Data-Driven Online Learning Algorithm for Optimal Linear Tracking Control over Unreliable Wireless MIMO Fading Channels

Minjie Tang*, Chenyuan Feng*, and Tony Q. S. Quek†

*Department of Communication Systems, EURECOM, France
† Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore
Emails: {Minjie.Tang, Chenyuan.Feng}@eurecom.fr, tonyquek@sutd.edu.sg

***Abstract*—This work explores the data-driven online tracking control problem for linear dynamic systems across multiple-input multiple-output (MIMO) fading channels. Initially, we address the optimal tracking control for a system with known plant dynamics, and design an innovative stochastic-approximation (SA)-based data-driven algorithm that leverage the instantaneous wireless channel state information (CSI). Subsequently, we extend this approach to accommodate unknown plant dynamics by proposing a novel normalized-stochastic-gradient-descent (NSGD)-based algorithm. This algorithm facilitates simultaneous system identification and control in an online setting using the real-time plant state as well as the CSI. Through Lyapunov drift analysis, we establish the asymptotic optimality of our proposed data-driven algorithms. Numerical results and analysis further demonstrate notable performance improvements compared to several leading learning techniques.**

***Index Terms*—Online training, linear tracking control, machine learning, MIMO communication, Unreliable transmission.**

I. INTRODUCTION

Optimal tracking control for linear systems via wireless networks has been a focal point of recent research, particularly due to its wide applications. This discipline is dedicated to the development of sophisticated remote tracking controllers that guide dynamic systems towards their desired states. A typical linear control system comprises three essential components: a potentially unstable *dynamic plant*, a *remote controller* and an *actuator* collocated with the *dynamic plant*, as shown in Fig. 1. The remote controller processes the real-time plant state and crafts the intermittent control signals. The control commands are then conveyed to the actuator via an unreliable wireless network, aiming to synchronize the internal plant states with predetermined target states. Despite this, the wireless network can introduce issues such as signal fading and packet loss, which may markedly diminish the tracking control efficacy of the system.

The design of tracking controllers for linear systems has been extensively studied [1], [2]. The linear system is modeled in the frequency domain, and offline proportional-integral-derivative (PID) control schemes are proposed [1]. These schemes rely on empirical PID coefficients, with control solutions generated at a remote controller based on the dynamic differences between the desired plant states and real-time states. However, PID control methods are inherently heuristic and lack robustness, making them suboptimal in many practical applications. To address these limitations and reduce reliance on trial-and-error tuning, online linear-quadratic-tracker (LQT) control approaches have been proposed in [2], where controllers and actuators are collocated. These LQT methods parameterize control actions using tracking control gains derived from solving the Riccati equation [3] in real-time, based on both the target and actual states of the dynamic systems. However, directly applying these LQT solutions becomes problematic for wireless communication. Such scenarios can lead to undesirable system behavior, as the tracking control gains need to account not only for the dynamic differences between target and real-time plant states but also for variations in wireless network conditions. Specifically, adapting the control gains to both wireless fading conditions and system dynamics is crucial for achieving accurate state trajectories. As a result, brute-force application of the LQT schemes from [1], [2] can result in significant mismatches between the desired and actual

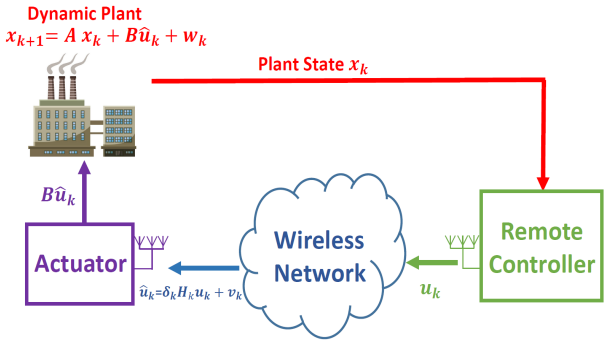

Fig. 1: Example of a linear system over the wireless channels.

This work was supported in part by the National Natural Science Foundation of China under Grant 62301328, and in part by the SNS JU project 6G-GOALS under the EU's Horizon program Grant Agreement No. 101139232.
Corresponding author: Chenyuan Feng.

plant states in wireless networked control systems.

Precise tracking control depends on accurate knowledge of the dynamic plant model, which is often unavailable in practice. Consequently, system identification of unknown dynamic plants becomes crucial for achieving high tracking performance. Traditional system identification methods for linear systems have been extensively studied in an offline setting, primarily through least-squares-based techniques [4]. However, these offline algorithms require exponentially large memory for sample storage, making them impractical for large-scale linear systems. To address the computational burden, online system identification methods have been developed, using recursive least squares [5] and projected online learning algorithms [6]. These methods reduce computational complexity by updating model parameters in real time. However, these approaches assume that controllers and actuators are collocated and do not account for plant noise, limiting their applicability in real-world scenarios. Applying these identification algorithms directly in wireless networked control systems will degrade identification performance due to the randomness in state samples used for system identification. To handle noisy data, stochastic gradient descent (SGD)-based algorithms have been widely used in parameter learning problems [7]. However, standard SGD-based approaches are unsuitable for identifying unstable dynamic plants, as they struggle with the unbounded variance of the state samples during training, leading to poor convergence and performance [8].

In this work, we propose a novel data-driven approach for online tracking control over wireless MIMO fading channels. The main contributions are as follows: ***(a) Data-Driven Online Tracking Control over the MIMO Fading Channels with Known Plant Dynamics:*** Leveraging the knowledge of plant dynamics and real-time channel state information (CSI), we develop an online tracking controller for linear systems using a novel stochastic approximation (SA)-based algorithm; ***(b) Data-Driven Online Tracking Control over the MIMO Fading Channels with Unknown Plant Dynamics:*** For scenarios where plant dynamics are unknown at the remote controller, we propose a new normalized stochastic gradient descent (NSGD)-based algorithm. This method enables simultaneous system identification and online learning of the optimal tracking control solution, utilizing real-time plant states and CSI; and ***(c) Optimal Convergence Performance Analysis:*** By providing Lyapunov-based analysis, we demonstrate that both the system identification and tracking control algorithms asymptotically achieve optimal performance.

II. System Model

A. Dynamic Plant Model

We consider a discrete-time linear system with S state variables, where the remote controller is equipped with N_t transmission antennas, and the actuator is equipped with N_r receiving antennas. The dynamics of the physical plant are:

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\widehat{\mathbf{u}}_k + \mathbf{w}_k, \quad k = 0, 1, 2, ..., \tag{1}$$

where $\mathbf{x}_k \in \mathbb{R}^{S\times 1}$ is the plant state variable, $\mathbf{A} \in \mathbb{R}^{S\times S}$ is the plant dynamics, $\mathbf{B} \in \mathbb{R}^{S\times N_r}$ is the control input matrix, $\widehat{\mathbf{u}}_k \in \mathbb{R}^{N_r\times 1}$ is the received control signal at the actuator and $\mathbf{w}_k \in \mathbb{R}^{S\times 1}$ is the plant noise with zero mean and finite noise covariance matrix $\mathbf{W} \in \mathbb{S}_+^S$. We assume the dynamic evolution (1) is potentially unstable [1]. The plant system $(\mathbf{A}, \mathbf{B})$ is assumed to be controllable.

B. Wireless MIMO Fading Channel Model

As depicted in Fig. 1, a wireless MIMO fading channel is considered between the remote controller and the actuator [9]. At each time slot k, the control signal received at the actuator, denoted by $\widehat{\mathbf{u}}_k \in \mathbb{R}^{N_r\times 1}$, is given by

$$\widehat{\mathbf{u}}_k = \delta_k \mathbf{H}_k \mathbf{u}_k + \mathbf{v}_k, \tag{2}$$

where $\mathbf{u}_k \in \mathbb{R}^{N_t\times 1}$ is the control action generated at the remote controller, and $\delta_k \in \{0, 1\}$ is the i.i.d. random access variable for the remote controller with $\Pr(\delta_k = 1) = p$. $\mathbf{H}_k \in \mathbb{R}^{N_r\times N_t} \sim \mathcal{N}(0, \mathbf{1}_{N_r})$ is the wireless MIMO fading matrix between the remote controller and the actuator, and $\mathbf{v}_k \sim \mathcal{N}(0, \mathbf{I}_{N_r})$ is the additive Gaussian noise.

III. Data-Driven Optimal Tracking Control for the Linear System with Known Plant Dynamics

The equivalent linear time-varying (LTV) system model can be derived by combining Eqs. (1) and (2), resulting in

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \delta_k \mathbf{B}\mathbf{H}_k \mathbf{u}_k + \mathbf{B}\mathbf{v}_k + \mathbf{w}_k. \tag{3}$$

Let $\mathbf{r}_k \in \mathbb{R}^{S\times 1}$ be the prior-known target state, which evolves according to $\mathbf{r}_{k+1} = \mathbf{G}\mathbf{r}_k$, where $\mathbf{G} \in \mathbb{R}^{S\times S}$ is the reference dynamics matrix. The goal of optimal tracking control for this LTV system can be framed as an ergodic tracking control problem over the aggregated state sequence $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, ...\}$, where the aggregated state at the k-th time slot is $\mathbf{S}_k = \{\mathbf{x}_k, \mathbf{r}_k, \delta_k \mathbf{H}_k\}$. The optimal tracking control problem can then be formulated as the following infinite-horizon ergodic optimization problem.

Problem 1 (Ergodic Optimal Tracking Control Problem):

$$\begin{aligned} \min_{\pi} \mathcal{J}^{\pi} = &\min_{\pi} \limsup_{K\to\infty} \frac{1}{K} \mathbb{E}[\sum_{k=1}^{K} \xi^k r_1(\mathbf{S}_k, \mathbf{u}_k)], \\ &s.t. \quad (3), \xi \in \mathbb{R}_+. \end{aligned} \tag{4}$$

where the control policy $\pi : \mathcal{S} \to \mathcal{U}$ is a mapping from $\mathbf{S}_k$ to $\mathbf{u}_k \in \mathcal{U}$. The per-stage reward function $r_1(\mathbf{S}_k, \mathbf{u}_k)$ is given by $r_1(\mathbf{S}_k, \mathbf{u}_k) = (\mathbf{x}_k - \mathbf{r}_k)^T \mathbf{Q}(\mathbf{x}_k - \mathbf{r}_k) + \mathbf{u}_k^T \mathbf{R}\mathbf{u}_k + (\delta_k \mathbf{H}_k \mathbf{u}_k)^T \mathbf{M}\mathbf{H}_k \mathbf{u}_k$. $\mathbf{Q} \in \mathbb{S}_+^S$, $\mathbf{R} \in \mathbb{S}_+^{N_t}$, and $\mathbf{M} \in \mathbb{S}_+^{N_r}$.

The following theorem summarizes the sufficient conditions for the existence of the solution to Problem 1.

Theorem *1: (The Sufficient Conditions for the Existence of the Solution to Problem 1)* Let the singular value decomposition (SVD) of $\delta_k \xi \mathbf{B}\mathbf{H}_k \left(\mathbf{H}_k^T \mathbf{M}\mathbf{H}_k + \mathbf{R}\right)^{-1} \mathbf{H}_k^T \mathbf{B}^T$ be $\delta_k \xi \mathbf{B}\mathbf{H}_k \left(\mathbf{H}_k^T \mathbf{M}\mathbf{H}_k + \mathbf{R}\right)^{-1} \mathbf{H}_k^T \mathbf{B}^T = \mathbf{V}_k^T \zeta_k \mathbf{V}_k$ with

[1] "The dynamic evolution (1) is potentially unstable" means that plant dynamics $\mathbf{A}$ contains possibly unstable eigenvalues, i.e., $\|\mathbf{A}\| > 1$.

the diagonal elements of ζ_k in descending order. Let $\text{rank}\left(\delta_k \mathbf{B}\mathbf{H}_k\mathbf{H}_k^T\mathbf{B}^T\right) = \gamma_k$ and $\mathbf{\Pi}_k = \begin{bmatrix}\mathbf{I}_{\gamma_k} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}_{S\times S}$. Problem 1 exists a unique solution if: (a) $0 < \xi\|\mathbf{G}\|^2 < 1$; and (b) $\|\mathbb{E}[\mathbf{A}^T\mathbf{V}_k^T\left(\mathbf{I}_S - \mathbf{\Pi}_k\right)\mathbf{V}_k\mathbf{A}]\| < \frac{1}{\xi}$.

Proof: Please refer to Appendix A. ∎

Note that the Problem 1 can be equivalently represented as a virtual ergodic optimal control problem as follows.

Problem 2 (Equivalent Virtual Ergodic Optimal Control Problem for Problem 1):

$$\begin{aligned}&\min_{\pi} \mathcal{J}^{\pi} = \min_{\pi} \limsup_{K\to\infty} \frac{1}{K}\mathbb{E}[\sum_{k=1}^{K} \xi^k r_2(\mathbf{S}_k, \mathbf{u}_k)]\\ &\text{s.t. } \widehat{\mathbf{x}}_{k+1} = \widehat{\mathbf{A}}\widehat{\mathbf{x}}_k + \widehat{\mathbf{B}}_k\mathbf{u}_k + \widehat{\mathbf{w}}_k,\end{aligned} \tag{5}$$

where the equivalent per-stage reward function for the remote controller $r_2(\mathbf{S}_k, \mathbf{u}_k)$ is given by $r_2(\mathbf{S}_k, \mathbf{u}_k) = \widehat{\mathbf{x}}_k^T\widehat{\mathbf{Q}}\widehat{\mathbf{x}}_k + \mathbf{u}_k^T\mathbf{R}\mathbf{u}_k + (\delta_k\mathbf{H}_k\mathbf{u}_k)^T\mathbf{M}\mathbf{H}_k\mathbf{u}_k$. $\widehat{\mathbf{x}}_k = \begin{bmatrix}\mathbf{x}_k\\ \mathbf{r}_k\end{bmatrix} \in \mathbb{R}^{2S\times 1}$, $\widehat{\mathbf{Q}} = \begin{bmatrix}\mathbf{Q} & -\mathbf{Q}\\ -\mathbf{Q} & \mathbf{Q}\end{bmatrix} \in \mathbb{R}^{2S\times 2S}$. $\widehat{\mathbf{A}} = \text{Diag}(\mathbf{A}, \mathbf{G}) \in \mathbb{R}^{2S\times 2S}$,$\widehat{\mathbf{B}}_k = \begin{bmatrix}\delta_k\mathbf{B}\mathbf{H}_k\\ \mathbf{0}_{S\times N_t}\end{bmatrix} \in \mathbb{R}^{2S\times N_t}$, and $\widehat{\mathbf{w}}_k = \begin{bmatrix}\mathbf{B}\mathbf{v}_k + \mathbf{w}_k\\ \mathbf{0}_{S\times 1}\end{bmatrix} \in \mathbb{R}^{2S\times 1}$. As a result, the solution to Problem 1 can be obtained via the solution of the *ergodic Bellman optimality equation* for Problem 2 as follows.

Theorem *2: (Ergodic Bellman Optimality Equation for Problem 1*) Under the sufficient conditions in Theorem 1, the ergodic Bellman optimality equation for Problem 2 is

$$\theta^* + V^*(\mathbf{S}_k) = \min_{\mathbf{u}_k}[r_2(\mathbf{S}_k, \mathbf{u}_k) + \xi\mathbb{E}[V^*(\mathbf{S}_{k+1})|\mathbf{S}_k, \mathbf{u}_k]], \tag{6}$$

where the expectation at R.H.S. of (6) is w.r.t. $\delta_{k+1}\mathbf{H}_{k+1}$, $\mathbf{w}_k$, and $\mathbf{v}_k$. $\theta^* = \mathcal{J}^* = \inf_{\pi}\mathcal{J}^{\pi}$ is the optimal average cost in Problem 2, and $V^*(\mathbf{S}_k)$ is the optimal value function for the remote controller.

Proof: Please refer to Chapter 6.7 of [10]. ∎

One might consider solving (6) using conventional iterative methods such as value iteration or Q-learning. However, these approaches encounter significant obstacles due to the "curse of dimensionality." This challenge stems from the necessity of learning the unstructured, high-dimensional value function $V^*(\mathbf{S}_k)$, which depends on the random variable $\mathbf{S}_k$ and spans an uncountable state space. To address these challenges, we leverage the fact that δ_k and $\mathbf{H}_k$ are i.i.d. random processes with respect to k, for $1 \le k \le K$. Based on this, we propose a structured approach by deriving an equivalent *reduced-state optimality equation*, which simplifies the problem and mitigates the dimensionality issues.

Theorem *3: (Structured Reduced-State Optimality Equation)* Under the sufficient conditions in Theorem 1, the optimal solution to Problem 2 is equivalent to the solution of the equivalent structured reduced-state optimality equation for Problem 2, which is given by

$$\begin{aligned}\widehat{\theta}^* + \widehat{V}^*(\widehat{\mathbf{x}}_k) = \mathbb{E}_{\delta_k\mathbf{H}_k}[&\min_{\mathbf{u}_k}[r_2(\mathbf{S}_k, \mathbf{u}_k)\\ &+ \xi\mathbb{E}_{\delta_{k+1}\mathbf{H}_{k+1},\mathbf{w}_k,\mathbf{v}_k}[\widehat{V}^*(\widehat{\mathbf{x}}_{k+1})|\mathbf{S}_k, \mathbf{u}_k]],\end{aligned} \tag{7}$$

where $\widehat{V}^*(\widehat{\mathbf{x}}_k) = \mathbb{E}_{\delta_k\mathbf{H}_k}[V^*(\mathbf{S}_k)|\mathbf{x}_k] = \widehat{\mathbf{x}}_k^T\mathbf{P}\widehat{\mathbf{x}}_k$ is the structured reduced-state value function and $\mathbf{P} \in \mathbb{S}^{2S}$. The optimal average cost $\widehat{\theta}^* = \theta^* = \inf_{\pi} J^{\pi} = \text{Tr}(\xi\mathbf{P}_{1:S}\mathbf{W} + \xi\mathbf{B}^T\mathbf{P}_{1:S}\mathbf{B})$, where $\mathbf{P}_{1:S}$ is the S-order leading principal submatrix of $\mathbf{P}$. The optimal control policy $\pi^* = \{\mathbf{u}_k^*, \forall k\}$, where $\mathbf{u}_k^*$ is the solution to Problem 2 given by

$$\mathbf{u}_k^* = -(\mathbf{R} + \mathbf{H}_k^T\mathbf{M}\mathbf{H}_k + \xi\widehat{\mathbf{B}}_k^T\mathbf{P}\widehat{\mathbf{B}}_k)^{-1}\xi\widehat{\mathbf{B}}_k^T\mathbf{P}\widehat{\mathbf{A}}\widehat{\mathbf{x}}_k. \tag{8}$$

Proof: Please refer to Appendix B. ∎

In contrast to solving (6) by learning the unstructured high-dimensional value function $V^*(\mathbf{S}_k)$, which encompasses an uncountable state space, solving (7) streamlines the process by focusing on the structured kernel $\mathbf{P}$ of the low-dimensional function $\hat{V}^*(\widehat{\mathbf{x}}_k)$. This shift in focus significantly reduces complexity and helps to circumvent the "curse of dimensionality." By leveraging the structural properties of $\widehat{V}^*(\widehat{\mathbf{x}}_k)$, $\widehat{\theta}^*$, and $\mathbf{u}_k^*$ as outlined in Theorem 3, the equivalent structured reduced-state optimality (7) can be expressed as $f(\mathbf{P}) = \mathbf{0}_{2S}$, where $f(\mathbf{P})$ is defined as follows:

$$\begin{aligned}f(\mathbf{P}) = \mathbb{E}_{\delta_k\mathbf{H}_k}[&\widehat{\mathbf{Q}} + \xi\widehat{\mathbf{A}}^T\mathbf{P}\widehat{\mathbf{A}} - \xi^2\widehat{\mathbf{A}}^T\mathbf{P}\widehat{\mathbf{B}}_k(\mathbf{R}+\\ &\mathbf{H}_k^T\mathbf{M}\mathbf{H}_k + \xi\widehat{\mathbf{B}}_k^T\mathbf{P}\widehat{\mathbf{B}}_k)^{-1}\widehat{\mathbf{B}}_k^T\mathbf{P}\widehat{\mathbf{A}}] - \mathbf{P}.\end{aligned} \tag{9}$$

To find the root of the equation $f(\mathbf{P}) = \mathbf{0}_{2S}$, we can employ the SA algorithm, as detailed in Algorithm 1[2].

Algorithm 1 Data-Driven Online Optimal Tracking Control for Linear Systems over MIMO Fading Channels

Initialization: Given a feasible positive semi-definite initial kernel value $\bar{\mathbf{P}}_1 \in \mathbb{S}^{2S}$, the initial estimated reduced-state value function is given by $\bar{V}_1(\widehat{\mathbf{x}}_1) = \widehat{\mathbf{x}}_1^T\bar{\mathbf{P}}_1\widehat{\mathbf{x}}_1$, and the estimated optimal tracking control solution at the initial timeslot is given by

$$\mathbf{u}_1 = -(\mathbf{R} + \mathbf{H}_1^T\mathbf{M}\mathbf{H}_1 + \xi\widehat{\mathbf{B}}_1^T\bar{\mathbf{P}}_1\widehat{\mathbf{B}}_1)^{-1}\xi\widehat{\mathbf{B}}_1^T\bar{\mathbf{P}}_1\widehat{\mathbf{A}}\widehat{\mathbf{x}}_1. \tag{10}$$

For $k = 2, 3, ...$

Step 1: (Update of the Reduced-State Value Function) Using $\bar{\mathbf{P}}_k$ updated at the $(k-1)$-th timeslot, the estimated reduced-state value function at the k-th timeslot is given by $\bar{V}_k(\widehat{\mathbf{x}}_k) = \widehat{\mathbf{x}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{x}}_k$.

Step 2: (Update of the Tracking Control Solution) Using $\bar{\mathbf{P}}_k$ updated at the $(k-1)$-th timeslot, the estimated optimal control solution at the k-th timeslot is given by

$$\mathbf{u}_k = -(\mathbf{R} + \mathbf{H}_k^T\mathbf{M}\mathbf{H}_k + \xi\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{B}}_k)^{-1}\xi\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{A}}\widehat{\mathbf{x}}_k \tag{11}$$

Step 3: (Update of the Learned $\bar{\mathbf{P}}_k$)
$\bar{\mathbf{P}}_{k+1}$ is updated using $\bar{\mathbf{P}}_k$ and $\delta_k\mathbf{H}_k$ given by

$$\begin{aligned}\bar{\mathbf{P}}_{k+1} = \bar{\mathbf{P}}_k + \alpha_k(&\widehat{\mathbf{Q}} + \xi\widehat{\mathbf{A}}^T\bar{\mathbf{P}}_k\widehat{\mathbf{A}} - \xi^2\widehat{\mathbf{A}}^T\bar{\mathbf{P}}_k\widehat{\mathbf{B}}_k(\mathbf{R}+\\ &\mathbf{H}_k^T\mathbf{M}\mathbf{H}_k + \xi\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{B}}_k)^{-1}\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{A}} - \mathbf{P}_k),\end{aligned} \tag{12}$$

where $\alpha_k > 0$ is the learning stepsize at k-th timeslot satisfying $\sum_{k=1}^{\infty}\alpha_k = \infty$, $\sum_{k=1}^{\infty}(\alpha_k)^2 < \infty$.

End

We formally summarize the convergence of Algorithm 1 in the following Theorem 4.

[2]In Steps 2 and 3 of Algorithm 1, the realization of the channel state $\delta_k\mathbf{H}_k$ is necessary. This can be obtained through standard channel estimation techniques at the actuator, which utilize the received pilot symbols from the controller and the subsequent channel feedback to the controller [11].

Theorem *4: (Almost Sure Convergence of Algorithm 1)* If the sufficient conditions in Theorem 1 are satisfied, then we have: (a) The learned $\bar{\mathbf{P}}_k$ via Algorithm 1 converges to the ground truth kernel $\mathbf{P}$ almost surely; (b) The learned $\bar{V}_k(\widehat{\mathbf{x}}_k)$ via Algorithm 1 converges to $\widehat{V}^*(\widehat{\mathbf{x}}_k) = \widehat{\mathbf{x}}_k^T\mathbf{P}\widehat{\mathbf{x}}_k$ almost surely; and (c) $\mathbf{u}_k$ via Algorithm 1 converges to $\mathbf{u}_k^*$ in (8) almost surely.

Proof: Please refer to Appendix C. ∎

Note that Algorithm 1 requires prior knowledge of the plant dynamics $\mathbf{A}$. In the subsequent sections, we will extend Algorithm 1 to accommodate scenarios with unknown plant dynamics. Specifically, we will demonstrate how both $\mathbf{A}$ and the optimal control input $\mathbf{u}_k^*$ can be learned simultaneously in an online manner.

IV. Online Optimal Tracking Control for the Linear System with Unknown Plant Dynamics

We propose an online remote system identification scheme at the remote controller using $\{\mathbf{x}_1, \mathbf{x}_2, ...\}$ and $\{\mathbf{u}_1, \mathbf{u}_2, ...\}$. The problem is formulated as follows.

Problem 3 (Identification of Plant Dynamics $\mathbf{A}$):

$$\min_{\widetilde{\mathbf{A}}} \limsup_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{w}_k,\mathbf{v}_k}[\|\mathbf{x}_{k+1} - \widetilde{\mathbf{A}}\mathbf{x}_k - \delta_k\mathbf{B}\mathbf{H}_k\mathbf{u}_k\|^2]. \tag{13}$$

Note that Problem 3 is a convex stochastic optimization problem, where $\widetilde{\mathbf{A}} = \mathbf{A}$ is the unique global optimal solutions. One may consider using the SGD algorithm to solve Problem 3. However, such algorithm cannot converge since the conditional variance $\mathbb{E}[\mathbf{x}_k\mathbf{x}_k^T|\widetilde{\mathbf{A}}_k]$ in the standard SGD update is not guaranteed to be bounded [12].

To tackle the convergence issue of SGD algorithm for system identification, we propose a novel NSGD algorithm. Specifically, let $\{\eta_k\}$ be the Lipschiz step-size sequence satisfying $\sum_{k=1}^{\infty}\eta_k = \infty$ and $\sum_{k=1}^{\infty}(\eta_k)^2 < \infty$. The learning step-size for plant dynamics $\widehat{\eta}_k$ is obtained by dynamic normalization on η_k according to the state realizations as: $\widehat{\eta}_k = \eta_k$ if $\|\mathbf{x}_{k-1}\|^2 < 1$ and $\widehat{\eta}_k = \frac{\eta_k}{\|\mathbf{x}_{k-1}\|^2}$ otherwise. Such state-dependent normalization ensures that the SGD update for system identification will be reduced by a smaller equivalent step size $\widehat{\eta}_k$ if the state $\mathbf{x}_k, \forall k \geq 1$, is drifting away. Based on the real-time plant state $\mathbf{x}_k$ and the CSI $\delta_k\mathbf{H}_k$ the update for identified $\widetilde{\mathbf{A}}_{k-1}$ at the k-th time slot is given by

$$\begin{aligned}\widetilde{\mathbf{A}}_k = \widetilde{\mathbf{A}}_{k-1} + \phi_k\widehat{\eta}_k(\mathbf{x}_k - \widetilde{\mathbf{A}}_{k-1}\mathbf{x}_{k-1} - \\ \delta_{k-1}\mathbf{B}\mathbf{H}_{k-1}\mathbf{u}_{k-1})\mathbf{x}_{k-1}^T, \quad \forall k \in \mathbb{Z}_+,\end{aligned} \tag{14}$$

where $\phi_k \in \{0, 1\}$ is an indicating function and $\phi_k = 1$ if and only if $\|\mathbb{E}[(\widetilde{\mathbf{A}}_k)^T\mathbf{V}_k^T(\mathbf{I} - \mathbf{\Pi}_k)\mathbf{V}_k\widetilde{\mathbf{A}}_k]\| < \frac{1}{\xi}$ and $\|\widetilde{\mathbf{A}}_k\| < \epsilon$ with $\epsilon \in \mathbb{R}_+$ being a finite prior-given truncation constant satisfying $\|\mathbf{A}\| << \epsilon < \infty$. $\mathbf{V}_k \in \mathbb{R}^{S\times S}$ and $\mathbf{\Pi}_k \in \mathbb{R}^{S\times S}$ are defined according to Theorem 1.

The identified $\widetilde{\mathbf{A}}_k$ via (14) can be applied to (11) to learn the optimal tracking control solution. We formally summarize the simultaneous online system identification and tracking control algorithm in the following Algorithm 2.

Algorithm 2 Online Identification and Tracking Control for Linear Systems over MIMO Fading Channels.

Initialization: The identified plant dynamics is initialized as $\widetilde{\mathbf{A}}_1 \in \mathbb{R}^{S\times S}$ with the condition for $\phi^1 = 1$ satisfied. The learned reduced-state value function is initialized as $\bar{V}_1(\widehat{\mathbf{x}}_1) = \widehat{\mathbf{x}}_1^T\bar{\mathbf{P}}_1\widehat{\mathbf{x}}_1$, where $\bar{\mathbf{P}}_1 \in \mathbb{S}^{2S}$ is a positive semi-definite matrix, $\widehat{\mathbf{x}}_1 = [\mathbf{x}_1^T, \mathbf{r}_1^T]^T \in \mathbb{R}^{2S\times 1}$ is the initial plant state, The control solution is initialized as

$$\mathbf{u}_1 = -(\mathbf{R} + \mathbf{H}_1^T\mathbf{M}\mathbf{H}_1 + \xi\widehat{\mathbf{B}}_1^T\bar{\mathbf{P}}_1\widehat{\mathbf{B}}_1)^{-1}\xi\widehat{\mathbf{B}}_1^T\bar{\mathbf{P}}_1\dot{\mathbf{A}}_1\widehat{\mathbf{x}}_1, \tag{15}$$

where $\dot{\mathbf{A}}_1 = \mathrm{Diag}(\widetilde{\mathbf{A}}_1, \mathbf{G}) \in \mathbb{R}^{2S\times 2S}$.

Step 1: (Update of the Identified Plant Dynamics) At k-th time slot, the identified plant dynamics $\widetilde{\mathbf{A}}_k$ is obtained via (14) at the remote controller.

Step 2: (Update of the Learned Reduced-State Value Function) Using the learned kernel of the reduced-state value function $\bar{\mathbf{P}}_k$ updated at $(k-1)$-th time slot, the learned reduced-state value function at k-th time slot is given by $\bar{V}_k(\mathbf{x}_k) = \widehat{\mathbf{x}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{x}}_k$.

Step 3: (Update of the Tracking Control Solution $\mathbf{u}_k$) The tracking control solution $\mathbf{u}_k$ at k-th time slot is given by

$$\mathbf{u}_k = -(\mathbf{R} + \mathbf{H}_k^T\mathbf{M}\mathbf{H}_k + \xi\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{B}}_k)^{-1}\xi\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\dot{\mathbf{A}}_k\widehat{\mathbf{x}}_k \tag{16}$$

where $\dot{\mathbf{A}}_k = \mathrm{Diag}(\widetilde{\mathbf{A}}_k, \mathbf{G})$, and $\bar{\mathbf{P}}_k$ is updated based on the real-time plant state $\mathbf{x}_k$ and the CSI $\delta_k\mathbf{H}_k$ as follows.

$$\begin{aligned}\bar{\mathbf{P}}_{k+1} = \bar{\mathbf{P}}_k + \alpha_k(\widehat{\mathbf{Q}} + \xi\dot{\mathbf{A}}_k^T\bar{\mathbf{P}}_k\dot{\mathbf{A}}_k - \xi^2\dot{\mathbf{A}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{B}}_k(\mathbf{R} + \\ \mathbf{H}_k^T\mathbf{M}\mathbf{H}_k + \xi\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{B}}_k)^{-1}\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\dot{\mathbf{A}}_k - \mathbf{P}_k),\end{aligned} \tag{17}$$

Let $k = k + 1$ and go to Step 1.

The convergence of Algorithm 2 can be achieved by the convergence of $\widetilde{\mathbf{A}}_k$ and $\bar{\mathbf{P}}_k$. We summarize the almost sure convergence of Algorithm 2 in the following theorem.

Theorem *5: (Convergence of Algorithm 2)* (a) $\widetilde{\mathbf{A}}_k$ via Algorithm 2 converges to $\mathbf{A}$ almost surely; (b) $\bar{\mathbf{P}}_k$ via Algorithm 2 converges to $\mathbf{P}$ almost surely; (c) $\bar{V}_k(\widehat{\mathbf{x}}_k)$ via Algorithm 2 converges to $\widehat{V}^*(\widehat{\mathbf{x}}_k) = \widehat{\mathbf{x}}_k^T\mathbf{P}\widehat{\mathbf{x}}_k$ almost surely; (d) $\mathbf{u}_k$ via Algorithm 2 converges to $\mathbf{u}_k^*$ in (8) almost surely.

Proof: Please refer to Appendix C. ∎

V. Numerical Results

To evaluate the performance of the proposed online learning algorithms, we compare them with the following baselines:

- **Baseline 1:** *(Known Plant Dynamics and Known Optimal Tracking Control Solutions [13])* $\mathbf{A}$ and $\mathbf{u}_k^*$ are known at remote controller. The remote controller outputs $\mathbf{u}_k^*$ at each timeslot.
- **Baseline 2:** *(Brute-Force Value-Iteration-based Tracking Control with Known Plant Dynamics [14])* Based on $\mathbf{x}_k$, the remote controller generates $\mathbf{u}_k$ via brute-force value iteration on (6) with the knowledge of $\mathbf{A}$.
- **Baseline 3:** *(Brute-Force Value-Iteration-based Tracking Control with SGD-based System Identification [15])* $\mathbf{A}$ is identified by standard SGD-based approach. The remote controller generates $\mathbf{u}_k$ via value iteration on (6) with the identified plant dynamics.

We consider a linear system paramterized by $\mathbf{A} = \begin{bmatrix}1.001 & 0.026 & 0.015\\ 0.002 & 0.9 & 0.03\\ 0.0025 & 0.004 & 0.985\end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix}1.61 & 1.43\\ 1.67 & 1.17\\ 1.77 & 1.42\end{bmatrix}$. $p = 0.8$. $N_t =$

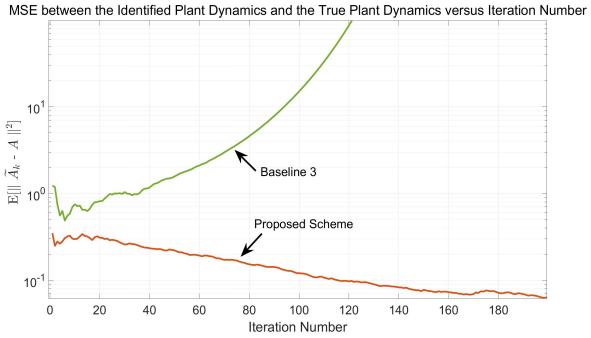

Fig. 2: MSE between the identified plant dynamics and the true plant dynamics.

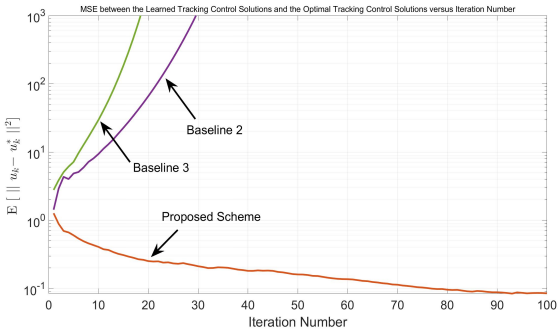

Fig. 3: MSE between the learned control solutions and the optimal control solutions.

$N_r = 2$. $\mathbf{R} = \mathbf{M} = 10^{-4}\mathbf{I}_2$, $\mathbf{Q} = 10^2\mathbf{1}_3$, $\gamma = 0.9$. $\mathbf{W} = 5\mathbf{I}_3 \in \mathbb{S}_+^3$, $\mathbf{G} = \mathbf{I}_3$, and $\mathbf{r}_1 = [100; 100; 100]$.

A. Convergence Performance for System Identification

Fig. 2 illustrates the mean square error (MSE) between the learned plant dynamics and the true plant dynamics as a function of the iteration number. As depicted in Fig. 2 , the learned plant dynamics using Baseline 3 do not converge to the true plant dynamics. This lack of convergence is attributed to the SGD-based identification algorithm employed in Baseline 3, which fails to ensure bounded conditional variance. Consequently, the algorithm does not achieve convergence. In contrast, the proposed scheme demonstrates asymptotic convergence in learning the plant dynamics, thanks to its normalized iterative update law.

B. Convergence Performance for Tracking Control

Fig. 3 reveals that the learned tracking control solutions obtained from Baseline 2 and Baseline 3 diverge from the optimal tracking solutions due to the "curse of dimensionality." In contrast, our proposed scheme asymptotically converges to the optimal tracking solutions, demonstrating its effectiveness in overcoming these challenges.

C. Stability Performance

Fig. 4 indicates that the plant states produced by Baseline 2 and Baseline 3 fail to track the target state due to their inability

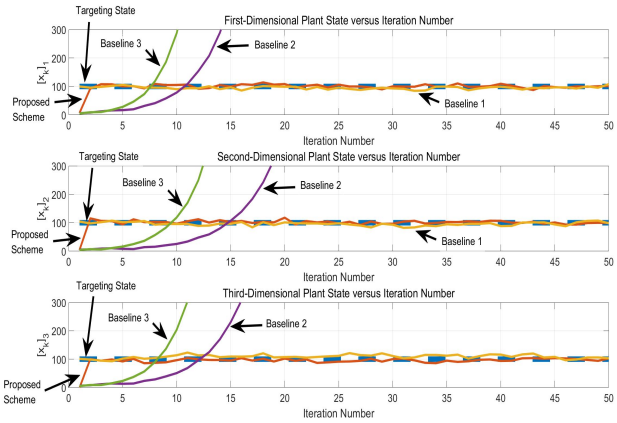

Fig. 4: Stability performance comparison.

to learn the optimal tracking control solutions. In contrast, the plant state achieved through our proposed scheme asymptotically converges to the target state over time, reflecting that the optimal tracking control solution is effectively attainable.

VI. Conclusion

This work addressed the data-driven online optimal tracking control problem for linear systems operating under wireless MIMO channels, considering both known and unknown plant dynamics. We proposed novel data-driven online learning algorithms for identifying plant dynamics and determining optimal tracking control solutions. Utilizing Lyapunov drift analysis, we demonstrated that the learned plant dynamics and tracking control solutions converge to their true values and optimal solutions, respectively, through the proposed learning algorithms. Numerical simulations further confirmed the superiority of our approach compared to various existing methods.

Appendix

A. Proof of Theorem 1

Problem 2 can be solved via the Markov decision process (MDP) techniques [13]. Specifically, using the similar approach as in [16], the optimality equation associated with Problem 2 can be represented as $\widehat{\theta}^* + \widehat{\mathbf{x}}_k^T\mathbf{P}\widehat{\mathbf{x}}_k = \mathbb{E}[\min_{\mathbf{u}_k}[\widehat{\mathbf{x}}_k^T\widehat{\mathbf{Q}}\widehat{\mathbf{x}}_k + \mathbf{u}_k^T(\mathbf{R} + \mathbf{H}_k^T\mathbf{M}\mathbf{H}_k)\mathbf{u}_k + \xi(\widehat{\mathbf{A}}\widehat{\mathbf{x}}_k + \widehat{\mathbf{B}}_k\mathbf{u}_k)^T\mathbf{P}(\widehat{\mathbf{A}}\widehat{\mathbf{x}}_k + \widehat{\mathbf{B}}_k\mathbf{u}_k) + \xi\mathrm{Tr}(\mathbf{P}_{1:S}\mathbf{W}) + \xi\mathrm{Tr}(\mathbf{B}^T\mathbf{P}_{1:S}\mathbf{B})]$, and $\mathbf{u}_k^*$ that achieves the minimum value of above equation is given by (8). Assuming $\widehat{V}^*(\widehat{\mathbf{x}}_k)$ exists, i.e., $\mathbf{P}$ exists, it follows that $\widehat{\theta}^* = \mathrm{Tr}\left(\xi\mathbf{P}_{1:S}\mathbf{W} + \xi\mathbf{B}^T\mathbf{P}_{1:S}\mathbf{B}\right)$. As a result, it suffices to prove that the solution $\mathbf{P}$ to the optimality equation for Problem 2 exists, then the solutions of $\widehat{\theta}^*$, $\widehat{V}^*(\widehat{\mathbf{x}}_k)$ and $\mathbf{u}_k^*$ to Problem 2 all exist.

We now prove the existence of unique $\mathbf{P}$. Specifically, note $f(\mathbf{P}) = \mathbf{0}_{2S}$ has all terms positive semi-definite. Hence, we can simultaneously diagonalize all terms for $f(\mathbf{P}) = \mathbf{0}_{2S}$. This follows that equations $\mathbf{P}_1 = \mathbf{Q} + \xi\mathbf{A}^T\mathbf{P}_1\mathbf{A} - \mathbb{E}[\xi^2\delta_k\mathbf{A}^T\mathbf{P}_1\mathbf{B}\mathbf{H}_k(\mathbf{R} + \mathbf{H}_k^T\mathbf{M}\mathbf{H}_k + \xi\mathbf{H}_k^T\mathbf{B}^T\mathbf{P}_1\mathbf{B}\mathbf{H}_k)^{-1}\mathbf{H}_k^T\mathbf{B}^T\mathbf{P}_1\mathbf{A}]$ and $\mathbf{P}_2 = \xi\mathbf{G}^T\mathbf{P}_2\mathbf{G} + \mathbf{Q}$

with unique positive semi-definite root $\mathbf{P}_1$ and $\mathbf{P}_2$ exists, respectively. It is easy to see that unique positive semi-definite $\mathbf{P}_2$ exists if $0 < \xi\|\mathbf{G}\|^2 < 1$. We now prove that unique positive semi-definite $\mathbf{P}_1$ exists under the sufficient conditions in Theorem 1. Specifically, by some derivations, we have $\mathbf{P}_1 \leq \xi\mathbf{A}^T\mathbb{E}[\mathbf{V}_k^T(\mathbf{I}-\mathbf{\Pi}_k)\mathbf{V}_k\mathbf{P}_1\mathbf{V}_k^T(\mathbf{I}-\mathbf{\Pi}_k)\mathbf{V}_k]\mathbf{A} + \bar{\mathbf{Q}}$, where $\bar{\mathbf{Q}} \in \mathbb{S}_+^S$ is a finite positive definite matrix. Under the sufficient conditions in Theorem 1, we use the monotonicity of the R.H.S. of the optimality equation w.r.t. $\mathbf{P}_1$, it follows that there is a unique positive semi-definite $\mathbf{P}_1$ such that the optimality equation is satisfied, which concludes the proof.

B. Proof of Theorem 3

Exploiting the i.i.d. property of $\mathbf{H}_k$ and δ_k, the optimality equation for Problem 2 in Theorem 2 can be represented as $\theta^* + V^*(\widehat{\mathbf{x}}_k, \delta_k\mathbf{H}_k) = \min_{\mathbf{u}_k}[r_2(\widehat{\mathbf{x}}_k, \delta_k\mathbf{H}_k, \mathbf{u}_k) + \xi\sum_{\widehat{\mathbf{x}}_{k+1}}\Pr(\widehat{\mathbf{x}}_{k+1}|\widehat{\mathbf{x}}_k, \delta_k\mathbf{H}_k, \mathbf{u}_k)\widehat{V}^*(\widehat{\mathbf{x}}_{k+1})]$. Taking the expectation of both sides of above equation over $\delta_k\mathbf{H}_k$, it follows (7) and concludes the proof.

C. Proof of Theorem 4 and Theorem 5

Note that Algorithm 1 is a special case of Algorithm 2. This follows to analyze the convergence of $\widetilde{\mathbf{A}}_k$ and $\bar{\mathbf{P}}_k$ in Algorithm 2 for convergence of both algorithms.

1) Convergence of $\widetilde{\mathbf{A}}_k$: We define the Lyapunov function as $V_k = \|\mathbf{A} - \widetilde{\mathbf{A}}_k\|_F^2$, with the associated Lyapunov drift given by $\Lambda(\widetilde{\mathbf{A}}_k) = \mathbb{E}[V_{k+1} - V_k|\widetilde{\mathbf{A}}_k]$. Substituting (14) into $\Lambda(\widetilde{\mathbf{A}}_k)$, and according to Lemma 2.1 of [17], it follows that $\lim_{k\to\infty}\mathbb{E}[\|\widetilde{\mathbf{A}}_k - \mathbf{A}\|_F^2] = 0$.

2) Convergence of $\bar{\mathbf{P}}_k$: The update rule for $\bar{\mathbf{P}}_k$ in (17) can be represented as $\bar{\mathbf{P}}_{k+1} = \bar{\mathbf{P}}_k - \alpha_k\widehat{g}(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k) = \bar{\mathbf{P}}_k - \alpha_k(f(\bar{\mathbf{P}}_k) + \mathcal{M}_k)$, where $\widehat{g}(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k) = \bar{\mathbf{P}}_k - g(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k)$, $f(\bar{\mathbf{P}}_k) = \mathbb{E}[\widehat{g}(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k)|\bar{\mathbf{P}}_k]$, $\mathcal{M}_k = \widehat{g}(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k) - f(\bar{\mathbf{P}}_k)$, $g(\bar{\mathbf{P}}_k) = \widehat{\mathbf{Q}} + \xi\dot{\mathbf{A}}_k^T\bar{\mathbf{P}}_k\dot{\mathbf{A}}_k - \xi^2\dot{\mathbf{A}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{B}}_k(\mathbf{R} + \mathbf{H}_k^T\mathbf{M}\mathbf{H}_k + \xi\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\widehat{\mathbf{B}}_k)^{-1}\widehat{\mathbf{B}}_k^T\bar{\mathbf{P}}_k\dot{\mathbf{A}}_k$. It can be verified that $\mathbb{E}[\|\bar{\mathbf{P}}_k\|_F^2] < \infty, \forall k$ and $\mathbb{E}[\mathrm{Tr}(\bar{\mathbf{P}}_k)] < \infty, \forall k$. Using the boundness of $\mathbb{E}[\|\bar{\mathbf{P}}_k\|_F^2]$ and the linearity of the trace functional on the monotonic $\widehat{g}^T(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k)$ w.r.t. $\bar{\mathbf{P}}_k$ [3], it follows that $\mathbb{E}[\|\bar{\mathbf{P}}_{k+1} - \mathbf{P}\|_F^2] = \mathbb{E}[\mathrm{Tr}((\bar{\mathbf{P}}_{k+1} - \mathbf{P})^T(\bar{\mathbf{P}}_{k+1} - \mathbf{P}))] = \mathbb{E}[\|\bar{\mathbf{P}}_{k+1} - \mathbf{P}\|_F^2] - 2\alpha_k\mathbb{E}[\mathrm{Tr}(\widehat{g}^T(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k)(\bar{\mathbf{P}}_k - \mathbf{P}))] + (\alpha_k)^2\mathbb{E}[\|\widehat{g}(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k)\|_F^2] \leq \mathbb{E}[\|\bar{\mathbf{P}}_k - \mathbf{P}\|_F^2] - 2\alpha_k\mathbb{E}[\mathrm{Tr}((\widehat{g}^T(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k) - \widehat{g}^T(\mathbf{P}, \mathbf{A}, \delta_k\mathbf{H}_k))(\bar{\mathbf{P}}_k - \mathbf{P}))] + (\alpha_k)^2c_1 \leq \mathbb{E}[\|\bar{\mathbf{P}}_k - \mathbf{P}\|_F^2] - 2\alpha_k\mathbb{E}[\mathrm{Tr}((\widehat{g}^T(\bar{\mathbf{P}}_k, \mathbf{A}, \delta_k\mathbf{H}_k) + \widehat{g}^T(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k) - \widehat{g}^T(\bar{\mathbf{P}}_k, \mathbf{A}, \delta_k\mathbf{H}_k) - \widehat{g}^T(\mathbf{P}, \mathbf{A}, \delta_k\mathbf{H}_k))(\bar{\mathbf{P}}_k - \mathbf{P}))] + (\alpha_k)^2c_2 \leq (1 - \alpha_kc_3)\mathbb{E}[\|\bar{\mathbf{P}}_k - \mathbf{P}\|_F^2] + (\alpha_k)^2c_4 + 2\alpha_k\mathbb{E}[\mathrm{Tr}((\widehat{g}^T(\bar{\mathbf{P}}_k, \mathbf{A}, \delta_k\mathbf{H}_k) - \widehat{g}^T(\bar{\mathbf{P}}_k, \widetilde{\mathbf{A}}_k, \delta_k\mathbf{H}_k))(\bar{\mathbf{P}}_k - \mathbf{P}))] \leq (1 - \alpha_kc_3)\mathbb{E}[\|\bar{\mathbf{P}}_k - \mathbf{P}\|_F^2] + (\alpha_k)^2c_4 + 2\alpha_k\mathbb{E}[\mathrm{Tr}((\mathbf{A}\mathbf{A}^T - \widetilde{\mathbf{A}}_k(\widetilde{\mathbf{A}}_k)^T)\widehat{g}^T(\bar{\mathbf{P}}_k, \mathbf{I}_{2S}, \delta_k\mathbf{H}_k)(\bar{\mathbf{P}}_k - \mathbf{P}))]$, where $0 < c_1, c_2, c_3, c_4 < \infty$. Further recursively applying the Von-Neumann's trace Inequality to the above inequality, we have $\mathbb{E}[\|\bar{\mathbf{P}}_{k+1} - \mathbf{P}\|_F^2] \leq (1 - \alpha_kc_5)\mathbb{E}[\|\bar{\mathbf{P}}_k - \mathbf{P}\|_F^2] + (\alpha_k)^2c_6 + \alpha_kc_7\mathbb{E}[\sum_{i=1}^S\sigma_s(\mathbf{A} - \widetilde{\mathbf{A}}_k)]$, where $0 < c_5, c_6, c_7 < \infty$, and $\sigma_i(\mathbf{A} - \widetilde{\mathbf{A}}_k)$ is the i-th decreasing-ordered singular value of $\mathbf{A} - \widetilde{\mathbf{A}}_k$.

The derivations in 1) shows that the sequence $\lim_{k\to\infty}\mathbb{E}[\sum_{i=1}^S\sigma_s^2(\mathbf{A} - \widetilde{\mathbf{A}}_k)] = 0$. According to the definition of the convergent sequence, it follows that for arbitrary given $c_8 > 0$, we have a finite timeslot $k_2 = 1, 2, ...,$ such that for all $k > k_2$, $\mathbb{E}[\sum_{i=1}^S\sigma_s(\mathbf{A} - \widetilde{\mathbf{A}}_k)] < c_8$. In other words, $\left\{\mathbb{E}[\sum_{i=1}^S\sigma_s(\mathbf{A} - \widetilde{\mathbf{A}}_k)]\right\}$ is also a convergent sequence that converges to 0. According to the definition of the convergent sequence that converges to 0, it follows that $\limsup_{K\to\infty}\mathbb{E}[\|\bar{\mathbf{P}}_k - \mathbf{P}\|_F^2] = 0$, which concludes the proof.

REFERENCES

[1] M. F. Miranda and K. G. Vamvoudakis, "Online optimal auto-tuning of pid controllers for tracking in a special class of linear systems," in *American Control Conf. (ACC)*, 2016, pp. 5443–5448.

[2] R. Moghadam and F. L. Lewis, "Output-feedback h quadratic tracking control of linear systems using reinforcement learning," *Int. J. Adaptive Control Signal Process.*, vol. 33, no. 2, pp. 300–314, 2019.

[3] R. R. Bitmead and M. Gevers, "Riccati difference and differential equations: Convergence, monotonicity and stability," in *The Riccati Equation*. Springer, 1991, pp. 263–291.

[4] M. Verhaegen and V. Verdult, *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.

[5] X. Wang, J. Zhou, S. Mou, and M. J. Corless, "A distributed algorithm for least squares solutions," *IEEE Trans. Autom. Control*, vol. 64, no. 10, pp. 4217–4222, 2019.

[6] Y. Kopsinis, K. Slavakis, and S. Theodoridis, "Online sparse system identification and signal reconstruction using projections onto weighted l-1 balls," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 936–952, 2010.

[7] K. Huang and S. Pu, "Improving the transient times for distributed stochastic gradient methods," *IEEE Trans. Autom. Control*, 2022.

[8] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[9] S. Guo, J. Ye, K. Qu, and S. Dang, "Green holographic mimo communications with a few transmit radio frequency chains," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 1, pp. 90–102, 2024.

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

[11] R. Ganesh, J. J. Kumari *et al.*, "A survey on channel estimation techniques in mimo-ofdm mobile communication systems," *Int. J. scientific engineering research*, vol. 4, no. 5, pp. 1850–1855, 2013.

[12] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, "Sgd: General analysis and improved rates," in *Int. conf. machine learning (ICML)*. PMLR, 2019, pp. 5200–5209.

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

[14] B. Pang and Z.-P. Jiang, "Adaptive optimal control of linear periodic systems: An off-policy value iteration approach," *IEEE Trans. Autom. Control*, vol. 66, no. 2, pp. 888–894, 2020.

[15] G. H. Elkaim, "System identification-based control of an unmanned autonomous wind-propelled catamaran," *Control Engineering Practice*, vol. 17, no. 1, pp. 158–169, 2009.

[16] B. Kiumarsi, F. L. Lewis, M.-B. Naghibi-Sistani, and A. Karimpour, "Optimal tracking control of unknown discrete-time linear systems using input-output measured data," *IEEE trans. cybern.*, vol. 45, no. 12, pp. 2770–2779, 2015.

[17] O. Sebbouh, R. M. Gower, and A. Defazio, "Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball," in *Conf. Learning Theory*, 2021, pp. 3935–3971.