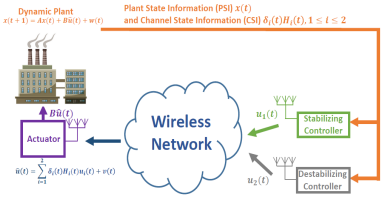


Fig. 1: A typical architecture of a non-cooperative zero-sum game for a linear system over the wireless network.

To address these challenges, we adopt a game-theoretic framework formulated as a zero-sum stochastic setting. Specifically, we consider a linear system architecture where two remote controllers interact with a dynamic plant through a shared actuator connected via wireless fading channels, as illustrated in Fig. 1. One controller is tasked with stabilizing the plant, while the other acts adversarially, aiming to destabilize it by injecting disruptive control signals. Both controllers transmit their control inputs wirelessly to the actuator, which then applies the combined signal directly to the plant. This architecture effectively captures the adversarial nature and uncertainty inherent in wireless control systems operating under unreliable or potentially malicious communication conditions.

Given the time-varying and unpredictable nature of wireless channels, the stabilizing controller must continuously adapt to real-time transmission conditions. In this context, online learning-based approaches, particularly those relying on reinforcement learning [8], provide a promising solution by enabling the controller to learn effective strategies directly from empirical channel realizations. However, applying these methods in wireless control systems remains highly challenging due to the complex and inherently stochastic nature of wireless environments. These systems typically involve uncountable state spaces and dynamic interference patterns, significantly hindering the scalability and real-time applicability of policy learning algorithms.

B. Related Works

Zero-sum games for linear systems over static channels have been extensively investigated in the existing literature [9]–[14].

Minjie Tang is with the Communication Systems Department, EURECOM, France (e-mail: Minjie.Tang@eurecom.fr).

Vincent K. N. Lau is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: eeknlau@ust.hk).

Corresponding Author: Vincent K. N. Lau.

For instance, recursive solutions such as value-iteration-based and policy-iteration-based algorithms over value functions were proposed in [9], [10] to address linear zero-sum games. Q-learning-based approaches for the Nash equilibrium of these games were introduced in [11] and [12]. In [13] and [14], stochastic zero-sum differential games were considered and the integral reinforcement learning algorithms derived from policy iteration were utilized to learn the Nash equilibrium. Note that the above works [9]–[14] all assume static channels between the actuator and the remote controllers. Brute-force application of these methods to games over wireless networks can lead to learned control policies that deviate from the Nash equilibrium. This deviation arises because such policies are typically static and fail to adapt to the stochastic and time-varying transmission opportunities inherent in wireless environments.

Recent research has focused on zero-sum games within linear systems that operate over wireless communication channels [15]–[20]. These studies incorporate Channel State Information (CSI) into the solution design to accommodate the randomness of channel fading. Specifically, in [20], the authors investigated on-off channels between the actuator and the remote controllers. The Nash equilibrium was achieved using a value-iteration-based approach. However, a fixed packet-loss pattern was assumed in the work, and extensions of the scheme to time-varying packet-loss channels pose challenges. In [18], the authors considered a linear stochastic zero-sum game with time-varying packet-loss channels. The control policy was learned by aggregating plant states at neighboring active slots, and can converge to the Nash equilibrium. However, this approach requires finite burst packet dropout. In [19], the Independent and Identically Distributed (i.i.d.) on-off channels were considered, and the authors derived the Nash equilibrium that can stabilize the system. However, it remains unclear how to practically learn the equilibrium point. Additionally, the modeling approach using a 0-1 process in [19] may over-simplify practical wireless communication channels. [15] and [16] tackled stochastic games over wireless fading channels with finite state space. The structural properties of the Nash equilibrium were developed by a set of coupled Bellman optimality equations. However, the assumption of a finite CSI state space is often unrealistic. Extending these methods to general fading channels with uncountably infinite CSI states encounters the "curse of dimensionality." To manage this, one may consider discretization of the CSI state space into a finite set. However, such an approach introduces errors that leads to the deviation of the control policy from the Nash equilibrium. [17] investigated a linear zero-sum game over non-zero-mean Gaussian fading channels, using CSI statistics to parameterize control strategies and learning them through policy iteration. However, brute-force applications of the approach in [17] over the zero-mean Gaussian fading channels will lead to instability of the dynamic systems. This occurs because the control solution derived from [17] fails to effectively account for instantaneous transmission opportunities arising from the time-varying wireless environment. Consequently, the system becomes uncontrollable under zero-mean Gaussian fading channels.

C. Motivations, Contributions, and Notations

In this work, we propose a novel structured online model-based learning algorithm to compute the Nash equilibrium of the stochastic game over the wireless network. The algorithm learns the parameters of the structured optimality condition via Stochastic Approximation (SA), leveraging both the known system parameters and real-time realizations of the wireless channel state. The key contributions of this work can be summarized as follows:

- **Structured Bellman Optimality Equations over Uncountable CSI State Space.** The coupling among control agents, system parameters, and random channel states is highly nonlinear, posing significant challenges for analyzing the Bellman optimality structure. We address this by first studying the finite-horizon stochastic game and, leveraging the monotonicity and contractivity of the nonlinear Bellman operator, derive structured Bellman equations through an asymptotic analysis tailored to our setting.
- **Structured Reduced-State Bellman Optimality Equations.** By exploiting the statistical independence between the fading process and the plant state, we derive reduced-state equations for computing the Nash equilibrium in the ergodic zero-sum game. Despite the uncountable CSI space, these equations involve only a single unknown, effectively mitigating the "curse of dimensionality."
- **Existence Conditions of the Nash Equilibrium.** Establishing equilibrium existence under uncountable CSI is challenging due to potential uncontrollability and time-varying non-cooperation. We address this by applying a positive semi-definite cone decomposition to the structured equations and derive explicit conditions based on system dynamics, channel statistics, and controller activation probabilities for the existence of a Nash equilibrium.
- **Online Model-based Structured Learning via Reduced-State Equations.** Standard SA algorithms require estimating full value functions, which is challenging in continuous state spaces. Leveraging the reduced-state structure, we propose an SA-based online learning algorithm that learns only the structured function kernel from system parameters and real-time CSI. To address the difficulty of constructing a Lyapunov function for convergence analysis, we adopt a non-standard Ordinary Differential Equation (ODE) approach where convergence is established by analyzing a virtual fixed-point iteration that approximates the associated ODE trajectory.

A preliminary version of this work appeared in IEEE CDC 2023 [21], focusing on algorithmic implementation. This paper significantly extends it by analyzing the Nash equilibrium, proposing efficient algorithmic techniques, and establishing convergence guarantees. Extensive simulations also show its superiority over state-of-the-art baselines.

Notation: Bold uppercase and lowercase letters denote matrices and vectors. The operators $(\cdot)^T$ and $\mathrm{Tr}(\cdot)$ denote the transpose and trace. $\mathbf{0}_{m\times n}$ and $\mathbf{0}_m$ are $m \times n$ and $m \times m$ zero matrices. $\mathbf{1}_S$ is the $S \times S$ identity matrix. $\mathrm{Diag}(a, b, \ldots)$

denotes a diagonal matrix with entries $a, b, \ldots$. $\mathbb{R}^{m\times n}$, $\mathbb{S}_+^m$, $\mathbb{S}^m$, $\mathbb{Z}_+$, and $\mathbb{R}_+$ denote the sets of $m \times n$ real matrices, $m\times m$ positive definite matrices, $m\times m$ positive semi-definite matrices, nonnegative integers, and positive real numbers. $\|\mathbf{A}\|$ and $\|\mathbf{a}\|$ are the spectral norm of matrix $\mathbf{A}$ and the Euclidean norm of vector $\mathbf{a}$. $[\mathbf{A}]_i$ denotes the ith principal submatrix and $[\mathbf{A}]_{a:b,c:d}$ is the submatrix of $\mathbf{A}$ from rows a to b and columns c to d. $\sigma_{\min}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the minimum singular value and eigenvalue of $\mathbf{A}$.

II. System Model

A. Dynamic Plant

We consider a time-slotted system with S-dimensional plant state $\mathbf{x}(t) \in \mathbb{R}^{S\times 1}$ controlled by two remote non-cooperative control players. The i-th remote controller ($i \in \{1, 2\}$) is equipped with $N_{i,t} \in \mathbb{Z}_+$ transmission antennas and the actuator is equipped with $N_r \in \mathbb{Z}_+$ receiving antennas. The plant system is governed by the first-order coupled linear difference equations, as follows[1].

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

where the initial plant state $\mathbf{x}(1) \in \mathbb{R}^{S\times 1}$ is randomly generated, with each element of $\mathbf{x}(1)$ sampled independently from a Gaussian distribution with zero mean and unit variance. $\widehat{\mathbf{u}}(t) \in \mathbb{R}^{N_r\times 1}$ represents the received noisy control signal at the actuator. $\mathbf{A} \in \mathbb{R}^{S\times S}$ is the system dynamics matrix, which characterizes the internal evolution of the dynamic plant, and $\mathbf{B} \in \mathbb{R}^{S\times N_r}$ is the actuator matrix. The additive plant noise $\mathbf{w}(t) \in \mathbb{R}^{S\times 1}$ follows a Gaussian distribution with zero mean and a finite covariance matrix $\mathbf{W} \in \mathbb{S}^S$.

B. Wireless MIMO Fading Channel Model

We model the communication channels between the actuator and the i-th remote controller as $N_r \times N_{i,t}$ wireless Multiple-Input Multiple-Output (MIMO) fading channels, where $N_r \in \mathbb{Z}_+$ denotes the number of receiving antennas at the actuator. The active controllers transmit control signals $\mathbf{u}_i(t) \in \mathbb{R}^{N_{i,t}\times 1}$ to the actuator through these wireless communication channels. The received signal $\widehat{\mathbf{u}}(t) \in \mathbb{R}^{N_r\times 1}$ at the actuator is given by

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

where $\mathbf{H}_i(t) \in \mathbb{R}^{N_r\times N_{i,t}}$ represents the wireless MIMO fading coefficients between the i-th remote controller and the actuator. The term $\mathbf{v}(t) \sim \mathcal{N}(\mathbf{0}_{S\times 1}, \mathbf{1}_{N_r})$ denotes the additive channel noise at the actuator. The random variable $\delta_i(t) \in \{0, 1\}$ models the random access activity of the i-th remote controller. It is i.i.d. across timeslots and remote controllers, with $\Pr(\delta_i(t) = 1) = p_i \in [0, 1]$.

Assumption 1 (Wireless MIMO Fading Channel Model [3])**.** *The coefficient of wireless MIMO fading channels $\mathbf{H}_i(t)$ between the i-th controller and the actuator remains quasi-static within each timeslot and each controller, and is i.i.d. across remote controllers and the timeslots. Moreover, each element of $\mathbf{H}_i(t)$ follows a Gaussian distribution with zero mean and unit variance.*

[1]We assume the control signal $\widehat{\mathbf{u}}(t)$ is generated and applied within the same timeslot, with negligible delay.

Assumption 1 on the fading coefficient $\mathbf{H}_i(t)$ reflects practical wireless systems. In reality, $\mathbf{H}_i(t)$ exhibits temporal correlation governed by the coherence time, within which it is highly correlated and becomes independent beyond. The i.i.d. assumption aligns with standard practice that approximates the timeslot duration to the coherence time [3]. Moreover, $\mathbf{H}_i(t)$ captures signal reinforcement and cancellation from scattering; with many scatterers, the central limit theorem justifies modeling $\mathbf{H}_i(t)$ as Gaussian. Finally, since the adversarial controller is typically distant or hidden from the stabilizer [22], $\mathbf{H}_1(t)$ and $\mathbf{H}_2(t)$ likely experience distinct scattering environments and are naturally modeled as i.i.d..

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

The dynamic system under wireless communication channels is linear and time-varying, with equivalent plant dynamics obtained by substituting (2) into (1), yielding:

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

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

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

where $\mathbf{Q} \in \mathbb{S}_+^S$, $\mathbf{R}_i \in \mathbb{S}_+^{N_{i,t}}$, and $\mathbf{M}_i \in \mathbb{S}_+^S$ are weighting matrices for the plant state cost, control cost, and actuation cost, respectively. $\gamma > 0$ is a positive constant penalizing the non-cooperation between the controllers. Furthermore, it satisfies $\gamma > \gamma^* > 0$, where γ^* is a critical threshold [11], [12], [23] representing the minimum value for which the zero-sum game is solvable. The explicit derivation of γ^* is provided in Section III.

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)**.**

Stabilizing Controller (Controller 1):

$$\begin{aligned} &\min_{\pi_1}\max_{\pi_2} \mathcal{J}^{\pi_1,\pi_2} \\ &= \min_{\pi_1}\max_{\pi_2} \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}^{\pi_1,\pi_2}[r(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t))] \end{aligned} \quad (5)$$

s.t. Dynamics (3).

Destabilizing Controller (Controller 2):

$$\max_{\pi_2}\min_{\pi_1}\mathcal{J}^{\pi_2,\pi_1} = \max_{\pi_2}\min_{\pi_1}\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}^{\pi_1,\pi_2}[r(\mathbf{S}(t),\mathbf{u}_1(t),\mathbf{u}_2(t))] \quad \text{s.t. Dynamics (3).} \tag{6}$$

The expectations in (5) and (6) are taken With Respect To (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 $\mathbf{v}(t)$.

The optimal solution to Problem 1 is referred to as the *Nash equilibrium* of Problem 1, as defined in the following sense.

Definition 1 (Nash Equilibrium [24])**.** *The control policies of the remote controllers, $\{\pi_1^*,\pi_2^*\}$, are said to constitute the Nash equilibrium of Problem 1 if*

$$\mathcal{J}^{\pi_1^*,\pi_2} \le \mathcal{J}^{\pi_1^*,\pi_2^*} \le \mathcal{J}^{\pi_1,\pi_2^*}. \tag{7}$$

According to Definition 1, the optimal policy π_1^* minimizes the state cost $\mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t)$, while π_2^* maximizes it. Thus, Controller 1 serves as the stabilizer, and Controller 2 acts as the destabilizer.

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

A. Structured Reduced-State Bellman Optimality Equations

Traditionally, solving Problem 1 to obtain the Nash equilibrium involves addressing the Bellman optimality equations [9]–[14], [23]. Instead of tackling these complex black-box equations, our goal is to derive structured Bellman optimality equations that reduce the unknowns and exploit problem-specific structure. However, this derivation is highly nontrivial due to the stochastic wireless setting: the channel state space is uncountable due to continuous fading and random access; the control signals and CSI are stochastically coupled, breaking separability; and the resulting Bellman operator is non-contractive, rendering standard dynamic programming tools inapplicable.

To address these challenges, we analyze a finite-horizon stochastic game and construct an induced Bellman operator that captures the expected value update under CSI randomness. By leveraging monotonicity and a cone-based contractivity argument under sufficient conditions (cf. Theorem 3), we establish asymptotic convergence of the operator. These steps lead to the structured Bellman optimality equations for Problem 1, formally stated in the following theorem.

Theorem 1 (Structured Bellman Optimality Equations for Problem 1)**.** *If the Nash equilibrium of Problem 1 exists under the sufficient conditions stated in Theorem 3, it can be obtained by solving a pair of structured Bellman optimality equations associated with Problem 1, given as follows:*

$$\theta_1^* + V_1^*(\mathbf{S}(t)) = \min_{\mathbf{u}_1(t)}\max_{\mathbf{u}_2(t)}\Big[r(\mathbf{S}(t),\{\mathbf{u}_i(t)\}_{i=1,2}) + \mathbb{E}\left[V_1^*(\mathbf{S}(t+1))\,|\mathbf{x}(t),\{\delta_i(t)\mathbf{H}_i(t),\mathbf{u}_i(t)\}_{i=1,2}\right]\Big], \tag{8}$$

$$\theta_2^* + V_2^*(\mathbf{S}(t)) = \max_{\mathbf{u}_2(t)}\min_{\mathbf{u}_1(t)}\Big[r(\mathbf{S}(t),\{\mathbf{u}_i(t)\}_{i=1,2}) + \mathbb{E}\left[V_2^*(\mathbf{S}(t+1))\,|\mathbf{x}(t),\{\delta_i(t)\mathbf{H}_i(t),\mathbf{u}_i(t)\}_{i=1,2}\right]\Big], \tag{9}$$

where

- $V_i^*(\mathbf{S}(t)) = \mathbf{x}^T(t)\bar{\mathbf{P}}(\{\delta_i(t)\mathbf{H}_i(t)\}_{i=1,2})\,\mathbf{x}(t)$ *is the optimal value function ($i \in \{1,2\}$), and $\bar{\mathbf{P}}(\{\delta_i(t)\mathbf{H}_i(t)\}_{i=1,2}) \in \mathbb{S}_+^S$ is a continuous function of the CSI $\{\delta_i(t)\mathbf{H}_i(t)\}_{i=1,2}$;*
- $\theta_i^* = \mathcal{J}^{\pi_1^*,\pi_2^*} = \mathrm{Tr}(\mathbf{B}^T\mathbf{P}\mathbf{B} + \mathbf{P}\mathbf{W})$ *is the optimal averaged cost of Problem 1, where $\mathbf{P} = \mathbb{E}[\bar{\mathbf{P}}(\{\delta_i(t)\mathbf{H}_i(t)\}_{i=1}^2)] \in \mathbb{S}_+^S$;*
- *The Nash equilibrium of Problem 1 is denoted as $\{\pi_i^*\}_{i=1,2} = \{\mathbf{u}_i^*(t)\}_{i=1,2}$, and $\{\mathbf{u}_i^*(t)\}_{i=1,2}$ corresponds to the optimal solutions to both (8) and (9).*

Proof: Please refer to Appendix A. ∎

When the Nash equilibrium of Problem 1 exists, one may consider classical iterative methods such as value iteration [9], [11], [12], [25] to solve the Bellman equations (8) and (9). However, direct application is challenging due to the curse of dimensionality induced by the uncountable CSI space $\{\delta_i(t)\mathbf{H}_i(t)\}_{i=1,2}$. Specifically, value iteration attempts to learn the value function $V_i^*(\mathbf{S}(t))$, which depends on the unknown matrix $\bar{\mathbf{P}}(\cdot)$ over a continuous domain. This results in an intractable number of unknowns. To address this, we exploit the statistical independence between the plant state $\mathbf{x}(t)$ and the CSI to derive equivalent structured reduced-state Bellman equations.

Theorem 2 (Structured Reduced-State Bellman Optimality Equations for Problem 1 [21])**.** *If the Nash equilibrium of Problem 1 exists under the sufficient conditions stated in Theorem 3, it can be derived from the solution of a pair of equivalent, structured, reduced-state Bellman optimality equations, given as follows:*

$$\theta_1^* + \tilde{V}_1^*(\mathbf{x}(t)) = \mathbb{E}\left[\min_{\mathbf{u}_1(t)}\max_{\mathbf{u}_2(t)}\Big[r(\mathbf{S}(t),\{\mathbf{u}_i(t)\}_{i=1,2}) + \mathbb{E}\left[\tilde{V}_1^*(\mathbf{x}(t+1))\,|\mathbf{x}(t),\{\delta_i(t)\mathbf{H}_i(t),\mathbf{u}_i(t)\}_{i=1,2}\right]\Big]\,\middle|\mathbf{x}(t),\{\mathbf{u}_i(t)\}_{i=1,2}\right], \tag{10}$$

$$\theta_2^* + \tilde{V}_2^*(\mathbf{x}(t)) = \mathbb{E}\left[\max_{\mathbf{u}_2(t)}\min_{\mathbf{u}_1(t)}\Big[r(\mathbf{S}(t),\{\mathbf{u}_i(t)\}_{i=1,2}) + \mathbb{E}\left[\tilde{V}_2^*(\mathbf{x}(t+1))\,|\mathbf{x}(t),\{\delta_i(t)\mathbf{H}_i(t),\mathbf{u}_i(t)\}_{i=1,2}\right]\Big]\,\middle|\mathbf{x}(t),\{\mathbf{u}_i(t)\}_{i=1,2}\right], \tag{11}$$

where

- $\tilde{V}_i^*(\mathbf{x}(t)) = \mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t)$ *represents the optimal reduced-state value function for $i \in \{1,2\}$, parameterized by a single kernel $\mathbf{P} \in \mathbb{S}_+^S$;*
- *The Nash equilibrium for Problem 1 is expressed as: $\{\pi_i^*\}_{i=1,2} = \{\mathbf{u}_1^*(t),\mathbf{u}_2^*(t),\forall t \in \mathbb{Z}_+\}$, where $\mathbf{u}_i^*(t) = \mathbf{K}_i(\mathbf{P},t)\mathbf{x}(t)$ represents the optimal solution to both (10) and (11), and the optimal feedback control gain $\mathbf{K}_i(\mathbf{P},t) \in \mathbb{R}^{N_{i,t}\times S}$ is given by*

$$\mathbf{K}_1(\mathbf{P},t) = -\Big(\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_1(t)\mathbf{B}\mathbf{H}_1(t)\Big)^{-1}\delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_1(t)\mathbf{A}, \tag{12}$$

and

$$\mathbf{K}_2(\mathbf{P},t) = (\gamma^2\delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t) + \gamma^2\mathbf{R}_2 - \delta_2(t)\mathbf{H}_2(t)\mathbf{B}^T\widetilde{\mathbf{P}}_2(t)\mathbf{B}\mathbf{H}_2(t))^{-1}\delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_2(t)\mathbf{A}, \tag{13}$$

where

$$\widetilde{\mathbf{P}}_1(t) = (\mathbf{P}^{-1} - \gamma^{-2}\delta_2(t)\mathbf{B}\mathbf{H}_2(t)\left(\mathbf{R}_2 + \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t)\right)^{-1}\mathbf{H}_2^T(t)\mathbf{B}^T)^{-1}, \tag{14}$$

and

$$\widetilde{\mathbf{P}}_2(t) = (\mathbf{P}^{-1} - \delta_1(t)\mathbf{B}\mathbf{H}_1(t)\left(\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t)\right)^{-1}\mathbf{H}_1^T(t)\mathbf{B}^T)^{-1}. \tag{15}$$

Compared to solving the Bellman equations in (8) and (9), which require learning the full value function $V_i^*(\mathbf{S}(t))$ over an uncountable CSI space, the reduced-state equations in (10) and (11) involve only $\tilde{V}_i^*(\mathbf{x}(t))$ with a single unknown $\mathbf{P}$. This alleviates the curse of dimensionality and makes learning the Nash equilibrium feasible. Section IV presents an online structured learning algorithm based on these equations. Moreover, the statistical independence between the CSI $\delta_i(t)\mathbf{H}_i(t)$ and the plant state $\mathbf{x}(t)$, required by Theorem 2, holds in many practical scenarios, such as nonlinear systems under Rayleigh fading. This broadens the applicability of our state-reduction technique beyond the linear Gaussian block-fading model considered in this work.

In the following, we discuss the structural form of the Nash equilibrium under several special cases based on Theorem 2.

1) Asymptotic Structural Form of the Nash Equilibrium: Note that a large $\gamma \in \mathbb{R}_+$ in Problem 1 imposes a strong penalty on the destabilizing controller (Controller 2), allowing the stabilizing controller (Controller 1) to dominate the system. The following corollary characterizes the simplified structure of the Nash equilibrium as $\gamma \to \infty$.

Corollary 1 (Asymptotic Structural Form of the Nash Equilibrium)**.** *If the Nash equilibrium* $\{\pi_1^*, \pi_2^*\} = \{\mathbf{u}_1^*(t), \mathbf{u}_2^*(t), \forall t \in \mathbb{Z}_+\}$ *of Problem 1 exists when* $\gamma \to \infty$*, then the optimal control solution for the remote controllers is given by* $\mathbf{u}_1^*(t) = \mathbf{K}_1(\mathbf{P},t)\mathbf{x}(t)$ *and* $\mathbf{u}_2^*(t) = \mathbf{0}_{N_{2,t}\times 1}$*, where*

$$\mathbf{K}_1(\mathbf{P},t) = -\left(\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_1(t)\right)^{-1}\delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{A}. \tag{16}$$

Proof: Please refer to Appendix B. ∎

Corollary 1 reveals a structural limit of the two-player zero-sum game as $\gamma \to \infty$. While the resulting control resembles an LQR-type solution, it generalizes classical LQR to systems with fading channels and uncountable CSI states. This result is rigorously derived from our Bellman framework and reflects a limiting equilibrium of the two-player game, rather than eliminating the adversary.

2) Homogeneous Structural Form of the Nash Equilibrium: We are also interested in the simplified structural form of the Nash equilibrium when remote controllers are homogeneous in the sense that $\delta_1(t) = \delta_2(t) = 1, \mathbf{H}_1(t) = \mathbf{H}_2(t) = \mathbf{H}(t), \mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}$, $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{M}$. This is summarized in the following corollary.

Corollary 2 (Homogeneous Structural Form of the Optimal Control Solution)**.** *If the Nash equilibrium* $\{\pi_1^*, \pi_2^*\} = \{\mathbf{u}_1^*(t), \mathbf{u}_2^*(t), \forall t \in \mathbb{Z}_+\}$ *of Problem 1 exists when remote controllers are homogeneous, the optimal control solution for remote controllers is given by* $\mathbf{u}_1^*(t) = \mathbf{K}_1(\mathbf{P},t)\mathbf{x}(t)$*,* $\mathbf{u}_2^*(t) = \mathbf{K}_2(\mathbf{P},t)\mathbf{x}(t)$*, where*

$$\mathbf{K}_1(\mathbf{P},t) = -(\mathbf{R} + \mathbf{H}^T(t)\mathbf{B}^T(t)\mathbf{M}\mathbf{B}\mathbf{H}(t) + \mathbf{H}^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_1(t)\mathbf{B}\mathbf{H}(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_1(t)\mathbf{A}, \tag{17}$$

$$\mathbf{K}_2(\mathbf{P},t) = (\gamma^2\mathbf{R} + \gamma^2\mathbf{H}^T(t)\mathbf{B}^T\mathbf{M}\mathbf{B}\mathbf{H}(t) - \mathbf{H}^T\mathbf{B}^T\widetilde{\mathbf{P}}_2(t)\mathbf{B}\mathbf{H}(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_2(t)\mathbf{A}, \tag{18}$$

and

$$\widetilde{\mathbf{P}}_i(t) = \left(\mathbf{P}^{-1} - \gamma^{-4+2i}\mathbf{B}\mathbf{H}(t)(\mathbf{R} + \mathbf{H}^T(t)\mathbf{B}^T\mathbf{M}\mathbf{B}\mathbf{H}(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T\right)^{-1}, i \in \{1,2\}. \tag{19}$$

Proof: Please refer to Appendix B. ∎

B. Existence of the Nash Equilibrium of Problem 1

The existence of the Nash equilibrium in Problem 1 requires the structured reduced-state Bellman equations in (10) and (11) to admit well-defined solutions for all $t \in \mathbb{Z}_+$. This requires the existence of $V_i^*(\mathbf{S}(t))$ and θ_i^*, along with strict convexity in $\mathbf{u}_1(t)$ and strict concavity in $\mathbf{u}_2(t)$ for the respective optimization problems.

These conditions are easier to verify under static channels with fixed $\delta_i(t)\mathbf{H}_i(t)$, by ensuring the standard controllability assumption and imposing a lower bound on the penalty parameter γ [11], [12], [23]. However, in stochastic wireless settings, random access and fading can render $(\mathbf{A}, \delta_1(t)\mathbf{B}\mathbf{H}_1(t))$ uncontrollable at times. Unlike the static case where a large γ ensures existence, it is nontrivial to check whether a given penalty γ guarantees feasibility under channel randomness. To address this, we derive a sufficient condition by analyzing the reduced-state Bellman equations via a positive semi-definite cone decomposition, leading to a closed-form expression for the existence of the Nash equilibrium.

Theorem 3 (Sufficient Conditions for the Existence of the Nash Equilibrium of Problem 1)**.** *Let* $\mathcal{H}(t) \triangleq \delta_1(t)\mathbf{B}\mathbf{H}_1(t)\left(\mathbf{R}_1 + \mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t)\right)^{-1}\mathbf{H}_1^T(t)\mathbf{B}^T - \gamma^{-2}\delta_2(t)\mathbf{B}\mathbf{H}_2(t)\left(\mathbf{R}_2 + \mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t)\right)^{-1}\mathbf{H}_2^T(t)\mathbf{B}^T$*. The Nash equilibrium of Problem 1 exists if the following conditions are satisfied:*

- *(a) Let the eigenvalue decomposition of* $\mathcal{H}(t)$ *be* $\mathcal{H}(t) = \mathbf{V}^T(t)\zeta(t)\mathbf{V}(t)$ *with the diagonal elements of* $\zeta(t)$ *in descending order. Let* $\Pi(t) = \mathrm{Diag}(1,...1,0,...0) \in \mathbb{S}_+^S$ *satisfy* $\mathrm{Tr}(\Pi(t)) = \gamma(t)$ *where* $\gamma(t) =$ *Rank*$(\zeta(t))$*, then* $\|\mathbb{E}[(\mathbf{A}^T(\mathbf{I} - \Pi(t))\mathbf{V}(t))^T\mathbf{A}^T(\mathbf{I} - \Pi(t))\mathbf{V}(t)]\| < 1$*;*
- *(b)* $\mathbb{E}[\lambda_{min}((\zeta(t))_{\gamma(t)})] \geq 0$*;*
- *(c)* $\gamma^2 \geq (\|\mathbf{Q}\|\|\mathbf{B}^T\mathbf{B}\|\lambda_{min}^{-1}(\mathbf{M}_2) + \|\mathbf{A}\|^2\mathbb{E}[\mathrm{Tr}((\zeta(t))_{\gamma(t)}^{-1})]\|\mathbf{B}^T\mathbf{B}\|\lambda_{min}^{-1}(\mathbf{M}_2))(1 - \|\mathbb{E}[(\mathbf{A}^T(\mathbf{I} - \Pi(t))\mathbf{V}(t))^T\mathbf{A}^T(\mathbf{I} - \Pi(t))\mathbf{V}(t)]\|)^{-1}$.

Proof: Please refer to Appendix C. ∎

The conditions in Theorem 3 depend explicitly on system parameters $(\mathbf{A}, \mathbf{B})$ and wireless channel statistics, including the distributions of $\mathbf{H}_i(t)$ and activation probabilities p_i. They are both *computationally feasible* through closed-form expressions and *physically interpretable*. Specifically, a less unstable plant (smaller $\|\mathbf{A}\|$), higher p_1, lower p_2, and more antennas at the stabilizer (larger $N_{1,t}$, smaller $N_{2,t}$) all improve feasibility. These factors reduce the spectral norm in Condition (a), increase the minimal eigenvalue in Condition (b), or lower the trace in Condition (c). Feasibility also increases with γ, as stronger penalization of the adversary makes all three conditions easier to meet. This is consistent with classical results in zero-sum stochastic control (e.g., [26]), where a large enough γ is often needed to suppress worst-case disturbances. Unlike prior works based on implicit algebraic or Riccati-type conditions, our framework provides *explicit and verifiable* criteria tied directly to system and channel parameters.

Notably, if the Nash equilibrium of Problem 1 exists for a finite γ, it also exists as $\gamma \to \infty$ due to the monotonic character of the Nash equilibrium w.r.t. γ [23]. This insight aids in identifying the necessary conditions for the Nash equilibrium, as outlined in the following theorem.

Theorem 4 (Necessary Conditions for the Existence of the Nash Equilibrium of Problem 1)**.** *If the Nash equilibrium of Problem 1 exists, then*

- $p_1 > 1 - \frac{1}{\|\mathbf{A}\|^2}$;
- *The pair* $\Big(\mathbf{A}, \big(\mathbb{E}[\mathbf{B}\mathbf{H}_1(t)(\mathbf{R}_1 + \mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t))^{-1}\mathbf{H}_1^T(t)\mathbf{B}^T]\big)^{\frac{1}{2}}\Big)$ *is controllable.*

Proof: Please refer to Appendix D. ∎

According to Theorem 4, the existence of the Nash equilibrium only requires the "average controllability" of the system w.r.t. the stabilizing controller.

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

A. Online Structured Learning Algorithm for the Nash Equilibrium of Problem 1

Using the structural form of the optimal reduced state value function $\tilde{V}_i^*(\mathbf{x}(t))$, the optimal averaged cost θ_i^* 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 (10) and (11) can be written into coupled nonlinear matrix equation, as follows.

$$\mathbf{P} = \mathbb{E}\left[g(\mathbf{P}, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t))\right], \tag{20}$$

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

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

where

$$\begin{aligned} \mathcal{M}_{11}(t) &= \mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) + \\ &\delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_1(t), \end{aligned} \tag{22}$$

$$\mathcal{M}_{12}(t) = \delta_1(t)\delta_2(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_2(t), \tag{23}$$

$$\mathcal{M}_{21}(t) = \mathcal{M}_{12}^T(t), \tag{24}$$

and

$$\begin{aligned} \mathcal{M}_{22}(t) &= -\gamma^2\delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t) + \delta_2(t) \\ &\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_2(t) - \gamma^2\mathbf{R}_2. \end{aligned} \tag{25}$$

Since (20) is a fixed-point equation w.r.t. the unknown variable $\mathbf{P}$, we can utilize the SA theory [27] to construct an online learning algorithm to learn the unknown variable $\mathbf{P}$ based on (20). The learned unknown variable $\mathbf{P}$ can then be applied to obtain the optimal reduced-state value function $\tilde{V}_i^*(\mathbf{x}(t))$, and the optimal control solution $\mathbf{u}_i^*(t)$ for the Nash equilibrium $\{\pi_i^*\}_{i=1,2}$ of Problem 1.

Algorithm 1 Online Structured Learning for the Nash Equilibrium of Problem 1

• **Step 1:** Given an arbitrary $S \times S$ dimensional positive definite matrix $\mathbf{P}^1 \in \mathbb{S}_+^S$ (e.g., $\mathbf{P}^1 = \mathbf{I}_S$), the initial estimated optimal reduced-state value function is given by

$$\tilde{V}_i^1(\mathbf{x}(1)) = \mathbf{x}^T(1)\mathbf{P}^1\mathbf{x}(1), i \in \{1, 2\}, \tag{26}$$

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

$$\mathbf{u}_i(1) = \mathbf{K}_i(\mathbf{P}^1, 1)\mathbf{x}(1), \ \ i \in \{1, 2\}. \tag{27}$$

• **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:

$$\mathbf{u}_i(t) = \mathbf{K}_i(\mathbf{P}^t, t)\mathbf{x}(t), \ \ i \in \{1, 2\}, \tag{28}$$

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

$$\tilde{V}_i^t(\mathbf{x}(t)) = \mathbf{x}^T(t)\mathbf{P}^t\mathbf{x}(t), \ \ i \in \{1, 2\}. \tag{29}$$

• **Step 3:** $\mathbf{P}^{t+1}$ is updated based on $\mathbf{P}^t$ via (31). Set $t = t + 1$ and proceed to **Step 2**.

Specifically, (20) can be further written into standard form $f(\mathbf{P}) = \mathbf{0}_S$, where $f(\mathbf{P})$ is given by

$$f(\mathbf{P}) = \mathbb{E}\left[g\left(\mathbf{P}, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)\right)\right] - \mathbf{P}. \tag{30}$$

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:

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \alpha^t\left(g\left(\mathbf{P}^t, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)\right) - \mathbf{P}^t\right), \tag{31}$$

where $\{\alpha^t\}_{t=1}^{\infty}$ 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 (20).

Remark 1 (SA Algorithm in Algorithm 1)**.** *Our Algorithm 1 differs from standard SA methods for game solutions (e.g.,*

[2]In Algorithm 1, the CSI $\{\delta_i(t)\mathbf{H}_i(t)\}_{i=1,2}$ can be obtained via standard channel estimation at the plant using pilot symbols and feedback from the remote controllers [28].

[26], [29], [30]), which assume finite-dimensional learning variables. In contrast, our SA is conducted over an infinite-dimensional reduced-state value function space. By leveraging its structure, we effectively learn the value function through its parameter $\mathbf{P}$.

B. Implementation Considerations

According to Section III-A, the structural form of the optimal control solution can be further simplified when $\gamma \to \infty$ and the remote controllers are homogeneous. This enables low-complexity implementation of Algorithm 1.

1) Asymptotic Behavior: When $\gamma \to \infty$, the coupled nonlinear matrix equation (20) can be simplified as follows.

$$\mathbf{P} = \mathbf{Q} + \mathbb{E}[\mathbf{A}^T(\mathbf{P}^{-1} + \delta_1(t)\mathbf{B}\mathbf{H}_1(t)(\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t))^{-1}\mathbf{H}_1^T(t)\mathbf{B}^T)\mathbf{A}]. \quad (32)$$

As a result, Step 3 of Algorithm 1 can be simplified into the per-stage update:

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \alpha^t(\mathbf{A}^T((\mathbf{P}^t)^{-1} + \delta_1(t)\mathbf{B}\mathbf{H}_1(t)(\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t))^{-1}\mathbf{H}_1^T(t)\mathbf{B}^T)^{-1}\mathbf{A} + \mathbf{Q} - \mathbf{P}^t). \quad (33)$$

Using the argument in Corollary 1, the learned control solution $\mathbf{u}_i(t)$ obtained at Step 2 of Algorithm 1 can be simplified as follows:

$$\mathbf{u}_1(t) = -\left(\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}^t\mathbf{B}\mathbf{H}_1(t)\right)^{-1}\delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}^t\mathbf{A}\mathbf{x}(t), \quad (34)$$

and $\mathbf{u}_2(t) = \mathbf{0}_{N_{2,t}\times 1}$. Hence, the computational complexity of Algorithm 1 can be reduced.

2) Homogeneous Behavior: When the remote controllers are homogeneous in the sense that $\delta_1(t) = \delta_2(t) = 1, \mathbf{H}_1(t) = \mathbf{H}_2(t) = \mathbf{H}(t), \mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}$ and $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{M}$, the coupled nonlinear matrix equation (20) can be simplified as follows.

$$\mathbf{P} = \mathbf{Q} + \mathbb{E}[\mathbf{A}^T(\mathbf{P}^{-1} + (1-\gamma^{-2})\mathbf{B}\mathbf{H}(t)(\mathbf{R} + \mathbf{H}^T(t)\mathbf{B}^T\mathbf{M}\mathbf{B}\mathbf{H}(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T)^{-1}\mathbf{A}]. \quad (35)$$

As a result, Step 3 of Algorithm 1 can be simplified into the per-stage update:

$$\mathbf{P}^{t+1} = \mathbf{P}^t + \alpha^t\left(\mathbf{A}^T((\mathbf{P}^t)^{-1} + (1-\gamma^{-2})\mathbf{B}\mathbf{H}(t)(\mathbf{R} + \mathbf{H}^T(t)\mathbf{B}^T\mathbf{M}\mathbf{B}\mathbf{H}(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T)^{-1}\mathbf{A} + \mathbf{Q} - \mathbf{P}^t\right). \quad (36)$$

Moreover, using the argument in Corollary 2, the learned control solution $\mathbf{u}_i(t)$ obtained at Step 2 of Algorithm 1 can be simplified as follows:

$$\mathbf{u}_1(t) = -(\mathbf{R} + \mathbf{H}^T(t)\mathbf{B}^T\mathbf{M}\mathbf{B}\mathbf{H}(t) + \mathbf{H}^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_1^t\mathbf{B}\mathbf{H}(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_1^t\mathbf{A}\mathbf{x}(t), \quad (37)$$

and

$$\mathbf{u}_2(t) = (\gamma^2\mathbf{R} + \gamma^2\mathbf{H}^T(t)\mathbf{B}^T\mathbf{M}\mathbf{B}\mathbf{H}(t) - \mathbf{H}^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_2^t\mathbf{B}\mathbf{H}(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T\widetilde{\mathbf{P}}_2^t\mathbf{A}\mathbf{x}(t), \quad (38)$$

where

$$\widetilde{\mathbf{P}}_i^t = \left((\mathbf{P}^t)^{-1} - \gamma^{-4+2i}\mathbf{B}\mathbf{H}(t)(\mathbf{R} + \mathbf{H}^T(t)\mathbf{B}^T\mathbf{M}\mathbf{B}\mathbf{H}(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T\right)^{-1}, i \in \{1,2\}. \quad (39)$$

As a result, the computational complexity of Algorithm 1 can be reduced.

C. Convergence Analysis

The ODE method serves as an important tool for analyzing the convergence of the SA algorithm [26]. While classical in SA literature, our analysis departs from standard treatments in two main ways. First, instead of tracking finite-dimensional variables, we study the evolution of a structured kernel $\mathbf{P} \in \mathbb{S}_+^S$ within an infinite-dimensional value function. Second, due to the nonlinearity and randomness from MIMO fading, we avoid direct Lyapunov analysis. Instead, we construct a virtual fixed-point iteration whose trajectory approximates the ODE and use it to prove convergence. This structure-aware and indirect approach allows us to rigorously analyze Algorithm 1 in a non-standard setting.

We now present the formal convergence statements and supporting lemmas. Specifically, the following lemma establishes that the stochastic evolution in Algorithm 1 asymptotically tracks a limiting ODE trajectory, which forms the basis of our convergence analysis.

Lemma 1 (The ODE Trajectory for the Stochastic Evolution (31))**.** *The stochastic evolution (31) will asymptotically track the ODE trajectory given by:*

$$\dot{\bar{\mathbf{P}}}_k = f(\bar{\mathbf{P}}_k), k \in [1,\infty], \ \bar{\mathbf{P}}_1 = \mathbf{P}^1. \quad (40)$$

In other words, $\Pr(\lim_{k\to\infty}\bar{\mathbf{P}}_k = \lim_{t\to\infty}\mathbf{P}^t) = 1$.

Proof: Please refer to Appendix E. ∎

As a result, we can analyze the convergence of the stochastic evolution (31) in Algorithm 1 by analyzing the convergence of the ODE trajectory (40).

Typically, Lyapunov stability analysis is employed to demonstrate the convergence of the ODE trajectory (40), but finding suitable analytical Lyapunov functions for our scenario is challenging. To address this, we construct a virtual fixed-point discrete iteration whose associated interpolated continuous trajectory closely approximates the ODE trajectory (40).

Lemma 2 (Virtual Fixed-Point Discrete Iteration)**.** *Let the virtual fixed-point discrete iteration* $\left\{\widehat{\mathbf{P}}^t, \widehat{\mathbf{P}}^1 = \mathbf{P}^1, t \in \mathbb{Z}_+\right\}$ *follows the recursion:*

$$\widehat{\mathbf{P}}^{t+1} = \widehat{\mathbf{P}}^t + \eta f(\widehat{\mathbf{P}}^t), \quad (41)$$

where $\eta > 0$ *is an arbitrary given positive constant. Further let the interpolated continuous trajectory* $\left\{\tilde{\mathbf{P}}_k, k \in [1,\infty], \tilde{\mathbf{P}}_1 = \widehat{\mathbf{P}}^1\right\}$ *for the virtual fixed-point discrete iteration* $\left\{\widehat{\mathbf{P}}^t, \widehat{\mathbf{P}}^1 = \mathbf{P}^1, t \in \mathbb{Z}_+\right\}$ *follows the dynamics:*

$$\tilde{\mathbf{P}}_k = \widehat{\mathbf{P}}^t + (k - k_t)f(\widehat{\mathbf{P}}^t), k \in [k_t, k_{t+1}], \quad (42)$$

where $k_t = t\eta$. We have:

$$\sup_{k\in[0,L]} \|\tilde{\mathbf{P}}_{l+k} - \bar{\mathbf{P}}^l_{l+k}\| = \mathcal{O}(\eta), \tag{43}$$

where $\bar{\mathbf{P}}^l_k$ is the limiting ODE (40) with the initial condition $\bar{\mathbf{P}}^l_l = \tilde{\mathbf{P}}_l$, $l \geq 1$ and $L > 0$.

Proof: Please refer to Appendix F. ∎

Since the constant value η can be made arbitrary small by letting $\eta \to 0$, the convergence of the stochastic iteration (31) for Algorithm 1 can be obtained by analyzing the convergence of virtual discrete fixed-point iteration $\left\{\widehat{\mathbf{P}}^t, \widehat{\mathbf{P}}^1 = \mathbf{P}^1, t \in \mathbb{Z}_+\right\}$. The full convergence result for Algorithm 1 is summarized in the following theorem.

Theorem 5 (Almost Sure Convergence of Algorithm 1)**.** *If the sufficient conditions in Theorem 3 are satisfied, and the online learning step-size sequence $\{\alpha^t\}$ satisfies $\sum_{t=1}^{\infty} \alpha^t = \infty$, $\sum_{t=1}^{\infty} (\alpha^t)^2 < \infty$, then we have:*

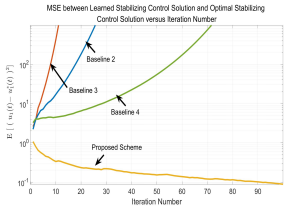
- *(a) The learned kernel $\mathbf{P}^t$ via Algorithm 1 converges to the ground truth kernel $\mathbf{P}$ almost surely, i.e., $\Pr(\lim_{t\to\infty} \mathbf{P}^t = \mathbf{P}) = 1$;*
- *(b) The learned reduced-state value function $\tilde{V}_i^t(\mathbf{x}(t))$ via Algorithm 1 converges to the optimal reduced-state value function $\tilde{V}_i^*(\mathbf{x}(t)) = \mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t)$ almost surely, i.e., $\Pr(\lim_{t\to\infty} \tilde{V}_i^t(\mathbf{x}(t)) = \tilde{V}_i^*(\mathbf{x}(t))) = 1, i \in \{1, 2\}$;*
- *(c) The control solution $\mathbf{u}_i(t)$ via Algorithm 1 converges to the optimal control solution $\mathbf{u}_i^*(t)$ in Theorem 2, i.e., $\Pr(\lim_{t\to\infty} \mathbf{u}_i(t) = \mathbf{u}_i^*(t)) = 1$, where $\mathbf{u}_1^*(t)$ and $\mathbf{u}_2^*(t)$ is the t-th element of the Nash equilibrium of Problem 1, i.e., $\{\pi_i^*\}_{i=1,2} = \{\mathbf{u}_1^*(t), \mathbf{u}_2^*(t), \forall t \in \mathbb{Z}_+\}$.*

Proof: Please refer to Appendix G. ∎

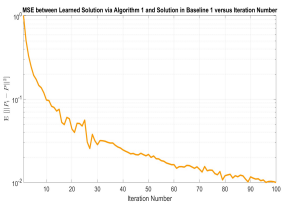
V. Numerical Results

In this section, we compare the proposed stabilizing control scheme in Algorithm 1 with various existing stabilizing control approaches for $\mathbf{u}_1(t)$ under external interference $\mathbf{u}_2(t)$ from two sources: (a) **Worst-case disturbance:** where the attacker uses the optimal destabilizing control from Theorem 2; and (b) **Random disturbance:** where a sine wave disturbance generator [31] applies $\mathbf{u}_2(t) = 10\sin(6t)\phi(t)$, with $\phi(t) \in \mathbb{R}^{N_{2,t}\times 1}$ drawn from a Gaussian distribution with zero mean and unit variance. The baseline schemes for $\mathbf{u}_1(t)$ are summarized below.

- **Baseline 1:** *(Prior-Known Nash Equilibrium [23])* The Nash equilibrium of Problem 1 is known at the stabilizing controller. Specifically, $\mathbf{P}$ that satisfies (20) is presumed to be known at the stabilizing controller. The optimal stabilizing control solution $\mathbf{u}_1^*(t)$, derived from Theorem 2, is implemented in the system.
- **Baseline 2:** *(Brute-Force Value Iteration without State Reduction [15])* 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 $(N_{1,t} \times N_r + N_{2,t} \times N_r) \times 2L$ finite intervals. The value for the optimal value function of the stabilizing controller $V_1^*(\mathbf{S}(t))$ is approximated by the value of the pseudo value function $\widehat{V}_1^d(\mathbf{x}(t)) = \mathbf{x}^T(t)\mathbf{P}^d\mathbf{x}(t), \mathbf{P}^d \in \mathbb{S}_+^S, 1 \leq d \leq (N_{1,t}\times N_r + N_{2,t}\times N_r)\times 2L$ 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 value function $\widehat{V}_1^d(\mathbf{x}(t)), 1 \leq d \leq (N_{1,t} \times N_r + N_{2,t} \times N_r) \times 2L$.
- **Baseline 3:** *(Brute-Force Value Iteration over Static Channels [12])* The optimal value function for the stabilizing controller is approximated by the pseudo value function $\widehat{V}_1^s(\mathbf{x}(t)) = \mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t), \mathbf{P} \in \mathbb{S}_+^S$. Based on $\mathbf{x}(t)$ at each t-th timeslot, the stabilizing controller learns the control policy via brute-force value iteration using the least square with $\{\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(t)\}$.
- **Baseline 4:** *(Naive LQR Control [32])* The stabilizing controller applies the naive LQR control solution without awareness of the disturbance.

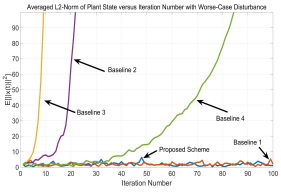


(a) MSE vs. iteration for optimal stabilizing control.

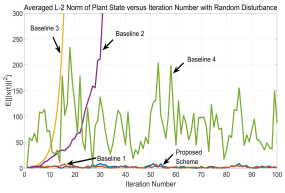


(b) MSE vs. iteration for Baseline 1 solution.

Fig. 2: The convergence analysis for the control schemes. The worse-case disturbance by the attacker is presented. 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 $\mathbf{B} = \begin{bmatrix} 1.61 & 0.67 \\ 0.74 & 0.52 \\ 1.02 & 0.56 \end{bmatrix}$. $p = 0.8$, $N_{1,t} = N_{2,t} = N_r = 2$, $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{I}_2$, $\mathbf{M}_1 = \mathbf{M}_2 = \mathbf{Q} = \mathbf{I}_3$, $\gamma = 10$, $\mathbf{W} = \mathbf{I}_3$, and $L = 2$. The Baseline 1 solution is $\mathbf{P} = \begin{bmatrix} 3.2564 & 0.3377 & 0.0479 \\ 0.3377 & 2.3516 & 0.6592 \\ 0.0479 & 0.6592 & 1.3388 \end{bmatrix}$.



(a) Averaged L_2-norm of plant state vs. iteration under worst-case disturbance.



(b) Averaged L_2-norm vs. iteration under sinusoidal random disturbance.

Fig. 3: Robustness analysis for the control schemes. The system parameters are configured as follows: $\mathbf{A} = \begin{bmatrix} 1.46 & 0.44 & 0.17 \\ 0.13 & 0.92 & 0.43 \\ 0.41 & 0.67 & 0.56 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1.41 & 0.43 \\ 0.67 & 0.44 \\ 0.92 & 0.51 \end{bmatrix}$. The Baseline 1 solution is $\mathbf{P} = \begin{bmatrix} 4.3580 & 0.2186 & -0.0981 \\ 0.2186 & 3.2266 & 1.2930 \\ -0.0981 & 1.2930 & 1.8146 \end{bmatrix}$.

A. Convergence Analysis

Fig. 2(a) shows the Mean Square Error (MSE) between the learned control $\mathbf{u}_1(t)$ and the Nash solution $\mathbf{u}_1^*(t)$ from Baseline 1. The proposed structured learning algorithm converges reliably, confirming the asymptotic optimality of Algorithm 1. In contrast, Baselines 2–4 diverge due to different modeling or approximation errors. Baseline 2 suffers from CSI discretization; Baseline 3 assumes static channels, leading to mismatch under dynamics; and Baseline 4 applies a fixed LQR gain without accounting for interference.

These trends can be understood via the structure $\mathbf{u}_1(t) = \mathbf{K}(\mathbf{P}_t, t)\mathbf{x}(t)$, where $\mathbf{P}_t$ is the evolving policy kernel. In our

algorithm, $\mathbf{P}_t$ is updated via structured SA and converges to the optimal $\mathbf{P}$, ensuring $\mathbf{u}_1(t) \to \mathbf{u}_1^*(t)$, as shown in Fig. 2(b). In Baselines 2 and 3, $\mathbf{P}_t$ fails to converge due to CSI discretization or model mismatch, leading to simultaneous instability in both policy $\mathbf{P}_t$ and state $\mathbf{x}(t)$. Baseline 4 uses a fixed kernel $\mathbf{P}_l$, avoiding instability from fluctuating $\mathbf{P}_t$, but the static gain $\mathbf{K}(\mathbf{P}_l, t)$, which reduces to $\mathbf{K}(\mathbf{P}_l)$, cannot adapt to dynamic channels or disturbances, resulting in a persistent suboptimality and gradually increasing deviation in state $\mathbf{x}(t)$ and control $\mathbf{u}_1(t) = \mathbf{K}(\mathbf{P}_t, t)\mathbf{x}(t)$ over time.

B. Robustness Analysis

Fig. 3(a) shows the averaged state norm $\mathbb{E}[\|\mathbf{x}(t)\|^2]$ under worst-case disturbance. The proposed method and Baseline 1 achieve the lowest values, reflecting robustness by following the true Nash equilibrium, consistent with Fig. 2. In contrast, Baselines 2 and 3 rely on approximate value iteration or static CSI assumptions and fail to mitigate adversarial effects, leading to instability. Baseline 4, which lacks disturbance modeling, also diverges under interference.

Fig. 3(b) presents results under random sinusoidal disturbances. Again, the proposed method and Baseline 1 maintain low state norms, showing resilience to stochastic interference. Baseline 2 considers CSI and disturbance but suffers from discretization errors. Baseline 3 neglects channel variation and fails to stabilize the plant. Baseline 4 performs poorly due to its fixed, disturbance-unaware policy, highlighting the importance of adaptive control under uncertainty.

VI. Conclusion

In this paper, we explored a zero-sum game for a linear system over wireless MIMO fading channels with uncountable CSI state space. We framed the problem as a stochastic ergodic game and derived structured reduced-state Bellman optimality equations to overcome the "curse of dimensionality" inherent in learning the Nash equilibrium from the uncountable CSI state space. Utilizing these equations, we derived the sufficient and necessary conditions for the existence of the Nash equilibrium and proposed a novel structured online learning algorithm that asymptotically achieves the Nash equilibrium via SA iteration. Numerical results demonstrated superior stability and convergence performance of the proposed scheme compared to the baseline approaches.

Appendix

A. Proof of Theorem 1

We begin by analyzing the finite-horizon zero-sum game and derive the structural properties of the optimality conditions within our framework through asymptotic analysis.

Problem 2 (The Finite-Horizon Game)**.**

Stabilizing Controller (Controller 1):

$$\min_{\pi_1} \max_{\pi_2} \tilde{\mathcal{J}}^{\pi_1,\pi_2}, \ \textit{s.t. (3).} \tag{44}$$

Destabilizing Controller (Controller 2):

$$\max_{\pi_2} \min_{\pi_1} \tilde{\mathcal{J}}^{\pi_2,\pi_1}, \ \textit{s.t. (3),} \tag{45}$$

where $\tilde{\mathcal{J}}^{\pi_1,\pi_2} = \mathbb{E}^{\pi_1,\pi_2}[\sum_{t=1}^{T} r(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)) + \mathbf{x}^T(T+1)\mathbf{Q}\mathbf{x}^T(T+1)]$.

The optimal solution to Problem 2 can be obtained by backward induction. We define the value functions $V_1(t, \mathbf{S}(t))$ and $V_2(t, \mathbf{S}(t))$ that satisfy $V_1(T+1, \mathbf{S}(T+1)) = V_2(T+1, \mathbf{S}(T+1)) = \mathbf{x}^T(T+1)\mathbf{Q}\mathbf{x}(T+1)$ as the solutions to:

$$V_1(t, \mathbf{S}(t)) = \min_{\mathbf{u}_1(t)} \max_{\mathbf{u}_2(t)} [r(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)) + \mathbb{E}[V_1(t+1, \mathbf{S}(t+1))|\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)]], \tag{46}$$

$$V_2(t, \mathbf{S}(t)) = \max_{\mathbf{u}_2(t)} \min_{\mathbf{u}_1(t)} [r(\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)) + \mathbb{E}[V_2(t+1, \mathbf{S}(t+1))|\mathbf{S}(t), \mathbf{u}_1(t), \mathbf{u}_2(t)]], t = 1, 2, ..., T. \tag{47}$$

Let $t = T$. We have:

$$\begin{aligned} &V_1(T, \mathbf{S}(T)) = V_2(T, \mathbf{S}(T)) = V(T, \mathbf{S}(T)) \\ &= \mathbf{x}^T(T)(\mathbf{A}^T\mathbf{Q}\mathbf{A} + \mathbf{Q})\mathbf{x}(T) + \mathrm{Tr}(\mathbf{Q}\mathbf{W} + \mathbf{B}^T\mathbf{Q}\mathbf{B}) \\ &+ \max_{\mathbf{u}_2(T)} \min_{\mathbf{u}_1(T)} (\mathbf{u}_1^T(T)(\mathbf{R}_1 + \delta_1(T)\mathbf{H}_1^T(T)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(T) \\ &+ \delta_1(T)\mathbf{H}_1^T(T)\mathbf{B}^T\mathbf{Q}\mathbf{B}\mathbf{H}_1(T))\mathbf{u}_1(T) + \mathbf{u}_2^T(T)(-\gamma^2\mathbf{R}_2 - \gamma^2\delta_2(T) \\ &\times \mathbf{H}_2^T(T)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(T) + \delta_2(T)\mathbf{H}_2^T(T)\mathbf{B}^T\mathbf{Q}\mathbf{B}\mathbf{H}_2(T))\mathbf{u}_2(T) \\ &+ 2\mathbf{u}_1^T(T)(\delta_1(T)\delta_2(T)\mathbf{H}_1^T(T)\mathbf{B}^T\mathbf{Q}\mathbf{B}\mathbf{H}_2(T))\mathbf{u}_2(T)), \end{aligned} \tag{48}$$

with the optimal solution in the Right-Hand Side (R.H.S.) of (48) given by

$$\begin{aligned} &\mathbf{u}_1^*(T) = -(\mathbf{R}_1 + \delta_1(T)\mathbf{H}_1^T(T)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(T) + \delta_1(T) \\ &\mathbf{H}_1^T(T)\mathbf{B}^T\widetilde{\mathbf{Z}}_1(T)\mathbf{B}\mathbf{H}_1(T))^{-1}\delta_1(T)\mathbf{H}_1^T(T)\mathbf{B}^T\widetilde{\mathbf{Z}}_1(T)\mathbf{A}\mathbf{x}(T), \end{aligned} \tag{49}$$

$$\begin{aligned} &\mathbf{u}_2^*(T) = (\gamma^2\delta_2(t)\mathbf{H}_2^T(T)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(T) + \gamma^2\mathbf{R}_2 - \delta_2(T) \\ &\mathbf{H}_2(T)\mathbf{B}^T\widetilde{\mathbf{Z}}_2(T)\mathbf{B}\mathbf{H}_2(T))^{-1}\delta_2(T)\mathbf{H}_2^T(T)\mathbf{B}^T\widetilde{\mathbf{Z}}_2(T)\mathbf{A}\mathbf{x}(T), \end{aligned} \tag{50}$$

where

$$\begin{aligned} &\widetilde{\mathbf{Z}}_1(T) = (\mathbf{Q}^{-1} - \gamma^{-2}\delta_2(T)\mathbf{B}\mathbf{H}_2(T)(\mathbf{R}_2 + \delta_2(T)\mathbf{H}_2^T(T) \\ &\times \mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(T))^{-1}\mathbf{H}_2^T(T)\mathbf{B}^T)^{-1}, \end{aligned} \tag{51}$$

$$\begin{aligned} &\widetilde{\mathbf{Z}}_2(T) = (\mathbf{Q}^{-1} - \delta_1(T)\mathbf{B}\mathbf{H}_1(T)(\mathbf{R}_1 + \delta_1(T)\mathbf{H}_1^T(T)\mathbf{B}^T\mathbf{M}_1 \\ &\times \mathbf{B}\mathbf{H}_1(T))^{-1}\mathbf{H}_1^T(T)\mathbf{B}^T)^{-1}. \end{aligned} \tag{52}$$

Let $\bar{\mathbf{Z}}(T+1) = \mathbf{Q}$. For $t = 1, 2, ..., T$, we have

$$V(t, \mathbf{S}(t)) = \mathbf{x}^T(t)\mathbf{Z}(t)\mathbf{x}(t) + \sum_{i=t}^{T} \mathrm{Tr}(\bar{\mathbf{Z}}(i+1)\mathbf{W} + \mathbf{B}^T\bar{\mathbf{Z}}(i+1)\mathbf{B}), \tag{53}$$

$$\mathbf{Z}(t) = g(\bar{\mathbf{Z}}(t+1), \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)), \tag{54}$$

where

$$\begin{aligned} &g(\bar{\mathbf{Z}}(t+1), \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)) = \mathbf{Q} + \mathbf{A}^T\bar{\mathbf{Z}}(t+1)\mathbf{A} - \mathbf{A}^T \\ &\times \begin{bmatrix} \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\bar{\mathbf{Z}}(t+1) \\ \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\bar{\mathbf{Z}}(t+1) \end{bmatrix}^T \begin{bmatrix} \mathcal{Z}_{11}(t) & \mathcal{Z}_{12}(t) \\ \mathcal{Z}_{21}(t) & \mathcal{Z}_{22}(t) \end{bmatrix}^{-1} \begin{bmatrix} \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\bar{\mathbf{Z}}(t+1) \\ \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\bar{\mathbf{Z}}(t+1) \end{bmatrix}\mathbf{A}, \end{aligned} \tag{55}$$

$$\begin{aligned} &\mathcal{Z}_{11}(t) = \mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) \\ &+ \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\bar{\mathbf{Z}}(t+1)\mathbf{B}\mathbf{H}_1(t), \end{aligned} \tag{56}$$

$$\mathcal{Z}_{12}(t) = \delta_1(t)\delta_2(t)\mathbf{H}_1^T(t)\mathbf{B}^T\bar{\mathbf{Z}}(t+1)\mathbf{B}\mathbf{H}_2(t), \tag{57}$$

$$\mathcal{Z}_{21}(t) = \mathcal{Z}_{12}^T(t), \tag{58}$$

$$\begin{aligned} &\mathcal{Z}_{22}(t) = -\gamma^2\delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t) \\ &+ \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\bar{\mathbf{Z}}(t+1)\mathbf{B}\mathbf{H}_2(t) - \gamma^2\mathbf{R}_2, \end{aligned} \tag{59}$$

and $\bar{\mathbf{Z}}(t) = \mathbb{E}_{\delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)}[\mathbf{Z}(t)]$.

Note that the optimal average cost in Problem 1 can be characterized by $\mathcal{J}^{\pi_1^*,\pi_2^*}$, where $\pi_i^* = \{\mathbf{u}_i^*(t), \ t \in \mathbb{Z}_+\}$ and $\mathbf{u}_i^*(t)$ is characterized in (49) and (50), as the horizon length $K \to \infty$. The structure of the optimality conditions for

Problem 1 can therefore be analyzed by investigating the asymptotic behavior of the following recursion:

$$\mathbf{Z}(i+1) = g(\bar{\mathbf{Z}}(i), \delta_1(i)\mathbf{H}_1(i), \delta_2(i)\mathbf{H}_2(i)), \mathbf{Z}(1) = \mathbf{Q}. \quad (60)$$

To analyze this, we consider the induced expected recursion:

$$\bar{\mathbf{Z}}(i+1) = \mathbb{E}[g(\bar{\mathbf{Z}}(i), \delta_1(i)\mathbf{H}_1(i), \delta_2(i)\mathbf{H}_2(i))], \bar{\mathbf{Z}}(1) = \mathbf{Q}. \quad (61)$$

Due to the monotonicity of the operator $\mathbb{E}[g(\cdot)]$ w.r.t. $\bar{\mathbf{Z}}(i)$ [23], and its contractivity under the conditions of Theorem 3, there exists an upper bound $\bar{\mathbf{Z}}^u$ such that for any $\bar{\mathbf{Z}} \succeq \bar{\mathbf{Z}}^u$, we have $\mathbb{E}[g(\bar{\mathbf{Z}})] \prec \bar{\mathbf{Z}}$.

Hence, based on the bounded convergence theorem, the sequence converges with $\limsup_{i\to\infty} \bar{\mathbf{Z}}(i) = \mathbf{P}$, and hence

$$\limsup_{i\to\infty} \bar{\mathbf{Z}}(i+1) = \bar{\mathbf{P}}(\delta_1(i+1)\mathbf{H}_1(i+1), \delta_2(i+1)\mathbf{H}_2(i+1)), \quad (62)$$

which is a continuous function of the i.i.d. variables $\{\delta_1(i+1)\mathbf{H}_1(i+1), \delta_2(i+1)\mathbf{H}_2(i+1)\}$. This gives the equilibrium structure, optimal cost, and value function in Theorem 1, completing the proof.

B. Proof of Corollary 1 and Corollary 2

Let $\mathbf{S}_1(t) = \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_1(t)$ and define $\mathbf{D}(t) = -\gamma^2\delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t) - \gamma^2\mathbf{R}_2 + \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_2(t)$. As $\gamma \to \infty$, we have:

$$\begin{aligned}
&\lim_{\gamma\to\infty} [\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) + \mathbf{S}_1(t) - \delta_1(t)\delta_2(t)\mathbf{H}_1^T(t) \\
&\times \mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_2(t)\mathbf{D}^{-1}(t)\delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_1(t)]^{-1} \\
&= [\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t) + \mathbf{S}_1(t)]^{-1}, \quad (63) \\
&\lim_{\gamma\to\infty} \mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_2(t)\mathbf{D}^{-1}(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{P}\mathbf{A} = \mathbf{0}_{N_{1,t}\times S}, \quad (64) \\
&\lim_{\gamma\to\infty} \mathbf{K}_1(\mathbf{P}, t) = [\mathbf{R}_1 + \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T(\mathbf{M}_1 + \mathbf{P})\mathbf{B}\mathbf{H}_1(t)]^{-1} \\
&\times \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{A}. \quad (65)
\end{aligned}$$

A symmetric argument gives $\lim_{\gamma\to\infty} \mathbf{K}_2(\mathbf{P}, t) = \mathbf{0}_{N_{2,t}\times S}$, completing the proof of Corollary 1.

When the controllers are homogeneous, we obtain:

$$\begin{aligned}
&\mathbf{K}_1(\mathbf{P}, t) = [\mathbf{R} + \mathbf{H}^T(t)\mathbf{B}^T(\mathbf{M} + \mathbf{P})\mathbf{B}\mathbf{H}(t) \\
&- \mathbf{H}^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}(t)(\mathbf{D}_\gamma(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}(t)]^{-1} \\
&\times [\mathbf{H}^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}(t)(\mathbf{D}_\gamma(t))^{-1}\mathbf{H}^T(t)\mathbf{B}^T\mathbf{P}\mathbf{A} - \mathbf{H}^T(t)\mathbf{B}^T\mathbf{P}\mathbf{A}],
\end{aligned} \quad (66)$$

where

$$\mathbf{D}_\gamma(t) = -\gamma^2\mathbf{H}^T(t)\mathbf{B}^T\mathbf{M}\mathbf{B}\mathbf{H}(t) - \gamma^2\mathbf{R} + \mathbf{H}^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}(t). \quad (67)$$

Taking $\gamma \to \infty$ in (66) yields Corollary 2.

C. Proof of Theorem 3

If $\mathbf{P} \in \mathbb{S}_+^S$ exists, (10) and (11) can be reformulated as follows.

$$\begin{aligned}
&\mathbf{x}^T(t)\mathbf{P}\mathbf{x}(t) = \mathbb{E}_{\{\delta_i(t)\mathbf{H}_i(t)\}}[\max_{\mathbf{u}_2(t)} \min_{\mathbf{u}_1(t)} \mathbf{x}^T(t)\mathbf{Q}\mathbf{x}(t) + \mathbf{u}_1^T(t)\mathbf{R}_1\mathbf{u}_1(t) \\
&- \gamma^2\mathbf{u}_2(t)\mathbf{R}_2\mathbf{u}_2(t) + (\delta_1(t)\mathbf{B}\mathbf{H}_1(t)\mathbf{u}_1(t))^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t)\mathbf{u}_1(t) \\
&- \gamma^2(\delta_2(t)\mathbf{B}\mathbf{H}_2(t)\mathbf{u}_2(t))^T\mathbf{M}_2\mathbf{B}\mathbf{H}_2(t)\mathbf{u}_2(t) + (\mathbf{A}\mathbf{x}(t) \\
&+ \sum_{i=1}^2 \delta_i(t)\mathbf{H}_i(t)\mathbf{u}_i(t))^T\mathbf{P}(\mathbf{A}\mathbf{x}(t) + \sum_{i=1}^2 \delta_i(t)\mathbf{H}_i(t)\mathbf{u}_i(t))].
\end{aligned} \quad (68)$$

The existence of a Nash equilibrium is ensured if $\gamma^2 \geq \frac{\|\mathbf{P}\|\,\|\mathbf{B}^T\mathbf{B}\|}{\lambda_{\min}(\mathbf{M}_2)}$. This condition highlights the need to analyze the existence of $\mathbf{P}$ and to characterize how $\mathbf{P}$ depends on the system parameters, assuming $\mathbf{P}$ exists.

Note that (10) and (11) define a fixed-point equation w.r.t. the $\mathbf{P}$ in (20), which can be further expressed as follows.

$$\begin{aligned}
\mathbf{P} &= \mathbb{E}\Bigg[\mathbf{A}^T\Bigg(\mathbf{P}^{-1} + \begin{bmatrix}\delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T \\ \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\end{bmatrix}^T \begin{bmatrix}\mathcal{N}_1(t) & \mathbf{0} \\ \mathbf{0} & \mathcal{N}_2(t)\end{bmatrix} \\
&\begin{bmatrix}\delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T \\ \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\end{bmatrix}\Bigg)^{-1}\mathbf{A}\Bigg] + \mathbf{Q} \\
&= \mathbf{Q} + \mathbb{E}[\mathbf{A}^T(\mathbf{P}^{-1} + \mathbf{B}\bar{\mathcal{M}}(t)\mathbf{B}^T)^{-1}\mathbf{A}],
\end{aligned} \quad (69)$$

where

$$\widehat{\mathcal{M}}_{11}(t) = \mathcal{M}_{11}(t) - \delta_1(t)\mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_1(t), \quad (70)$$

$$\widehat{\mathcal{M}}_{22}(t) = \mathcal{M}_{22}(t) - \delta_2(t)\mathbf{H}_2^T(t)\mathbf{B}^T\mathbf{P}\mathbf{B}\mathbf{H}_2(t), \quad (71)$$

$$\mathcal{N}_1(t) = \widehat{\mathcal{M}}_{11}^{-1}(t), \mathcal{N}_2(t) = \widehat{\mathcal{M}}_{22}^{-1}(t), \quad (72)$$

$$\bar{\mathcal{M}}(t) = \delta_1(t)\mathbf{H}_1(t)\mathcal{N}_1(t)\mathbf{H}_1^T(t) + \delta_2(t)\mathbf{H}_2(t)\mathcal{N}_2(t)\mathbf{H}_2^T(t). \quad (73)$$

Let

$$\mathbf{P}^c(t) = \mathbf{V}^T(t)\begin{bmatrix}[\mathbf{P}(t)]_{\gamma(t)} & [\mathbf{P}(t)]_{\gamma(t)}\mathbf{L}(t) \\ \mathbf{L}^T(t)[\mathbf{P}(t)]_{\gamma(t)} & \mathbf{L}^T(t)[\mathbf{P}(t)]_{\gamma(t)}\mathbf{L}(t)\end{bmatrix}\mathbf{V}(t), \quad (74)$$

$$\begin{aligned}
\mathbf{P}^u(t) &= \mathbf{V}^T(t)(\mathbf{I}_S - \Pi(t))\mathbf{V}(t)\mathbf{P}\mathbf{V}^T(t)(\mathbf{I}_S - \Pi(t))\mathbf{V}(t) \\
&- \mathbf{V}^T(t)\begin{bmatrix}\mathbf{0}_{S-\gamma(t)} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}^T(t)[\mathbf{P}(t)]_{\gamma(t)}\mathbf{L}(t)\end{bmatrix}\mathbf{V}(t),
\end{aligned} \quad (75)$$

$$\mathbf{L}(t) = [\mathbf{P}(t)]_{\gamma(t)}(\mathbf{V}(t)\mathbf{P}(t)\mathbf{V}^T(t))_{1:\gamma(t),\gamma(t)+1:S}, \quad (76)$$

where $[\mathbf{P}(t)]_{\gamma(t)}$ denotes the $\gamma(t)$-th order principal submatrix of $\mathbf{V}(t)\mathbf{P}\mathbf{V}^T(t)$.

It then gives that

$$\begin{aligned}
\mathbf{P} &= \mathbf{Q} + \mathbf{A}^T\mathbb{E}[\mathbf{P}^u(t)]\mathbf{A} + \mathbf{A}^T\mathbb{E}[\mathbf{V}^T(t)\widehat{\zeta}(t)^{-1}\big(\widehat{\zeta}(t)\mathbf{V}(t)\mathbf{P}^c(t) \\
&\times \mathbf{V}^T(t)\widehat{\zeta}(t)\Pi(t) + \mathbf{I}\big)^{-1}\widehat{\zeta}(t)\mathbf{V}(t)\mathbf{P}^c(t)\mathbf{V}^T(t)\widehat{\zeta}(t)\widehat{\zeta}^{-1}(t)\mathbf{V}(t)]\mathbf{A} \\
&= \mathbf{Q} + \mathbf{A}^T\mathbb{E}[\mathbf{V}^T(t)(\mathbf{I} - \Pi(t))\mathbf{V}(t)\mathbf{P}\mathbf{V}^T(t)(\mathbf{I} - \Pi(t))\mathbf{V}(t)]\mathbf{A} \\
&+ \mathbf{A}^T\mathbb{E}[\mathbf{V}^T(t)\begin{bmatrix}(\zeta(t))_{\gamma(t)}^{-\frac{1}{2}}\mathcal{P}_{11}(t)(\zeta(t))_{\gamma(t)}^{-\frac{1}{2}} & \mathcal{P}_{12}(t) \\ \mathcal{P}_{21}(t) & \mathcal{P}_{22}(t) - \mathbf{L}^T(t)[\mathbf{P}(t)]_{\gamma(t)}\mathbf{L}(t)\end{bmatrix} \\
&\mathbf{V}(t)]\mathbf{A}.
\end{aligned} \quad (77)$$

Given Condition (b), we have:

$$\begin{aligned}
&\mathbb{E}\left[\begin{bmatrix}(\zeta(t))_{\gamma(t)}^{-\frac{1}{2}}\mathcal{P}_{11}(t)(\zeta(t))_{\gamma(t)}^{-\frac{1}{2}} & \mathcal{P}_{12}(t) \\ \mathcal{P}_{21}(t) & \mathcal{P}_{22}(t) - \mathbf{L}^T(t)[\mathbf{P}(t)]_{\gamma(t)}\mathbf{L}(t)\end{bmatrix}\right] \\
&\preceq \mathbb{E}[\mathrm{Diag}((\zeta(t))_{\gamma(t)}^{-1}, \mathbf{0}_{S-\gamma(t)})] \preceq \mathbb{E}[\mathrm{Tr}(\zeta^{-1}(t))]\,\mathbf{I}_S,
\end{aligned} \quad (78)$$

$$\begin{aligned}
\mathbf{P} &\preceq \mathbf{Q} + \mathbf{A}^T\mathbb{E}[\mathbf{V}^T(t)(\mathbf{I} - \Pi(t))\mathbf{V}(t)\mathbf{P}\mathbf{V}^T(t)(\mathbf{I} - \Pi(t))\mathbf{V}(t)]\mathbf{A} \\
&+ \|\mathbf{A}\|^2\,\mathbb{E}[\mathrm{Tr}(\zeta^{-1}(t))]\,\mathbf{I}_S.
\end{aligned} \quad (79)$$

Further, under Condition (a), there exists an upper bound $\mathbf{P}^{\mathrm{up}}$ satisfying

$$\begin{aligned}
&(\|\mathbf{Q}\| + \|\mathbf{A}\|^2\,\mathbb{E}[\mathrm{Tr}(\zeta^{-1}(t))]) \cdot \|\mathbf{I}_S - \mathbf{A}^T\,\mathbb{E}[\mathbf{V}^T(t)(\mathbf{I} - \Pi(t))\mathbf{V}(t) \\
&\times \mathbf{P}^{\mathrm{up}}\,\mathbf{V}^T(t)(\mathbf{I} - \Pi(t))\mathbf{V}(t)]\mathbf{A})^{-1}\| \,\mathbf{I}_S \prec \mathbf{P}^{\mathrm{up}},
\end{aligned} \quad (80)$$

such that $\mathbb{E}[g(\mathbf{P}^{\mathrm{up}})] \prec \mathbf{P}^{\mathrm{up}}$.

To analyze the fixed-point equation, we construct two matrix sequences for $t \in \mathbb{Z}_+$:

$$\mathbf{P}^{(1)}(t+1) = \mathbb{E}[g(\mathbf{P}^{(1)}(t))], \quad \mathbf{P}^{(1)}(1) = \mathbf{0}_S, \quad (81)$$

$$\mathbf{P}^{(2)}(t+1) = \mathbb{E}[g(\mathbf{P}^{(2)}(t))], \quad \mathbf{P}^{(2)}(1) \succeq \mathbf{P}^{\mathrm{up}}. \quad (82)$$

Due to the monotonicity of $\mathbb{E}[g(\mathbf{P})]$ w.r.t. $\mathbf{P}$ [23], we have

$$\mathbf{P}^{(1)}(t+1) \succeq \mathbf{P}^{(1)}(t), \quad \forall t \geq 1, \tag{83}$$

$$\mathbf{P}^{(2)}(t+1) \preceq \mathbf{P}^{(2)}(t), \quad \forall t \geq 1, \tag{84}$$

$$\mathbf{P}^{(1)}(t) \preceq \mathbf{P}^{(1)}(t+1) \preceq \mathbf{P}^{(2)}(t+1) \preceq \mathbf{P}^{(2)}(t) \preceq \mathbf{P}^{(2)}. \tag{85}$$

As a result, the sequence $\{\mathbf{P}^{(1)}(t)\}$ is monotonically increasing and bounded above, while $\{\mathbf{P}^{(2)}(t)\}$ is decreasing and bounded below. Hence, by the bounded convergence theorem:

$$\lim_{t\to\infty} \mathbf{P}^{(1)}(t) = (\mathbf{P}^{(1)})^* = \mathbb{E}[g((\mathbf{P}^{(1)})^*)], \tag{86}$$

$$\lim_{t\to\infty} \mathbf{P}^{(2)}(t) = (\mathbf{P}^{(2)})^* = \mathbb{E}[g((\mathbf{P}^{(2)})^*)]. \tag{87}$$

To prove the uniqueness of the Nash equilibrium of Problem 1, assume that the fixed-point equation (69) admits two distinct solutions $\mathbf{P}_1^* \neq \mathbf{P}_2^*$ such that $\mathbb{E}[g(\mathbf{P}_1^*)] = \mathbf{P}_1^*$ and $\mathbb{E}[g(\mathbf{P}_2^*)] = \mathbf{P}_2^*$. Then, there exists a scalar $\gamma \in (0,1)$ such that $\mathbf{P}_1^* \succeq \gamma \mathbf{P}_2^*$ but $\mathbf{P}_1^* \not\succeq \gamma' \mathbf{P}_2^*$ for some $\gamma' > \gamma$. Note that due to the monotonicity of $g(\cdot)$ and the positive definiteness of $\mathbf{Q}$,

$$\mathbb{E}[g(\gamma \mathbf{P}_2^*)] \succeq \gamma \mathbb{E}[g(\mathbf{P}_2^*)] + \gamma c_1 \mathbf{P}_2^* = (1+c_1)\gamma \mathbf{P}_2^*, \tag{88}$$

where $c_1 = \frac{\gamma \sigma_{\min}(\mathbf{Q})}{\|\mathbb{E}[g(\mathbf{P}_2^*)]\|} > 0$ and $\sigma_{\min}(\mathbf{Q}) = 1/\|\mathbf{Q}^{-1}\|$.

Therefore,

$$\mathbf{P}_1^* \succeq \mathbb{E}[g(\gamma \mathbf{P}_2^*)] \succeq (1+c_1)\gamma \mathbf{P}_2^*, \tag{89}$$

which contradicts the assumption that $\mathbf{P}_1^* \not\succeq \gamma' \mathbf{P}_2^*$ for some $\gamma' > \gamma$, since $(1+c_1)\gamma > \gamma'$. Hence, the fixed-point solution must be unique: $\mathbf{P}_1^* = \mathbf{P}_2^*$.

Further, given the condition $\gamma^2 \geq \|\mathbf{P}\|\|\mathbf{B}^T\mathbf{B}\|/\lambda_{\min}(\mathbf{M}_2)$, the existence of the Nash equilibrium in Problem 1 is guaranteed. Note that

$$\|\mathbf{P}\| \leq (\|\mathbf{Q}\| + \|\mathbf{A}\|^2 \mathbb{E}[\mathrm{Tr}(\zeta^{-1}(t))]) \times (1 - \|\mathbb{E}[(\mathbf{A}^T(\mathbf{I}-\Pi(t))\mathbf{V}(t))^T \mathbf{A}^T(\mathbf{I}-\Pi(t))\mathbf{V}(t)]\|)^{-1}. \tag{90}$$

As a result, we need

$$\gamma^2 \geq (\|\mathbf{Q}\|\|\mathbf{B}^T\mathbf{B}\|\lambda_{\min}^{-1}(\mathbf{M}_2) + \|\mathbf{A}\|^2 \mathbb{E}[\mathrm{Tr}(\zeta^{-1}(t))]\|\mathbf{B}^T\mathbf{B}\|\lambda_{\min}^{-1}(\mathbf{M}_2))(1 - \|\mathbb{E}[(\mathbf{A}^T(\mathbf{I}-\Pi(t))\mathbf{V}(t))^T \mathbf{A}^T(\mathbf{I}-\Pi(t))\mathbf{V}(t)]\|)^{-1}. \tag{91}$$

This establishes Condition (c) and completes the proof.

D. Proof of Theorem 4

Let the R.H.S. of (69) be denoted as $h(\mathbf{P}, \gamma)$. According to Lemma 3.5 in [23], $h(\mathbf{P}, \gamma)$ is a decreasing function w.r.t. the non-cooperation penalty γ. Consequently, if the Nash equilibrium exists, the solution $\mathbf{P}$ to the fixed-point equation $\mathbf{P} = h(\mathbf{P}, \infty)$ must also exist. This further implies that the following iteration is stable:

$$\mathbf{P}^{(3)}(t+1) = h(\mathbf{P}^{(3)}(t), \infty), \quad \mathbf{P}^{(3)}(1) \in \mathbb{S}_+^S. \tag{92}$$

Using similar arguments as in Theorem 2 of [33], we derive two stable lower bounds, $\mathbf{P}^{(3a)}(t)$ and $\mathbf{P}^{(3b)}(t)$, for $\mathbf{P}^{(3)}(t)$. These bounds are governed by the following dynamics, respectively:

$$\mathbf{P}^{(3a)}(t+1) = (1-p_1)\,\mathbf{A}^T \mathbf{P}^{(3a)}(t)\,\mathbf{A}, \quad \mathbf{P}^{(3a)}(1) = \mathbf{P}^{(3)}(1),$$

$$\mathbf{P}^{(3b)}(t+1) = p_1\,\mathbf{A}^T(2\,\mathbb{E}\big[\mathbf{B}\mathbf{H}_1(t)(\mathbf{R}_1 + \mathbf{H}_1^T(t)\mathbf{B}^T\mathbf{M}_1\mathbf{B}\mathbf{H}_1(t))^{-1} \cdot \mathbf{H}_1^T(t)\mathbf{B}^T\big] + (\mathbf{P}^{(3b)}(t))^{-1})^{-1}\mathbf{A}, \quad \mathbf{P}^{(3b)}(1) = \mathbf{P}^{(3)}(1). \tag{93}$$

This establishes the conditions stated in Theorem 4 and completes the proof.

E. Proof of Lemma 1

It is straightforward to verify that $f(\mathbf{P})$ in (30) satisfies the following conditions:

(a) *(Lipschitz Continuity)*: For $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{S}_+^S$, the function $f(\mathbf{P})$ satisfies

$$\|f(\mathbf{P}_1) - f(\mathbf{P}_2)\| \leq (1 + \|\mathbf{A}\|^2)\|\mathbf{P}_1 - \mathbf{P}_2\|, \quad \forall. \tag{94}$$

(b) *(Martingale Difference Noise)*: Let $\widehat{f}(\mathbf{P}^t) = g(\mathbf{P}^t) - \mathbf{P}^t$ and define the noise term $\mathbf{N}(t) = \widehat{f}(\mathbf{P}^t) - f(\mathbf{P}^t)$. Then, the sequence $\{\mathbf{N}(t), t \in \mathbb{Z}_+\}$ forms a martingale difference sequence w.r.t. the filtration

$$\mathcal{F}(t) \triangleq \sigma(\mathbf{P}^1, \ldots, \delta_1(t)\mathbf{H}_1(t), \delta_2(t)\mathbf{H}_2(t)), \tag{95}$$

satisfying the condition $\mathbb{E}[\mathbf{N}(t+1) \mid \mathcal{F}(t)] = \mathbf{0}_S$.

(c) *(Square Integrability)*: The sequence $\{\mathbf{N}(t), t \in \mathbb{Z}_+\}$ is square-integrable and satisfies

$$\mathbb{E}\left[\|\mathbf{N}(t+1)\|^2 \mid \mathcal{F}(t)\right] \leq 2\|\mathbf{A}\|^2\left(1 + \|\mathbf{P}(t)\|^2\right), \quad \forall t > 0. \tag{96}$$

As stated in Chapter 2 of [26], the stochastic evolution in (31) will asymptotically follow the ODE trajectory described in (40). This completes the proof of Lemma 1.

F. Proof of Lemma 2

Let $L = N\eta \in \mathbb{R}_+$ for some $N \in \mathbb{Z}_+$. For $k \in \mathbb{R}_+$, define $[k] = \min\{t\eta : t \in \mathbb{R}_+, t\eta < k\}$ and $k_t = t\eta$.

For $t \in \mathbb{Z}_+$ and $0 \leq l \leq N-1$, we have:

$$\tilde{\mathbf{P}}_{k_{t+l}} = \tilde{\mathbf{P}}_{k_t} + \int_{k_t}^{k_{t+l}} f(\tilde{\mathbf{P}}_{[k]})\,dk, \tag{97}$$

$$\bar{\mathbf{P}}_{k_{t+l}}^{k_t} = \tilde{\mathbf{P}}_{k_t} + \int_{k_t}^{k_{t+l}} f(\bar{\mathbf{P}}_{[k]}^{k_t})\,dk + \int_{k_t}^{k_{t+l}} (f(\bar{\mathbf{P}}_k^{k_t}) - f(\bar{\mathbf{P}}_{[k]}^{k_t}))dk. \tag{98}$$

This yields the following bound:

$$\sup_{0\leq j\leq l} \|\tilde{\mathbf{P}}_{k_{t+j}} - \bar{\mathbf{P}}_{k_{t+j}}^{k_t}\| \leq c_1\eta(1 + \|\tilde{\mathbf{P}}_{k_t}\|) + \eta L\|\mathbf{A}\|^2 \sum_{m=0}^{l-1} \sup_{j\leq m} \|\tilde{\mathbf{P}}_{k_{t+j}} - \bar{\mathbf{P}}_{k_{t+j}}^{k_t}\|, \tag{99}$$

where $c_1 > 0$ is a constant. By applying Grönwall's inequality, it follows that $\sup_{t\leq j\leq t+N-1} \|\tilde{\mathbf{P}}_{k_j} - \bar{\mathbf{P}}_{k_j}^{k_t}\|^2 \leq c_2\eta$ and $\sup_{k\in[0,L]} \|\tilde{\mathbf{P}}_{l+k} - \bar{\mathbf{P}}_{l+k}^{l}\| \leq c_3\eta$ for some constants $c_2, c_3 > 0$. This completes the proof.

G. Proof of Theorem 5

The convergence of Algorithm 1 can be established by analyzing the virtual fixed-point process $\{\widehat{\mathbf{P}}^t, t \geq 1\}$ under an arbitrarily small step size $\xi \to 0$. Define the mapping $\widetilde{g}(\mathbf{P}) = \mathbf{P} + \xi f(\mathbf{P})$, which corresponds to the iterative update $\widehat{\mathbf{P}}^{t+1} = \widetilde{g}(\widehat{\mathbf{P}}^t)$. As shown in Appendix C, the fixed-point equation $\widetilde{\mathbf{P}}^* = \widetilde{g}(\mathbf{P}^*)$ admits a unique solution $\mathbf{P}^*$. According to Section 3.4 of [34], the operator $\mathbf{P} \mapsto \mathbf{P} + \xi f(\mathbf{P})$ is contractive. Following the bounding technique from Appendix C, we construct two matrix bounds: a lower bound $\widehat{\mathbf{P}}^{(1)} = \mathbf{0}_S$ satisfying $\widehat{\mathbf{P}}^{(1)} \preceq \widetilde{g}(\widehat{\mathbf{P}}^{(1)})$, and an upper bound $\widehat{\mathbf{P}}^{(2)} \prec \infty$ such that $\widehat{\mathbf{P}}^{(2)} \succeq \widetilde{g}(\widehat{\mathbf{P}}^{(2)})$. Then for $t = 1, 2, ...$, define the following two matrix sequences:

$$\widehat{\mathbf{P}}^{(1)}(t+1) = \widetilde{g}(\widehat{\mathbf{P}}^{(1)}(t)), \quad \widehat{\mathbf{P}}^{(1)}(1) = \mathbf{0}_S, \tag{100}$$

$$\widehat{\mathbf{P}}^{(2)}(t+1) = \widetilde{g}(\widehat{\mathbf{P}}^{(2)}(t)), \quad \widehat{\mathbf{P}}^{(2)}(1) = \widehat{\mathbf{P}}^{(2)}. \tag{101}$$

By construction, the virtual iteration satisfies $\widehat{\mathbf{P}}^{(1)}(t) \preceq \widehat{\mathbf{P}}^t \preceq \widehat{\mathbf{P}}^{(2)}(t)$. Taking the limit as $t \to \infty$ and using the uniqueness of $\mathbf{P}^*$, we obtain $\limsup_{t\to\infty} \widehat{\mathbf{P}}^t = \mathbf{P}^*$. Consequently, the associated functions $\widetilde{V}^t(\mathbf{x}(t))$, $\mathbf{u}_1(t)$, and $\mathbf{u}_2(t)$ converge almost surely to their optimal values. This concludes the proof of Theorem 5.

REFERENCES

[1] Z. Liu, W. Zhan, X. Liu, Y. Zhu, M. Qi, J. Leng, L. Wei, S. Han, X. Wu, and X. Yan, "A wireless controlled robotic insect with ultrafast untethered running speeds," *Nat. Commun*, vol. 15, no. 1, p. 3815, 2024.

[2] T. Wigren, K. Lau, R. A. Delgado, and R. H. Middleton, "Globally stable delay alignment for feedback control over wireless multipoint connections," *IEEE Trans. Control Netw. Syst.*, vol. 7, no. 4, pp. 1633–1642, 2020.

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

[4] R. Qian, Z. Duan, Y. Qi, T. Peng, and W. Wang, "Formation-control stability and communication capacity of multiagent systems: A joint analysis," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 2, pp. 917–927, 2020.

[5] H. Xu, J. Zhang, Z. Sun, and H. Yang, "Event-based wireless tracking control for a wheeled mobile robot against reactive jamming attacks," *IEEE Trans. Control Netw. Syst.*, vol. 10, no. 4, pp. 1925–1936, 2023.

[6] D. Liuzza, D. V. Dimarogonas, and K. H. Johansson, "Generalized PID synchronization of higher order nonlinear systems with a recursive lyapunov approach," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 4, pp. 1608–1621, 2017.

[7] P. Duan, L. He, Z. Duan, and L. Shi, "Distributed cooperative LQR design for multi-input linear systems," *IEEE Trans. Control Netw. Syst.*, vol. 10, no. 2, pp. 680–692, 2022.

[8] X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang, X. Xu, B. Dai, and Q. Miao, "Deep reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 4, pp. 5064–5078, 2024.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.*, vol. 11, no. 3, pp. 706–714, 2014.

[14] S. A. A. Rizvi and Z. Lin, "Output feedback adaptive dynamic programming for linear differential zero-sum games," *Automatica*, vol. 122, p. 109272, 2020.

[15] 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.

[16] J. Song, S. He, Z. Ding, and F. Liu, "A new iterative algorithm for solving H infinity 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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. Syst.*, 2021.

[21] M. Tang and V. K. Lau, "Online learning algorithms for zero-sum games of linear systems over wireless MIMO fading channels with uncountable state space," *in Proc. 62nd IEEE Conf. Decis. Control*, pp. 8751–8756, 2023.

[22] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. 47th Annu. Allerton Conf. Commun., Control*. IEEE, 2009, pp. 911–918.

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

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

[25] F. Zhao, W. Gao, T. Liu, and Z.-P. Jiang, "Event-triggered robust adaptive dynamic programming with output feedback for large-scale systems," *IEEE Trans. Control Netw. Syst.*, vol. 10, no. 1, pp. 63–74, 2022.

[26] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.

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

[28] X. Rao and V. K. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, 2014.

[29] A. Jalilzadeh, F. Yousefian, and M. Ebrahimi, "Stochastic approximation for estimating the price of stability in stochastic nash games," *ACM Trans. Model. Comput. Simul.*, vol. 34, no. 2, pp. 1–24, 2024.

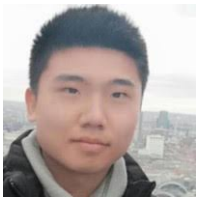
[30] J. Lei and U. V. Shanbhag, "Stochastic nash equilibrium problems: Models, analysis, and algorithms," *IEEE Control Syst. Mag.*, vol. 42, no. 4, pp. 103–124, 2022.

[31] Y. Yang, Z. Guo, H. Xiong, D.-W. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3735–3747, 2019.

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

[33] S. Cai and V. K. Lau, "Zero MAC latency sensor networking for cyber-physical systems," *IEEE Trans. Signal Process.*, vol. 66, no. 14, pp. 3814–3823, 2018.

[34] P. J. West, *Algorithms and software for solving coupled discrete-time riccati equations via the las language (computer-aided design, control, game theory)*. University of Illinois at Urbana-Champaign, 1986.



Minjie Tang (Member, IEEE) received the B.Eng. degree in information and communication engineering from The Huazhong University of Science and Technology (HUST), Wuhan, China, in 2018, and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2024. He is currently a Postdoctoral Research Fellow at Communication Systems Department, EURECOM, France. His current research interests include semantic communications for control, reinforcement learning, networked control systems and industrial IoT.

Vincent K. N. Lau (Fellow, IEEE) received the B.Eng. (Hons.) degree from The University of Hong Kong, in 1992, and the Ph.D. degree from Cambridge University in 1997. He was with Bell Labs from 1997 to 2004, and the Department of Electronic and Computer Engineering (ECE), The Hong Kong University of Science and Technology (HKUST), in 2004. Currently, he is the Chair Professor and the Founding Director of Huawei-HKUST Innovation Laboratory, HKUST. His research interests include robust and delay-optimal cross layer optimization for MIMO/OFDM wireless systems, interference mitigation techniques for wireless networks, massive MIMO, M2M, and network control systems.