Temperature Control for Cyber-Physical Thermal Systems over Wireless Networks: A Model-Assisted Deep Reinforcement Learning Approach

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

***Abstract*—The integration of cyber-physical systems (CPS) with Industrial Internet of Things (IIoT) requires reliable and efficient thermal management, yet existing control methods often degrade under nonlinear thermal dynamics and unreliable wireless links. This paper investigates cyber-physical control of a nonlinear thermal system, where a remote controller regulates the temperature of a thermal plant over a wireless communication network. We first establish a cyber-physical thermal system (CPTS) model that explicitly incorporates nonlinear heat-transfer mechanisms (conduction, convection, and radiation) together with wireless transmission impairments (fading and noise). Based on this model, we formulate an optimal temperature tracking control problem and characterize the structural properties of the optimal solution using the homotopy perturbation method. To enable practical implementation under real-time constraints, we develop a model-assisted structured deep reinforcement learning (DRL) framework, in which a deep neural network (DNN) approximates only the residual high-order terms of the control law while structured update rules guide the effective learning process. The almost sure convergence of the proposed learning scheme is established using Lyapunov stability analysis. Numerical evaluations are performed under a furnace temperature control setup using simulation data generated from the proposed CPTS model parameterized by typical furnace settings, which accurately capture the underlying furnace dynamics. The proposed scheme achieves a best tracking error of about 0.01 in terms of mean square error (MSE) and converges within 50 iterations. This corresponds to a 20 dB reduction in MSE and a twofold improvement in convergence speed compared with state-of-the-art control schemes, thereby demonstrating both the effectiveness and robustness of the proposed approach for wireless CPTS.**

***Index Terms*—Cyber-physical systems, wireless networks, real-time signal processing, nonlinear thermal control, reinforcement learning, Lyapunov stability.**

I. Introduction

CYBER-physical thermal systems (CPTS) are integral to numerous industrial applications, including advanced manufacturing, data center cooling, and energy-efficient building climate control [1]. With the proliferation of Internet-of-Things (IoT) technologies [2], thermal management devices can be rapidly deployed, wirelessly interconnected, and intelligently coordinated, facilitating adaptive, distributed, and real-time temperature regulation. We consider a representative CPTS comprising a *thermal plant* and a *remote temperature controller*, as illustrated in Fig. 1. The thermal plant consists of a *heat source* and a *target object* requiring precise temperature control. The remote controller employs real-time signal processing of temperature feedback from the thermal plant to generate thermal control commands, which are transmitted over a wireless communication channel. The heat source promptly adjusts the temperature of the target object according to the received noisy commands. However, wireless transmission introduces fading, interference, and noise that compromise signal integrity and degrade CPTS temperature control, making effective mitigation of these impairments crucial for robust performance.

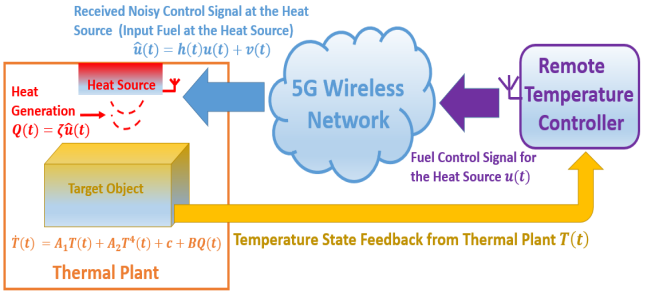


Fig. 1: Architecture of a cyber-physical thermal control system over a wireless network.

Temperature control in CPTS is challenging, largely because accurate regulation depends on precise thermal modeling, whereas many existing works oversimplify system dynamics. Some studies [3] reduce CPTS behavior to linear or AutoRegressive Moving Average with eXogenous inputs (ARMAX) models, while [4] relies on deep neural networks (DNNs) trained purely on empirical data without theoretical guarantees. Physically informed approaches [5], [6] improve fidelity, but [5] considers only heat conduction and overlooks source–object interactions, whereas [6] models convection but insufficiently captures key spatiotemporal dynamics. More comprehensive formulations [7] integrate both conduction and convection, yet nonlinear thermal phenomena such as radiative

This work was supported by the National Natural Science Foundation of China under Grant 62503418, the Zhejiang Provincial Natural Science Foundation of China under Grant LQN26F030001, the National Foreign Expert Project under Project H20240994, and the Research Grants Council of Hong Kong under the Areas of Excellence (AoE) Scheme under Grant AoE/E-601/22-R. *(Corresponding Authors: Songfu Cai; Vincent K. N. Lau)*

Minjie Tang is with the Communication Systems Department, EURECOM, France (e-mail: Minjie.Tang@eurecom.fr). Songfu Cai is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and is also with the Zhejiang Key Laboratory of Multimodal Communication Networks and Intelligent Information Processing, Hangzhou 310027, China (e-mail: sfcai@zju.edu.cn). 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).

transfer remain inadequately represented. Consequently, these modeling simplifications limit the robustness of temperature control schemes in practical CPTS settings, particularly under dynamic thermal loads, external disturbances, and wireless-induced signal distortions.

Second, most existing works on CPTS controller design assume oversimplified communication channels between remote controllers and physical thermal plants. For instance, [8] employs offline pole placement, while [9], [10] adopts PID control to regulate system temperatures over static channels. However, these heuristic approaches lack optimality. To address this limitation, some recent works [11], [12] propose optimal control schemes based on linear quadratic tracker (LQT) and model predictive control (MPC) designs, yet the approach still relies on the assumption of fixed-gain connections between controllers and thermal plants. Notably, all these works [8]–[11] neglect the impacts of time-varying wireless fading and nonlinear thermal interactions, both of which critically affect the control performance of CPTS operating over wireless networks.

To address the nonlinear dynamics inherent in cyber-physical systems (CPS), reinforcement learning (RL) has emerged as a promising strategy [13]. However, classical RL methods often suffer from the curse of dimensionality [14] due to continuous state and action spaces, and discretization-based approaches such as [15] introduce quantization errors that degrade performance. Deep reinforcement learning (DRL) alleviates these limitations by using DNNs to approximate control solutions, thereby reducing the problem to learning a finite set of parameters [16]–[18]. For instance, [19] applied DRL to intrusion detection in IoT and demonstrated the effectiveness of double deep Q-networks (DDQN). In [20], deep Q-networks (DQN) were used for anomaly detection with partially labeled data, improving generalization, while [21] employed an evolutionary DRL-based detection model for insider misuse with limited training data. Despite these advances, the use of arbitrarily structured DNNs in DRL often leads to slow convergence and inefficient learning because they fail to exploit problem-specific structures. Although model-assisted DRL can incorporate system knowledge to generate synthetic data and improve efficiency [18], the design of DRL architectures tailored explicitly for optimal CPS control remains largely underexplored. This challenge becomes even more pronounced in CPTS over wireless networks, where nonlinear thermal dynamics and stochastic, time-varying wireless channels further exacerbate the convergence issues of conventional black-box DRL methods.

There are some recent works that consider DRL-based control over oversimplified wireless communication channels [22], [23]. Specifically, [22] models the communication link as an on–off packet-drop channel without accounting for practical wireless fading. In [23], the channel is represented by an oversimplified, location-dependent, time-invariant SNR model that depends solely on vehicle positions and does not capture any temporal evolution of the fading process; once the positions are fixed, the channel remains constant. Therefore, neither [22] nor [23] incorporates a realistic time-varying wireless fading model into the control loop. Under a realistic time-varying fading channel between the temperature controller and the thermal plant, the channel gain evolves stochastically over time. As a result, the associated optimality conditions must be reformulated to explicitly include the wireless fading state. Consequently, existing control approaches [22], [23] cannot be directly applied because they do not adapt to temporal fluctuations in the fading process and would experience significant degradation in control performance.

In this work, we address temperature control in CPTS with nonlinear thermal dynamics over wireless fading channels by proposing a *structure-aware* RL algorithm *tailored for CPTS*. Unlike the generic applications of reinforcement learning commonly adopted in existing works [16]–[18], we explicitly exploit the intrinsic structure of the optimal solution derived from first-principles thermal laws and develop a structure-aware RL framework that only approximates the unknown residual structured component of the CPTS control solution. This structure-exploiting design substantially reduces learning complexity and achieves significantly faster convergence compared with generic RL-based approaches [16]–[18]. Our key contributions are summarized as follows.

- **Comprehensive Physical Thermal Modeling for CPTS.** To enable effective thermal controller design, it is essential to develop a CPTS thermal model that captures the key heat transfer mechanisms (conduction, convection, and radiation) together with the impact of wireless impairments on control-signal delivery. However, deriving an explicit model that incorporates all these effects is challenging due to the strong and stochastic nonlinear couplings involved. To address this issue, we start from first-principle energy balance equations, apply a weighted residual (Galerkin) method to discretize the governing PDEs, and obtain a tractable state-space representation via reduced-order modeling augmented with random fading effects.
- **Structured Optimality Condition for CPTS Control with Nonlinear Dynamics over Wireless Channels.** Conventional RL methods characterize optimal control via the Hamilton–Jacobi–Bellman (HJB) equation of the value or Q-function, together with a separate link between the co-function and these functions. This two-step structure is inefficient for our setting, where wireless channel impairments are typically ignored. To address this, we seek a structured optimality condition defined directly in terms of the co-function, explicitly capturing both nonlinear thermal dynamics and wireless channel uncertainty. The main difficulty lies in the coupling between nonlinear thermal effects and stochastic channel gains, which renders the dynamics highly nonlinear and time-varying. We overcome this by applying the Pontryagin maximum principle to the joint system–channel state and exploiting closed-form thermal models to derive a tractable structured optimality condition. To the best of our knowledge, this is the first structured optimality condition tailored to CPTS control over wireless networks that jointly incorporates fundamental thermal laws and wireless fading effects.

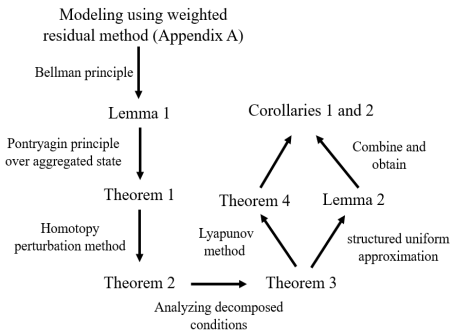


Fig. 2: Connections between the key theorems, lemmas and corollaries.

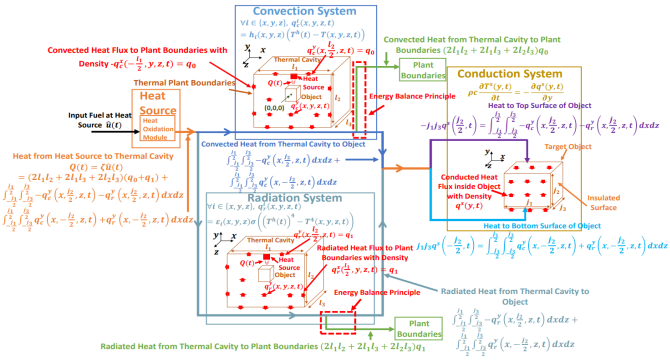


Fig. 3: Diagram of the heat flow in the thermal plant.

- **Structured CPTS Control Solution with Nonlinear Dynamics over Wireless Channels.** To facilitate efficient learning of the optimal control solution, we leverage the structural properties of the co-function induced by the partially nonlinear thermal dynamics. Specifically, we decompose the co-function into a linear kernel, corresponding to the purely linear thermal dynamics, and a higher-order perturbation term that satisfies a decomposed optimality condition. Since the linear kernel can be derived analytically offline, only the perturbation term needs to be learned, which substantially reduces the complexity of the learning process. However, this decomposition is challenging because the linear and nonlinear terms of the co-function are coupled in the original optimality condition. To address this, we employ a homotopy-based method that systematically separates the co-function into a deterministic linear kernel and a residual perturbation.
- **Structured Online RL Algorithm for CPTS Control over Wireless Channels.** The structured form of the co-function, together with the decomposed optimality condition, allows us to approximate only the higher-order perturbation component using a lightweight neural network (NN). The network parameters are optimized by enforcing the structured optimality condition within the proposed RL framework, and convergence is ensured through Lyapunov analysis. Compared with black-box methods that approximate the entire co-function, our approach (i) achieves faster convergence, as reflected by the faster decay of the gradient norm during learning, and (ii) requires significantly lower computational complexity than the quartic black-box bound, since only the residual perturbation is learned with far fewer parameters under structured update rules.

A flowchart that elucidates the connections between the key lemmas, theorems and corollaries in this work is provided in Fig. 2. The notations that will be used throughout this article are summarized as follows. Bold uppercase and lowercase letters denote matrices and vectors, respectively. The operators $(\cdot)^T$ and $\mathrm{Tr}(\cdot)$ represent the transpose and trace of a matrix. $\mathrm{Diag}(a, b, ...)$ denotes a diagonal matrix with elements $\{a, b, ...\}$. The sets $\mathbb{R}^{m\times n}$, $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{S}_+^a$, $\mathbb{S}^a$, $\mathbb{Z}_+$, and $\mathbb{Z}$ represent $m\times n$ real matrices, real numbers, positive real numbers, $a\times a$ positive definite matrices, $a\times a$ positive semi-definite matrices, positive integers, and non-negative integers, respectively. The spectral norm and the maximum eigenvalue of $\mathbf{A}$ are $\|\mathbf{A}\|$ and $\lambda(\mathbf{A})$, respectively. $[\mathbf{A}]_{a,b}$ refers to the element at the a-th row and b-th column. The Hadamard product of matrices $\mathbf{A}$ and $\mathbf{B}$ is $\mathbf{A}\odot\mathbf{B}$. The b-th element of vector $\mathbf{a}$ is $[\mathbf{a}]_b$, and $\|\mathbf{a}\|$ denotes its L_2 norm. Given $\mathbf{a} = [a_1, ..., a_l]^T \in \mathbb{R}^{l\times 1}$, its element-wise b-th power is $\mathbf{a}^b = [a_1^b, ..., a_l^b]^T$. Identity and zero matrices are denoted as $\mathbf{I}_s$ (size $s\times s$), $\mathbf{I}_{a\times b}$ (all ones, size $a\times b$), $\mathbf{0}_s$ (zero matrix, size $s\times s$), and $\mathbf{0}_{a\times b}$ (zero matrix, size $a\times b$). The inner product of vectors $\mathbf{a}$ and $\mathbf{b}$ is $\langle \mathbf{a}, \mathbf{b}\rangle$. The indicator function $\mathbf{1}_{\{a\geq 0\}} \in \{0, 1\}$ satisfies $\mathbf{1}_{\{a\geq 0\}} = 1$ if and only if $a \geq 0$.

II. CPTS Model

A. Heat Convection between Target Object and Heat Source

We consider a CPTS thermal plant with an effective thermal cavity of dimensions $l_1 \times l_2 \times l_3$, as shown in Fig. 3. At time $t \in \mathbb{R}_+$, heat convection within the thermal cavity is described by the heat convection law:

$$q_c^i(x, y, z, t) = h_i(x, y, z)(T^h(t) - T(x, y, z, t)), i \in \{x, y, z\}, \quad (1)$$

where $q_c^i(x, y, z, t)$ and $h_i(x, y, z)$ denote heat flux density and heat transfer coefficient for convection in the i-direction, respectively; $T^h(t)$ and $T(x, y, z, t)$ represent the temperatures of the heat source and thermal cavity. The triplet of (x, y, z) denotes the three-dimensional Cartesian coordinates.

The Neumann boundary condition [24] for (1) is given by:

$$-q_c^i(-\tfrac{l_i}{2}, t) = q_c^i(\tfrac{l_i}{2}, t) = q_0, \quad i \in \{x, y, z\}. \quad (2)$$

B. Heat Radiation in the Thermal Cavity

For $i \in \{x, y, z\}$, heat radiation within the thermal cavity is governed by the Stefan-Boltzmann law:

$$q_r^i(x, y, z, t) = \sigma\epsilon_i(x, y, z)\left((T^h(t))^4 - T^4(x, y, z, t)\right), \quad (3)$$

where $q_r^i(x, y, z, t)$ denotes the heat flux density for radiation in the i-direction, $\epsilon_i(x, y, z) \in (0, 1)$ is the emissivity, and $\sigma = 6.7 \times 10^{-8}$ is the Stefan-Boltzmann constant.

The Neumann boundary condition for (3) is given by:

$$-q_r^i(-\tfrac{l_i}{2}, t) = q_r^i(\tfrac{l_i}{2}, t) = q_1, \quad i \in \{x, y, z\}. \quad (4)$$

C. Heat Conduction in the Target Object

We model heat conduction within a target object of dimensions $j_1 \times j_2 \times j_3$, centrally positioned within the thermal cavity, under the following assumption.

Assumption 1 (Target Object Properties [25]): The surfaces of the object in the xy and yz planes are insulated, and its temperature field is uniform across these planes. ■

Consequently, the temperature field $T^s(y,t)$ along the spatial dimension $y \in [-\frac{j_2}{2}, \frac{j_2}{2}]$ follows the transient heat conduction equation:

$$\rho c \frac{\partial T^s(y,t)}{\partial t} = \lambda_s \frac{\partial^2 T^s(y,t)}{\partial y^2}, \tag{5}$$

where ρ, c, and λ_s denote mass density, heat capacity, and thermal conductivity of the object, respectively.

The Neumann boundary condition for (5) is given by:

$$\begin{aligned} q^s\left(\pm\frac{j_2}{2}, t\right) &= -\lambda_s \frac{\partial T^s(y,t)}{\partial y}\bigg|_{y=\pm\frac{j_2}{2}} \\ &= q_r^y(x, \pm\tfrac{j_2}{2}, z, t) + q_c^y(x, \pm\tfrac{j_2}{2}, z, t), \end{aligned} \tag{6}$$

where $x \in [-\frac{j_1}{2}, \frac{j_1}{2}]$, $z \in [-\frac{j_3}{2}, \frac{j_3}{2}]$.

The boundary temperatures satisfy:

$$T^s\left(\pm\frac{j_2}{2}, t\right) = T\left(x, \pm\frac{j_2}{2}, z, t\right). \tag{7}$$

The conduction heat flux $q^s(y,t)$, which governs the temperature within the object, is driven by the surface radiative and convective fluxes $q_r^i(x,y,z,t)$ and $q_c^i(x,y,z,t)$, where $i \in \{x,y,z\}$, within the thermal cavity.

D. Heat Process at the Heat Source

The thermal energy $\widehat{u}(t) \in \mathbb{R}$ supplied by the heat source generates effective heat energy $Q(t) \in \mathbb{R}$, which is transferred to the thermal cavity. However, due to inefficiencies and losses, only a fraction of this energy contributes effectively:

$$Q(t) = \zeta \widehat{u}(t), \tag{8}$$

where $\zeta \in (0,1)$ denotes the heat source efficiency.

According to the energy balance principle, the generated heat $Q(t)$ is equal to the sum of the heat transferred to the cavity boundaries and the target object:

$$\begin{aligned} Q(t) = &\underbrace{(2l_1l_2 + 2l_1l_3 + 2l_2l_3)(q_0 + q_1)}_{\text{Boundary heat losses}} + \\ &\underbrace{j_1 j_3 \left[q^s(\tfrac{j_2}{2}, t) + q^s(-\tfrac{j_2}{2}, t)\right]}_{\text{Heat to target object}}. \end{aligned} \tag{9}$$

E. Overall Dynamic Model of the Thermal Plant

The overall dynamic model of the CPTS incorporates heat conduction within the target object, heat convection and radiation in the thermal cavity, and heat generation at the heat source. Applying the weighted residual method [26] to equations (1)–(9), we obtain the CPTS dynamics:[1]

$$\dot{\mathbf{T}}(t) = \mathbf{A}_1 \mathbf{T}(t) + \mathbf{A}_2 \mathbf{T}^4(t) + \mathbf{B}\hat{u}(t) + \mathbf{c}, \mathbf{T}(0) = \mathbf{T}_0, t \geq 0, \tag{10}$$

[1]See Appendix A for the detailed derivation of (10) and the resulting properties $\lambda(\mathbf{A}_1) < 0$ and $\lambda(\mathbf{A}_2) = 0$.

where $\mathbf{T}(t) = [T^s(\frac{j_2}{2}, t), T^s(0,t), T^s(-\frac{j_2}{2}, t)]^T \in \mathbb{R}^{3\times 1}$ represents the temperature state of the target object, and $\mathbf{T}^4(t) = [(T^s(\frac{j_2}{2}, t))^4, (T^s(0,t))^4, (T^s(-\frac{j_2}{2}, t))^4]^T \in \mathbb{R}^{3\times 1}$ denotes its fourth-power temperature. The initial thermal state is given by $\mathbf{T}_0 \in \mathbb{R}^{3\times 1}$. The thermal transition matrices are defined as $\mathbf{A}_1 = \begin{bmatrix} \frac{-54\lambda_s}{\rho c j_2^2} & \frac{95\lambda_s}{\rho c j_2^2} & \frac{-42\lambda_s}{\rho c j_2^2} \\ \frac{30\lambda_s}{\rho c j_2^2} - \frac{3h_0}{\rho c j_2} & \frac{-60\lambda}{\rho c j_2^2} & \frac{30\lambda}{\rho c j_2^2} + \frac{3h_0}{\rho c j_2} \\ \frac{54\lambda_s}{\rho c j_2^2} & \frac{-96\lambda_s}{\rho c j_2^2} & \frac{42\lambda_s}{\rho c j_2^2} \end{bmatrix}$ and $\mathbf{A}_2 = \begin{bmatrix} 0 & 0 & 0 \\ \frac{-3\epsilon_0\sigma}{\rho c j_2} & 0 & \frac{3\epsilon_0\sigma}{\rho c j_2} \\ 0 & 0 & 0 \end{bmatrix}$. The actuation matrix and thermal bias are given by $\mathbf{B} = \begin{bmatrix} \frac{\zeta}{\rho c j_1 j_2 j_3} \\ 0 \\ \frac{15\zeta}{2\rho c j_1 j_2 j_3} \end{bmatrix}$ and $\mathbf{c} = \begin{bmatrix} \frac{(2l_1l_2+2l_2l_3+2l_1l_3)(q_0+q_1)}{\rho c j_1 j_2 j_3} \\ 0 \\ \frac{(15l_1l_2+15l_2l_3+15l_1l_3)(q_0+q_1)}{\rho c j_1 j_2 j_3} \end{bmatrix}$. The heat transfer coefficient and emissivity satisfy $h_0 = -h_y(x, \frac{j_2}{2}, z) = h_y(x, -\frac{j_2}{2}, z)$ and $\epsilon_0 = -\epsilon_y(x, \frac{j_2}{2}, z) = \epsilon_y(x, -\frac{j_2}{2}, z)$, for $x \in [-\frac{j_1}{2}, \frac{j_1}{2}]$ and $z \in [-\frac{j_3}{2}, \frac{j_3}{2}]$.

Remark 1 (Key Properties of $\lambda(\mathbf{A}_1) < 0$ and $\lambda(\mathbf{A}_2) = 0$)**.** *Note that $\lambda(\mathbf{A}_1) < 0$ and $\lambda(\mathbf{A}_2) = 0$ hold for all physically meaningful values of ρ, c, and λ_s. Specifically, for any physically meaningful (i.e., strictly positive) ρ, c, and λ_s, we have $\mathbf{A}_1 = \frac{\lambda_s}{\rho c j_2^2}\mathbf{L}$, which is a positive scaling of the matrix $\mathbf{L} = \begin{bmatrix} -54 & 95 & -42 \\ 30 - 3h_0 j_2/\lambda_s & -60 & 30 + 3h_0 j_2/\lambda_s \\ 54 & -96 & 42 \end{bmatrix}$. As rigorously proved in Appendix A, $\lambda(\mathbf{L}) < 0$ for all such parameter values, and it follows that $\lambda(\mathbf{A}_1) < 0$ always holds. Moreover, $\mathbf{A}_2$ is a nilpotent matrix satisfying $\mathbf{A}_2^2 = \mathbf{0}_3$, and thus all of its eigenvalues are zero. Therefore, $\lambda(\mathbf{A}_2) = 0$ for any physically meaningful ρ, c, and λ_s. It is worth noting that the fact that $\lambda(\mathbf{A}_1) < 0$ and $\lambda(\mathbf{A}_2) = 0$ is fully consistent with the heat balance law and physical intuition. Without external heat input, the temperature of a thermal object must remain bounded over time rather than diverging to infinity. This fundamental behavior holds regardless of the specific values of the mass density, heat capacity, or thermal conductivity of the object.*

F. Wireless Communication Model

We account for wireless signal distortions and model the communication link between the remote controller and the heat source as a wireless fading channel. Based on the temperature feedback $\mathbf{T}(t)$, the remote controller generates a thermal control signal $u(t) \in \mathbb{R}$, which is transmitted to the heat source. The received signal at the heat source is given by:

$$\hat{u}(t) = h(t)u(t) + v(t), \tag{11}$$

where $h(t) \sim \mathcal{N}(0, \xi)$ denotes wireless fading with variance $\xi > 0$, and $v(t) \sim \mathcal{N}(0,1)$ is additive white Gaussian noise (AWGN).

It is worth noting that the thermal radiation inside the furnace does not affect the quality of wireless communications between the plant and the remote temperature controller, since it lies in the infrared spectrum (typically 0.7–20 μm), whereas industrial wireless systems operate in the sub-6 GHz bands

with carrier wavelengths on the order of 5–50 cm. As the radiation wavelength is several orders of magnitude shorter than the carrier wavelength, its direct coupling into the radio-frequency band can be ignored. By contrast, the working conditions of industrial sites (e.g., metallic furnace walls, surrounding equipment, or heavy machinery) may introduce multipath propagation, shadowing, and time-varying attenuation that impact the communication quality. These effects are well captured by the fading channel model $h(t)$ adopted in our formulation, whose variance ξ characterizes the signal-to-noise ratio (SNR) and thus the communication quality.

Remark 2 (Industrial Applicability of the CPTS Model)**.** *The modeling framework in Section II explicitly incorporates the three fundamental heat transfer mechanisms: conduction, convection, and radiation, which universally govern industrial thermal processes. These mechanisms arise in diverse applications, including regulating fuel input in reheating furnaces, controlling temperature trajectories in ceramic sintering, adjusting mold or chamber temperatures in composite manufacturing, and maintaining safe operating temperatures in battery modules. Therefore, the proposed control scheme is broadly applicable to industrial thermal systems by specifying the geometry, material properties, and operating parameters for each specific application.*

III. Structured Temperature Control for CPTS over Wireless Networks

A. Problem Formulation for Optimal Temperature Control in CPTS over Wireless Networks

Let $t \geq 0$ and $\mathbf{T}(0) = \mathbf{T}_0$. The evolution of the temperature state $\mathbf{T}(t)$ in the CPTS with partially nonlinear time-varying (PNTV) dynamics is obtained by combining (10) and (11), given by:

$$\dot{\mathbf{T}}(t) = \mathbf{A}_1\mathbf{T}(t) + \mathbf{A}_2\mathbf{T}^4(t) + \mathbf{c} + h(t)\mathbf{B}u(t) + \mathbf{B}v(t). \quad (12)$$

Suppose the target temperature profile $\mathbf{r}(t) \in \mathbb{R}^{3\times 1}$ evolves as:

$$\dot{\mathbf{r}}(t) = \mathbf{G}\mathbf{r}(t), \quad t \geq 0, \quad \mathbf{r}(0) = \mathbf{r}_0, \quad (13)$$

where $\mathbf{G} \in \mathbb{R}^{3\times 3}$ characterizes the target thermal transition matrix. The optimal temperature control problem for the PNTV CPTS is formulated below.

Problem 1 (CPTS Temperature Control Problem)**.**

$$\begin{aligned} &\min_{\pi} \limsup_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{E}\left[r(\mathbf{T}(t),\mathbf{r}(t),h(t),u(t))\right]dt \\ &\textit{s.t.} \quad \dot{\mathbf{T}}(t) = \mathbf{A}_1\mathbf{T}(t) + \mathbf{A}_2\mathbf{T}^4(t) + \mathbf{c} + h(t)\mathbf{B}u(t) + \mathbf{B}v(t), \\ &\qquad \dot{\mathbf{r}}(t) = \mathbf{G}\mathbf{r}(t), \quad \mathbf{T}(0) = \mathbf{T}_0, \quad \mathbf{r}(0) = \mathbf{r}_0, \end{aligned} \quad (14)$$

where the temperature control policy is $\pi = \{u(t), t \geq 0\}$. The per-stage reward function $r(\mathbf{T}(t),\mathbf{r}(t),h(t),u(t))$ is defined as:

$$\begin{aligned} r(\mathbf{T}(t),\mathbf{r}(t),h(t),u(t)) &= (\mathbf{T}(t)-\mathbf{r}(t))^T\mathbf{Q}(\mathbf{T}(t)-\mathbf{r}(t)) \\ &\quad + (R + Mh^2(t))u^2(t), \end{aligned} \quad (15)$$

where $(\mathbf{T}(t)-\mathbf{r}(t))^T\mathbf{Q}(\mathbf{T}(t)-\mathbf{r}(t))$ represents the temperature tracking error cost, modeling the instantaneous deviation between the target temperature state and the real-time temperature state of the target object. $(R + Mh^2(t))u^2(t)$ consists of: (i) the control cost $Ru^2(t)$, representing the transmission power over the wireless interface from the remote controller to the heat source, and (ii) the thermal input cost $Mh^2(t)u^2(t)$, modeling the thermal energy consumption at the heat source. The weighting coefficients are $\mathbf{Q} \in \mathbb{S}_+^3$, $R \in \mathbb{R}_+$, and $M \in \mathbb{R}_+$. The expectation in the objective function is taken with respect to (w.r.t.) the random CSI $h(t)$ and the AWGN $v(t)$ at the heat source.

Remark 3 (Feasibility of Problem 1 and Stability of the Thermal Dynamics (10))**.** *Problem 1 is feasible provided that the target trajectory $\mathbf{r}(t)$ remains bounded, owing to the* internal stability *of the thermal dynamics* (10)*. Note that $\lambda(\mathbf{A}_1) < 0$ and $\lambda(\mathbf{A}_2) = 0$, as discussed in detail in Remark 1 and rigorously proved in Appendix A. Consequently, under the optimal input $u^*(t) \in \mathbb{R}$, the closed-loop system remains stable, ensuring a finite optimal cost, i.e., $\limsup_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{E}[r(\mathbf{T}(t),\mathbf{r}(t),h(t),u^*(t))]dt < \infty$. This aligns with the energy-balance principle, which guarantees bounded temperature evolution even without external heat injection.Nonetheless, our objective extends beyond ensuring stability alone; we aim to achieve precise temperature tracking and efficient energy usage by minimizing the long-term cost $\limsup_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{E}[r(\mathbf{T}(t),\mathbf{r}(t),h(t),u(t))]dt$ rather than solely focusing on stability.*

B. Structured Optimality Condition for CPTS Control

Traditionally, optimal CPTS temperature control is obtained by solving the Hamilton-Jacobi-Bellman (HJB) equation. Specifically, Problem 1 can be equivalently reformulated as a linear quadratic regulator (LQR) problem.

Problem 2 (Equivalent Formulation for Problem 1)**.**

$$\begin{aligned} &\min_{\pi} \limsup_{T\to\infty} \frac{1}{T}\int_{t=0}^T \mathbb{E}[r^a(\hat{\mathbf{T}}(t),h(t),u(t))]dt \\ &\textit{s.t.} \quad \dot{\hat{\mathbf{T}}}(t) = \hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) + \hat{\mathbf{c}} + \hat{\mathbf{B}}(t)u(t) + \hat{\mathbf{v}}(t), \\ &\qquad \hat{\mathbf{T}}(0) = \hat{\mathbf{T}}_0, \end{aligned} \quad (16)$$

where $\hat{\mathbf{T}}_0 = [\mathbf{T}_0^T,\, \mathbf{r}_0]^T \in \mathbb{R}^{6\times 1}$ is the aggregated initial state. The expectation in (16) *is taken w.r.t. the random CSI $h(t)$ and the AWGN $v(t)$ at the actuator. The equivalent per-stage reward function $r^a(\hat{\mathbf{T}}(t),\, h(t),\, u(t))$ is given by*

$$r^a(\hat{\mathbf{T}}(t),h(t),u(t)) = \hat{\mathbf{T}}^T(t)\hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + Ru^2(t) + Mh^2(t)u^2(t), \quad (17)$$

where $\hat{\mathbf{T}}(t) = [\mathbf{T}^T(t),\, \mathbf{r}^T(t)]^T \in \mathbb{R}^{6\times 1}$ is the aggregated temperature state, $\hat{\mathbf{Q}} = \begin{bmatrix} \mathbf{Q} & -\mathbf{Q} \\ -\mathbf{Q} & \mathbf{Q} \end{bmatrix} \in \mathbb{S}^6$ is the aggregated weighting matrix, $\hat{\mathbf{A}}_1 = \mathrm{Diag}(\mathbf{A}_1, \mathbf{G}) \in \mathbb{R}^{6\times 6}$ and $\hat{\mathbf{A}}_2 = \mathrm{Diag}(\mathbf{A}_2, \mathbf{0}_3) \in \mathbb{R}^{6\times 6}$ are the linear and nonlinear thermal transition matrix, $\hat{\mathbf{B}}(t) = [h(t)\,\mathbf{B}^T,\ \mathbf{0}_{1\times 3}]^T \in \mathbb{R}^{6\times 1}$ is the aggregated actuation matrix, $\hat{\mathbf{v}}(t) = [\mathbf{v}^T(t)\,\mathbf{B}^T,\ \mathbf{0}_{1\times 3}]^T \in \mathbb{R}^{6\times 1}$ is the aggregated noise, and $\hat{\mathbf{c}} = [\,\mathbf{c}^T,\ \mathbf{0}_{1\times 3}\,]^T \in \mathbb{R}^{6\times 1}$ is the aggregated bias.

Consequently, the optimal temperature control solution to Problem 1 can be obtained by solving the HJB equation for Problem 2, as stated in the following lemma.

Lemma 1 (The HJB Equation and Optimal Solution to Problem 2)**.** *The optimal solution to Problem 2 is given by*

$$u^*(t) \;=\; -\bar{R}(t)\,\hat{\mathbf{B}}^T(t)\,\lambda\big(\hat{\mathbf{T}}(t)\big), \tag{18}$$

where $\bar{R}(t) = \big(2R + 2M\,h^2(t)\big)^{-1}$ *and* $\lambda(\hat{\mathbf{T}}(t)) = \frac{\partial V(\hat{\mathbf{T}}(t))}{\partial \hat{\mathbf{T}}(t)}$. *The value function* $V(\hat{\mathbf{T}}(t))$ *satisfies the HJB equation for Problem 2:*

$$-\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial t} \;=\; \mathbb{E}_{h(t)}\Big[\min_{u(t)}\Big(\hat{\mathbf{T}}^T(t)\,\hat{\mathbf{Q}}\,\hat{\mathbf{T}}(t) \;+\; (R + M\,h^2(t))\; u^2(t) \;+\; \mathbb{E}\Big[\Big\langle \lambda(\hat{\mathbf{T}}(t)), \dot{\hat{\mathbf{T}}}(t)\Big\rangle \Big| \hat{\mathbf{T}}(t),\, h(t),\, u(t)\Big]\Big)\Big]. \tag{19}$$

Proof: See Appendix B. ∎

Note that the unknown co-function $\lambda(\hat{\mathbf{T}}(t))$ in $u^*(t)$ of (18) is the state-derivative of the solution $V(\hat{\mathbf{T}}(t))$ to the HJB equation (19). As a result, one may consider using traditional RL algorithms, such as value iteration or Q-learning, to solve (19) and subsequently derive the optimal solution $u^*(t)$ in (18) based on their relationship. However, this two-stage approach is inefficient. To enable efficient learning, we derive a structured optimality condition for $\lambda(\hat{\mathbf{T}}(t))$ using Pontryagin's principle.

Theorem 1 (Optimality Condition w.r.t. the Co-Function for CPTS Control)**.** *The optimality condition (19) for Problem 2 can be represented w.r.t. the co-function* $\lambda(\hat{\mathbf{T}}(t))$*, given by*

$$\frac{\partial \lambda\big(\hat{\mathbf{T}}(t)\big)}{\partial \hat{\mathbf{T}}(t)}\Big(\hat{\mathbf{A}}_1\,\hat{\mathbf{T}}(t) - \mathbb{E}\big[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)\big]\,\lambda\big(\hat{\mathbf{T}}(t)\big) + \hat{\mathbf{A}}_2 \times \hat{\mathbf{T}}^4(t) + \hat{\mathbf{c}}\Big) + \hat{\mathbf{Q}}\,\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T\,\lambda\big(\hat{\mathbf{T}}(t)\big) + \big(4\,\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t) \times \mathbf{I}_{1\times 6})\big)\lambda(\hat{\mathbf{T}}(t)) = \mathbf{0}_{6\times 1}, \forall \hat{\mathbf{T}}(t) \in \mathbb{R}^{6\times 1}. \tag{20}$$

Proof: See Appendix C. ∎

C. Structured Optimal Solution for CPTS Control

Note that the unknown co-function $\lambda(\hat{\mathbf{T}}(t))$ is defined on a continuous state space $\hat{\mathbf{T}}(t) \in \mathbb{R}^{6\times 1}$, making brute-force solutions of (20) computationally infeasible. Hence, we exploit the inherent structure of $\lambda(\hat{\mathbf{T}}(t))$ by modeling its structured kernel. However, deriving an explicit form is challenging due to the strong nonlinear coupling involving the cubic state term $\mathbf{T}^3(t)$ in $\big(4\,\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t)\mathbf{I}_{1\times 6})\big)\lambda(\hat{\mathbf{T}}(t))$. To address this difficulty, we observe the Riccati structures inherent in linear optimality conditions and recognize CPTS dynamics as linear systems with higher-order nonlinear perturbations. By applying the homotopy perturbation method [27] to (20), we derive the following theorem, characterizing a structured decomposition of $\lambda(\hat{\mathbf{T}}(t))$.

Theorem 2 (Structured Decomposition of the Co-Function)**.** *The co-function* $\lambda(\hat{\mathbf{T}}(t))$ *admits the decomposition:*

$$\lambda\big(\hat{\mathbf{T}}(t)\big) \;=\; \sum_{n=0}^{\infty} \lambda_n\big(\hat{\mathbf{T}}(t)\big), \tag{21}$$

where each component $\lambda_n\big(\hat{\mathbf{T}}(t)\big)$ *satisfies the PDE*

$$\mathcal{D}_n\big(\hat{\mathbf{T}}(t), \lambda_n\big(\hat{\mathbf{T}}(t)\big)\big) + \mathcal{P}_n\big(\hat{\mathbf{T}}(t), \lambda_{n-1}\big(\hat{\mathbf{T}}(t)\big)\big)\,\mathbf{1}_{\{n\geq 1\}} = \mathbf{0}_{6\times 1}, \tag{22}$$

subject to the initial condition $\lambda_n\big(\mathbf{0}_{6\times 1}\big) = \mathbf{0}_{6\times 1}$. *The dominant term* $\mathcal{D}_n$ *is*

$$\mathcal{D}_n\big(\hat{\mathbf{T}}(t), \lambda_n\big(\hat{\mathbf{T}}(t)\big)\big) = \frac{\partial \lambda_n\big(\hat{\mathbf{T}}(t)\big)}{\partial \hat{\mathbf{T}}(t)}\Big[\hat{\mathbf{A}}_1\,\hat{\mathbf{T}}(t) - \mathbb{E}\big[\bar{R}(t)\,\hat{\mathbf{B}}(t) \times \hat{\mathbf{B}}^T(t)\big]\,\lambda_n\big(\hat{\mathbf{T}}(t)\big)\Big] + \hat{\mathbf{Q}}\,\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T\,\lambda_n\big(\hat{\mathbf{T}}(t)\big), \tag{23}$$

and the perturbation term $\mathcal{P}_n$ *is*

$$\mathcal{P}_n\big(\hat{\mathbf{T}}(t), \lambda_{n-1}\big(\hat{\mathbf{T}}(t)\big)\big) = \Big[\frac{\partial \lambda_{n-1}\big(\hat{\mathbf{T}}(t)\big)}{\partial \hat{\mathbf{T}}(t)}\,\hat{\mathbf{A}}_2\,\hat{\mathbf{T}}^4(t) + \big(4\,\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t)\,\mathbf{I}_{1\times 6})\big)\,\lambda_{n-1}\big(\hat{\mathbf{T}}(t)\big)\Big]\,\mathbf{1}_{\{n\geq 1\}} \;+\; \hat{\mathbf{c}}\,\mathbf{1}_{\{n=1\}}, \tag{24}$$

where $\lambda_{-1}(\cdot) = \mathbf{0}_{6\times 1}$.

Proof: See Appendix D. ∎

By analyzing each component $\lambda_n(\hat{\mathbf{T}}(t))$ and its relationship to the overall co-function $\lambda(\hat{\mathbf{T}}(t))$ from Theorem 2, we derive the following theorem characterizing the structure of $\lambda(\hat{\mathbf{T}}(t))$ and the corresponding optimal control $u^*(t)$.

Theorem 3 (Structure of Optimal Control Solution)**.** *The optimal solution has a structured form as follows.*

$$u^*(t) = -\bar{R}(t)\hat{\mathbf{B}}^T(t)\lambda(\hat{\mathbf{T}}(t)), \tag{25}$$

where $\lambda(\hat{\mathbf{T}}(t))$ *is the structured co-function given by*

$$\lambda(\hat{\mathbf{T}}(t)) = \mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + g(\hat{\mathbf{T}}(t)). \tag{26}$$

$\mathbf{C}_0 \in \mathbb{R}^{6\times 1}$, $\mathbf{C}_1 \in \mathbb{R}^{6\times 6}$, *and* $g(\hat{\mathbf{T}}(t))$ *represents higher-order nonlinear terms, given by*

$$g(\hat{\mathbf{T}}(t)) = \sum_{m=3}^{\infty} \sum_{\{a,\ldots,f\}\in\mathcal{C}_m} \mathbf{C}_m(a,...,f)[[\hat{\mathbf{T}}(t)]_1^a, \ldots, [\hat{\mathbf{T}}(t)]_6^f]^T. \tag{27}$$

Furthermore, they satisfy the optimality conditions as follows.

$$-\mathbf{C}_1\mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)]\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{c}} + \hat{\mathbf{A}}_1^T\mathbf{C}_0 = \mathbf{0}_{6\times 1}, \tag{28}$$

$$\mathbf{C}_1\hat{\mathbf{A}}_1 + \hat{\mathbf{A}}_1^T\mathbf{C}_1 - \mathbf{C}_1\mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)]\mathbf{C}_1 + \hat{\mathbf{Q}} = \mathbf{0}_{6\times 6}. \tag{29}$$

and

$$\left(\frac{\partial g(\hat{\mathbf{T}}(t))}{\partial \hat{\mathbf{T}}(t)}\right)\Big(\hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) - \mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)]\Big(\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + g(\hat{\mathbf{T}}(t))\Big) + \hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) + \hat{\mathbf{c}}\Big) + \hat{\mathbf{A}}_1^T g(\hat{\mathbf{T}}(t)) + (4\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t) \mathbf{I}_{1\times 6}))\Big(\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + g(\hat{\mathbf{T}}(t))\Big) + \mathbf{C}_1(-\mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)] g(\hat{\mathbf{T}}(t)) + \hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t)) = \mathbf{0}_{6\times 1}, \forall \hat{\mathbf{T}}(t) \in \mathbb{R}^{6\times 1}. \tag{30}$$

Proof: See Appendix E of the supplementary material. ∎

Since the optimal CPTS control solution, given by $u^*(t) = -\bar{R}(t)\hat{\mathbf{B}}^T\lambda(\hat{\mathbf{T}}(t))$, depends on the co-function $\lambda(\hat{\mathbf{T}}(t))$, it is desirable to approximate $\lambda(\hat{\mathbf{T}}(t))$ in order to compute $u^*(t)$. This leads to the estimation of $\Big\{\mathbf{C}_0, \mathbf{C}_1, g(\hat{\mathbf{T}}(t))\Big\}$. While the linear kernel $\{\mathbf{C}_0, \mathbf{C}_1\}$, satisfying the optimality conditions (28) and (29), can be efficiently computed offline

using standard solvers such as the Schur method and the Newton–Kleinman iteration [28], obtaining the nonlinear residual $g(\hat{\mathbf{T}}(t))$ efficiently in an offline manner remains challenging. In the following section, we propose an online structured learning algorithm to approximate the optimal CPTS control solution $u^*(t)$ by learning the nonlinear component $g(\hat{\mathbf{T}}(t))$ in real time.

IV. Structured Online RL for Temperature Control of CPTS over Wireless Networks

A. Black-box RL for CPTS Control

We learn the optimal CPTS solution $u^*(t)$ in (18) by approximating $\lambda(\hat{\mathbf{T}}(t))$ using an unstructured NN $f_b(\hat{\mathbf{T}}(t);\theta_b)$, whose parameter $\theta_b \in \mathbb{R}^{l_b\times 1}$ (with $l_b \in \mathbb{Z}_+$ neurons) is learned online. Given that $\lambda(\hat{\mathbf{T}}(t))$ satisfies the optimality condition (20), the parameter optimization problem is formulated as:

Problem 3 (Black-box RL for CPTS Control)**.**

$$\min_{\theta_b} L_b(\theta_b) = \min_{\theta_b} \limsup_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{E}\big[e_b\big(t;\theta_b\big)\big]\,dt, \tag{31}$$

where the expectation is taken over the CSI $h(t)$ and the AWGN $v(t)$. The term $e_b(t;\theta_b)\in\mathbb{R}$ is defined as

$$\begin{aligned} e_b(t;\theta_b) &= \tfrac{1}{2}\left\|\bar{\mathbf{e}}_b\big(t,\theta_b\big)\right\|^2 \\ &= \tfrac{1}{2}\Big\|\tfrac{\partial f_b\left(\hat{\mathbf{T}}(t);\theta_b\right)}{\partial\hat{\mathbf{T}}(t)}\big(\hat{\mathbf{A}}_1\,\hat{\mathbf{T}}(t) - \bar{R}(t)\,\hat{\mathbf{B}}(t)\,\hat{\mathbf{B}}^T(t)\,f_b\big(\hat{\mathbf{T}}(t);\theta_b\big) \\ &\quad + \hat{\mathbf{A}}_2\,\hat{\mathbf{T}}^4(t) + \hat{\mathbf{c}}\big) \;+\; \hat{\mathbf{Q}}\,\hat{\mathbf{T}}(t) \;+\; \hat{\mathbf{A}}_1^T\,f_b\big(\hat{\mathbf{T}}(t);\theta_b\big) \\ &\quad + \Big(4\,\hat{\mathbf{A}}_2^T \odot \big(\hat{\mathbf{T}}^3(t)\,\mathbf{I}_{1\times 6}\big)\Big)\,f_b\big(\hat{\mathbf{T}}(t);\theta_b\big)\Big\|^2. \end{aligned} \tag{32}$$

By using temporal difference (TD) learning [29], [30], we can iteratively update the NN parameter $\theta_{b,k} = [\theta_{b,k}^1,...,\theta_{b,k}^{l_b}]^T \in \mathbb{R}^{l_b\times 1}$ at each sampling instant k based on the real-time CSI $h(k\tau)$ and the measured temperature state $\mathbf{T}(k\tau)$ from the thermal plant, as follows [2].

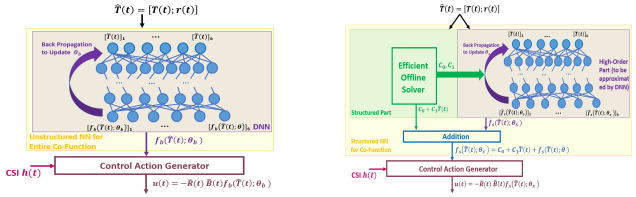
$$\begin{aligned} \theta_{b,k+1} &= \theta_{b,k} - \alpha_k \nabla_{\theta_{b,k}} e_b(k\tau;\theta_{b,k}) \\ &= \theta_k - \alpha_k \mathbf{m}_k, \end{aligned} \tag{33}$$

where $\tau > 0$ is the sampling period. The step size $\alpha_k > 0$ satisfies $\sum_{k=0}^{\infty}\alpha_k = \infty$ and $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$. The increment is defined as $\mathbf{m}_k = [\,\mathbf{m}_{1,k}^T\bar{\mathbf{e}}_b(k\tau;\theta_{b,k}),\ldots,\mathbf{m}_{l_b,k}^T\bar{\mathbf{e}}_b(k\tau;\theta_{b,k})\,]^T \in \mathbb{R}^{l_b\times 1}$, with each $\mathbf{m}_{i,k}\in\mathbb{R}^{6\times 1}$ given by:

$$\begin{aligned} \mathbf{m}_{i,k} &= \frac{\partial^2 f_b(\hat{\mathbf{T}}(t);\theta_{b,k})}{\partial\hat{\mathbf{T}}(t)\,\partial\theta_{b,k}^i}\Big|_{t=k\tau}(\hat{\mathbf{A}}_1\,\hat{\mathbf{T}}(k\tau) - \bar{R}(k\tau)\,\hat{\mathbf{B}}(k\tau)\,\hat{\mathbf{B}}^T(k\tau) \\ & f_b(\hat{\mathbf{T}}(k\tau);\theta_{b,k}) + \hat{\mathbf{A}}_2\,\hat{\mathbf{T}}^4(k\tau) + \hat{\mathbf{c}}) - \bar{R}(k\tau)\,\frac{\partial f_b(\hat{\mathbf{T}}(t);\theta_{b,k})}{\partial\hat{\mathbf{T}}(t)}\Big|_{t=k\tau} \\ & \hat{\mathbf{B}}(k\tau)\,\hat{\mathbf{B}}^T(k\tau)\,\frac{\partial f_b(\hat{\mathbf{T}}(t);\theta_{b,k})}{\partial\theta_{b,k}^i}\Big|_{t=k\tau} + (\hat{\mathbf{A}}_1^T + 4\,\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t) \\ & \mathbf{I}_{1\times 6}))\,\frac{\partial f_b(\hat{\mathbf{T}}(t);\theta_{b,k})}{\partial\theta_{b,k}^i}\Big|_{t=k\tau}. \end{aligned} \tag{34}$$

The learned co-function $f_b(\hat{\mathbf{T}}(t);\theta_{b,k})$ is then applied at the remote temperature controller to generate the CPTS control solution $u(t)$. Algorithm 1 and Fig. 4(a) summarize online CPTS control scheme using the black-box RL.

[2]The implementation of Algorithms 1 and 2 requires the real-time CSI $h(k\tau)$, which can be obtained in practice via standard pilot-based channel estimation at the thermal plant, followed by channel feedback to the remote temperature controller [31].



(a) Summary of $\mathcal{A}_1$. (b) Summary of $\mathcal{A}_2$.

Fig. 4: Summary of online CPTS control using $\mathcal{A}_1$ and $\mathcal{A}_2$. $\mathcal{A}_2$ uses a DNN $f_s(\hat{\mathbf{T}}(t);\theta_s)$ to approximate only the high-order term $g(\hat{\mathbf{T}}(t))$, resulting in faster convergence and lower computational complexity compared to $\mathcal{A}_1$.

It is worth noting that an unstructured NN may result in slow convergence and high computational complexity. To overcome these limitations, we propose a *structured* learning strategy in the following subsection.

B. Structured RL for CPTS Control

1) Network Architecture: Note that the co-function $\lambda(\hat{\mathbf{T}}(t))$ admits a structured form, $\lambda(\hat{\mathbf{T}}(t)) = \mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + g(\hat{\mathbf{T}}(t))$, as shown in Theorem 3. Since the linear kernel $\{\mathbf{C}_0,\mathbf{C}_1\}$ can be computed offline analytically, online learning of $\lambda(\hat{\mathbf{T}}(t))$ reduces to learning the nonlinear component $g(\hat{\mathbf{T}}(t))$. Specifically, we approximate the co-function using a structured NN $f_s(\hat{\mathbf{T}}(t);\theta_s)$:

$$f_s(\hat{\mathbf{T}}(t);\theta_s) = \mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + \bar{f}_s(\hat{\mathbf{T}}(t);\theta_s), \tag{35}$$

where the NN $\bar{f}_s(\hat{\mathbf{T}}(t);\theta_s)\in\mathbb{R}^{6\times 1}$ approximates $g(\hat{\mathbf{T}}(t))$ in $\lambda(\hat{\mathbf{T}}(t))$ and is parameterized by $\theta_s\in\mathbb{R}^{l_s\times 1}$, with $l_s\in\mathbb{Z}_+$ denoting the number of neurons.

2) Design of Reward Function: Since the co-function $\lambda(\hat{\mathbf{T}}(t))$ satisfies (30), we reformulate it into the per-stage reward function $e_s(t;\theta_s)$ for θ_s as:

$$\begin{aligned} e_s(t;\theta_s) &= \frac{1}{2}\|\bar{\mathbf{e}}_s(t;\theta_s)\|^2 \\ &= \frac{1}{2}\Big\|\frac{\partial\bar{f}_s(\hat{\mathbf{T}}(t);\theta_s)}{\partial\hat{\mathbf{T}}(t)}\Big(\hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) - \bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t) \\ &\Big(\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + \bar{f}_s(\hat{\mathbf{T}}(t);\theta_s)\Big) + \hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) + \hat{\mathbf{c}}\Big) + \hat{\mathbf{A}}_1^T \\ &\bar{f}_s(\hat{\mathbf{T}}(t);\theta_s) + 4\hat{\mathbf{A}}_2^T \odot \hat{\mathbf{T}}^3(t)\mathbf{I}_{1\times 6}\Big(\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + \bar{f}_s(\hat{\mathbf{T}}(t);\theta_s)\Big) \\ &- \bar{R}(t)\mathbf{C}_1(\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)\bar{f}_s(\hat{\mathbf{T}}(t);\theta_s) + \hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t))\Big\|^2. \end{aligned} \tag{36}$$

The optimization problem for θ_s can be formulated as follows:

Problem 4 (Optimization for Structured NN)**.**

$$\min_{\theta_s} L_s(\theta_s) = \min_{\theta_s}\limsup_{T\to\infty}\frac{1}{T}\int_{t=0}^{T}\mathbb{E}[e_s(t;\theta_s)]dt, \tag{37}$$

where the expectation in (37) is w.r.t. the CSI $h(t)$ and the AWGN $v(t)$.

3) Training Process: To learn the CPTS control solution, TD learning is employed to optimize θ_s in Problem 4. Specifically, the NN parameter $\theta_{s,k} = [\theta_{s,k}^1,\ldots,\theta_{s,k}^{l_s}]^T\in\mathbb{R}^{l_s\times 1}$ is

Algorithm 1 Online CPTS Control via Black-Box Approach

Initialization:

- Initializing $f_b(\hat{\mathbf{T}}(t);\theta_{b,0}) \leftarrow$ using $\theta_{b,0} = [\theta_{b,0}^1, ..., \theta_{b,0}^{l_b}]^T$.
- Initializing $u(t) \leftarrow -\bar{R}(0)\hat{\mathbf{B}}^T(0) f_b(\hat{\mathbf{T}}(0);\theta_{b,0}), 0 \le t < \tau$, using $f_b(\hat{\mathbf{T}}(t);\theta_{b,0})$.

For $k = 1, 2, 3, ...$**:**

- **Step 1 (Update of the Unstructured NN):** $\theta_{b,k+1} \leftarrow$ according to (33) using $\hat{\mathbf{T}}(k\tau)$, $h(k\tau)$, and $\theta_{b,k}$.
- **Step 2 (Update of the CPTS Solution):** $u(t) \leftarrow -\bar{R}(k\tau)\hat{\mathbf{B}}^T(k\tau) f_b(\hat{\mathbf{T}}(k\tau);\theta_{b,k}), k\tau \le t < (k+1)\tau$.

End

updated at each sampled timeslot k using real-time CSI $h(k\tau)$ and temperature feedback $\mathbf{T}(k\tau)$, as follows.

$$\begin{aligned}\theta_{s,k+1} &= \theta_{s,k} - \nabla_{\theta_{s,k}} e_s(k\tau;\theta_{s,k}) \\ &= \theta_{s,k} - \alpha_k \mathbf{n}_k.\end{aligned} \tag{38}$$

where $\mathbf{n}_k = [\mathbf{n}_{1,k}^T \bar{\mathbf{e}}_s(k\tau;\theta_{s,k}), ..., \mathbf{n}_{l_s,k}^T \bar{\mathbf{e}}_s(k\tau;\theta_{s,k})]^T \in \mathbb{R}^{l_s \times 1}$ with each $\mathbf{n}_{i,k} \in \mathbb{R}^{6\times 1}$ given by:

$$\begin{aligned}\mathbf{n}_{i,k} &= \frac{\partial^2 \bar{f}_s(\hat{\mathbf{T}}(t);\theta_{s,k})}{\partial \hat{\mathbf{T}}(t) \partial \theta_{s,k}^i}|_{t=k\tau} \Big(\hat{\mathbf{A}}_1 \hat{\mathbf{T}}(k\tau) - \bar{R}(k\tau)\hat{\mathbf{B}}(k\tau)\hat{\mathbf{B}}^T(k\tau) \\ &\Big(\mathbf{C}_0 + \mathbf{C}_1 \hat{\mathbf{T}}(k\tau) + \bar{f}_s(\hat{\mathbf{T}}(k\tau);\theta_{s,k})\Big) + \hat{\mathbf{A}}_2 \hat{\mathbf{T}}^4(k\tau) + \hat{\mathbf{c}}\Big) \\ &- \frac{\partial \bar{f}_s(\hat{\mathbf{T}}(t);\theta_{s,k})}{\partial \hat{\mathbf{T}}(t)}|_{t=k\tau} \bar{R}(k\tau)\hat{\mathbf{B}}(k\tau)\hat{\mathbf{B}}^T(k\tau) \frac{\partial \bar{f}_s(\hat{\mathbf{T}}(t);\theta_{s,k})}{\partial \theta_{i,k}}|_{t=k\tau} \\ &+ \hat{\mathbf{A}}_1^T \frac{\partial \bar{f}_s(\hat{\mathbf{T}}(t);\theta_{s,k})}{\partial \theta_{s,k}^i}|_{t=k\tau} + \Big(4\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(k\tau)\mathbf{I}_{1\times 6})\Big) \\ &\frac{\partial \bar{f}_s(\hat{\mathbf{T}}(t);\theta_{s,k})}{\partial \theta_{s,k}^i}|_{t=k\tau} - \bar{R}(k\tau)\mathbf{C}_1\hat{\mathbf{B}}(k\tau)\hat{\mathbf{B}}^T(k\tau) \\ &\frac{\partial \bar{f}_s(\hat{\mathbf{T}}(t);\theta_{s,k})}{\partial \theta_{s,k}^i}|_{t=k\tau}, 1 \le i \le l_s.\end{aligned} \tag{39}$$

The learned structured co-function $f_s(\hat{\mathbf{T}}(t);\theta_{s,k})$ can be deployed at the remote temperature controller to determine the CPTS control solution $u(t)$. The structured learning algorithm for CPTS control is outlined in Algorithm 2 and Fig. 4(b).

Algorithm 2 Online CPTS Control via Structured Approach

Initialization:

- Initializing $f_s(\hat{\mathbf{T}}(t);\theta_{s,0}) \leftarrow \mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + \bar{f}_s(\hat{\mathbf{T}}(t);\theta_{s,0})$ using $\theta_{s,0} = [\theta_{s,0}^1, ..., \theta_{s,0}^{l_s}]^T$.
- Initializing $u(t) \leftarrow -\bar{R}(0)\hat{\mathbf{B}}^T(0) f_s(\hat{\mathbf{T}}(0);\theta_{s,0}), 0 \le t < \tau$ using $f_s(\hat{\mathbf{T}}(t);\theta_{s,0})$.

For $k = 1, 2, 3, ...$**:**

- **Step 1 (Update of the Structured NN):** $\theta_{s,k+1} \leftarrow$ according to (38) using $\hat{\mathbf{T}}(k\tau)$, $h(k\tau)$, and $\theta_{s,k}$.
- **Step 2 (Update of the Control Solution):** $u(t) \leftarrow -\bar{R}(k\tau)\hat{\mathbf{B}}^T(k\tau) \bar{f}_s(\hat{\mathbf{T}}(k\tau);\theta_{s,k}), k\tau \le t < (k+1)\tau$.

End

By incorporating the structure of the co-function $\lambda(\hat{\mathbf{T}}(t))$ into the design of the structured NN $f_s(\hat{\mathbf{T}}(t);\theta_s)$, Algorithm 2 ensures fast convergence and reduced computational complexity, as analyzed in the following subsection.

Remark 4 (RL Category of Our Method and its Advantages Compared to Conventional RL Approaches)**.** *The proposed scheme belongs to the class of continuous-state value-based reinforcement learning methods. Instead of directly parameterizing the control policy $u(t)$ in an end-to-end manner, we exploit the structural properties of the thermal tracking problem and learn the co-function $\lambda(\hat{\mathbf{T}}(t))$, which is the state derivative of the value function $V(\hat{\mathbf{T}}(t))$. By leveraging the fixed-point optimality condition of the HJB equation, we construct a structured parameterization of $\lambda(\cdot)$ that incorporates problem-specific forms and approximate only the residual unknown terms using a NN. This design yields faster convergence, lower computational complexity, and better data efficiency compared with conventional methods such as Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimization (PPO), and Soft Actor–Critic (SAC), which rely on generic black-box policy and value function approximators without exploiting problem structure.*

C. Convergence Analysis

Denote Algorithm 1 and Algorithm 2 as $\mathcal{A}_b$ and $\mathcal{A}_s$, respectively. The convergence of $\mathcal{A}_i$ depends on that of the learned co-functions $f_i(\hat{\mathbf{T}}(t);\theta_i)$, where $i \in \{b, s\}$. Thus, we analyze the convergence of $\mathcal{A}_i$ by examining $f_i(\hat{\mathbf{T}}(t);\theta_i)$ via Lyapunov analysis [32], as stated in the following theorem.

Theorem 4 (Convergence of $\mathcal{A}_b$ and $\mathcal{A}_s$)**.** *Let $i \in \{b, s\}$. If the following conditions hold:*

- *$f_i(\hat{\mathbf{T}}(t);\theta_i)$ is Lipschitz continuous w.r.t. θ_i, i.e., there exists $L > 0$ such that for all $\theta_{i,1}, \theta_{i,2} \in \mathbb{R}^{l_i \times 1}$,*
$$|f_i(\hat{\mathbf{T}}(t);\theta_{i,1}) - f_i(\hat{\mathbf{T}}(t);\theta_{i,2})|^2 \le L\|\theta_{i,1} - \theta_{i,2}\|^2.$$
- *The gradient satisfies:*
$$\mathbb{E}[\|\nabla_{\theta_i} f_i(\hat{\mathbf{T}}(t);\theta_i)\|^2] < c < \infty.$$

Then, $\mathcal{A}_i$ converges to the optimal point or a saddle point of the loss function $L_i(\theta_i)$ w.p.1, provided that $L_i(\theta_i)$ is convex or non-convex w.r.t. θ_i, respectively. Furthermore, if $\mathcal{A}_i$ converges to the optimal point θ_i^* of $L_i(\theta_i)$ with $L_i(\theta_i^*) = 0$, then the control solution $u(t)$ obtained via $\mathcal{A}_i$ converges to the optimal control solution $u^*(t)$ in (18) w.p.1, i.e.,

$$\limsup_{k\to\infty} \Pr(u(k\tau) = u^*(k\tau)) = 1.$$

Proof: See Appendix F. ∎

Theorem 4 reveals that reliable convergence of online RL algorithms ($\mathcal{A}_b$, $\mathcal{A}_s$) for CPTS over wireless networks critically depends on the smoothness and bounded gradient of NNs $f_i(\hat{\mathbf{T}}(t);\theta_i)$. Such conditions naturally arise in neural approximators such as fully-connected networks with bounded activation functions (e.g., sigmoid, tanh) or ReLU networks with weight clipping [33], ensuring robust RL-based temperature control despite nonlinear dynamics and wireless signal distortions.

We compare the convergence speeds of $\mathcal{A}_b$ and $\mathcal{A}_s$ when they converge, quantifying the performance gain of the structured approach over the black-box approach for CPTS control. To ensure fairness, we impose the same NN approximation performance for the co-function $\lambda(\hat{\mathbf{T}}(t))$ in both methods, where the unstructured NN $f_b(\hat{\mathbf{T}}(t);\theta_b)$ in $\mathcal{A}_b$ and the structured NN $f_s(\hat{\mathbf{T}}(t);\theta_s)$ in $\mathcal{A}_s$ approximate $\lambda(\hat{\mathbf{T}}(t))$ with equal accuracy.

By leveraging the universal approximation theorem [34], we establish the following lemma, which characterizes the relationship between the structures of the NNs $f_i(\hat{\mathbf{T}}(t);\theta_i), i \in \{b,s\}$, and their respective approximation error bounds for $\lambda(\hat{\mathbf{T}}(t))$ under the structured and black-box approaches.

Lemma 2 (Universal Approximation for the Co-Function)**.** *The co-function $\lambda(\hat{\mathbf{T}}(t))$ can be approximated by the NN $f_i(\hat{\mathbf{T}}(t);\theta_i), i \in \{b,s\}$ with arbitrary accuracy over a compact set Θ. Specifically, for any $\epsilon > 0$, there exists a structured NN $f_s(\hat{\mathbf{T}}(t);\theta_s) = \mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + f_s(\hat{\mathbf{T}}(t);\theta_s)$, where $f_s(\hat{\mathbf{T}}(t);\theta_s)$ is a DNN component with width $W_s \in [W_s^*, \infty)$ and depth $L = 2\lfloor \log_2 6 \rfloor + 2$, or an unstructured NN $f_b(\hat{\mathbf{T}}(t);\theta_b)$ with width $W_b \in [\lceil \frac{W_s \times L+42}{L} \rceil, \infty)$ and depth L, such that $\sup_{\hat{\mathbf{T}}(t)\in\Theta} \|f_i(\hat{\mathbf{T}}(t);\theta_i) - \lambda(\hat{\mathbf{T}}(t))\|_2^2 \leq \epsilon$, where $W_s^* \in \mathbb{Z}_+$ is the minimum achievable width.*

Proof: See Appendix G. ∎

In both the CPTS control problem solved by the black-box RL method (Problem 3) and that solved by the proposed structure-aware method (Problem 4), the optimality condition satisfies $\nabla_{\theta_i} L_i(\theta_i^*) = 0$, where θ_i^* denotes the optimal parameter of the corresponding problem. Thus, the squared gradient norm $\mathbb{E}[\|\nabla_{\theta_{i,k}} L_i(\theta_{i,k})\|^2]$ serves as a standard measure of the optimization error, indicating how far the current iterate is from the optimum. Consequently, we employ an upper bound on the gradient norm (also referred to as a gradient-error bound in stochastic optimization) to quantify and compare the convergence speeds of $\mathcal{A}_b$ and $\mathcal{A}_s$, as summarized below.

Corollary 1 (Convergence Rate of Learning Algorithms)**.** *Let $i \in \{b,s\}$ and let $f_i(\hat{\mathbf{T}}(t);\theta_i)$ denote a NN of width W_i and depth L with approximation error ϵ (Lemma 2). Let θ_b^* and θ_s^* be the optimal solutions to Problems 3 and 4, respectively. If both $\mathcal{A}_b$ and $\mathcal{A}_s$ converge, then the convergence rate of algorithm $\mathcal{A}_i$ is given by:*

$$\inf_{0\leq k\leq K} \mathbb{E}\big[\|\nabla_{\theta_{i,k}} L_i(\theta_{i,k})\|^2\big] \leq \frac{L_i(\theta_{i,0}) - L_i(\theta_i^*)}{c\sum_{k=1}^K \alpha_k} + \frac{\sum_{k=1}^K \alpha_k^2\, \mathbb{E}\big[\|\nabla_{\theta_{i,k}} e_i(k\tau;\theta_{i,k})\|^2\big]}{c\sum_{k=1}^K \alpha_k}. \quad (40)$$

Furthermore, the difference in convergence rates between $\mathcal{A}_b$ and $\mathcal{A}_s$ satisfies

$$\mathbb{E}\left[\|\nabla_{\theta_{b,k}} L_b(\theta_{b,k})\|^2\right] - \mathbb{E}\left[\|\nabla_{\theta_{s,k}} L_s(\theta_{s,k})\|^2\right] \sim \mathcal{O}\left(\frac{\sum_{t=0}^k \alpha_t^2}{\sum_{t=0}^k \alpha_t}\right). \quad (41)$$

Proof: See Appendix F. ∎

The result in Corollary 1 indicates that the structured approach $\mathcal{A}_s$ achieves a lower gradient norm at each iteration compared to the black-box approach $\mathcal{A}_b$. Since gradient norm reduction is directly linked to convergence speed of the algorithms, this implies that $\mathcal{A}_s$ converges faster than $\mathcal{A}_b$.

D. Computational Complexity Analysis

We further compare the computational complexities of $\mathcal{A}_b$ and $\mathcal{A}_s$ under the condition that their NNs attain the same approximation accuracy. The result is summarized in the following corollary.

Corollary 2 (Computational Complexity Difference between $\mathcal{A}_b$ and $\mathcal{A}_s$)**.** *Let $d = \dim(\hat{\mathbf{T}}(t))$, and let the NN structures for $f_i(\hat{\mathbf{T}};\theta_i)$ be the same as in Corollary 1. The computational complexity ratio between $\mathcal{A}_b$ and $\mathcal{A}_s$ is $O((W_s(2\lfloor \log_2 d \rfloor + 2))^{\gamma-2})$, where $\gamma \geq 4$.*

Proof: See Appendix H. ∎

Corollary 2 highlights the computational advantage of the proposed structured method over the black-box approach. Since $\gamma \geq 4$, the complexity reduction is at least $O(W_s^2 L^2)$, demonstrating substantial efficiency gains. For large W_s, the benefit becomes even more pronounced, indicating that the proposed structured scheme scales more favorably and achieves significantly lower computational cost as the network width increases.

V. Numerical Results

In this section, we investigate a representative CPTS control scenario, namely the temperature regulation of a furnace system for slab reheating over wireless networks. Specifically, the furnace system consists of a furnace plant (corresponding to the thermal plant in Fig. 1) and a remote temperature controller, which may be implemented on a programmable logic controller (PLC), a distributed control system (DCS), or an edge computing device. The furnace plant is composed of a heater (the heat source in Fig. 1) and a reheating slab (the target object in Fig. 1) equipped with thermal sensors that measure the top-surface temperature $[\mathbf{T}(t)]_1$, the center-point temperature $[\mathbf{T}(t)]_2$, and the bottom-surface temperature $[\mathbf{T}(t)]_3$ of the slab. The slab temperature state is represented as $\mathbf{T}(t) = \big[[\mathbf{T}(t)]_1,\ [\mathbf{T}(t)]_2,\ [\mathbf{T}(t)]_3\big]^T$, which is fed back to the remote controller for generating the fuel control command $u(t)$. This command is transmitted to the heater through an unreliable wireless network, and the received noisy command $\widehat{u}(t)$ is subsequently used by the heater to regulate the fuel output and provide heat for reheating the slab. The control objective is to regulate the slab temperature $\mathbf{T}(t)$ toward the desired temperature profile $\mathbf{r}(t)$ over time by adjusting the heater fuel supply $\widehat{u}(t)$ according to the control signals $u(t)$ generated by the remote controller.

Under this typical industrial control practice, we validate the performance advantages of the proposed structured control algorithm by benchmarking it against several baseline methods, which are summarized below.

- **Baseline 1** *(PID-based Control [10])*: $u(t) = \mathbf{K}_p\mathbf{e}(k) + \mathbf{K}_i\sum_{i=0}^k \mathbf{e}(i-1) + \mathbf{K}_d(\mathbf{e}(k) - \mathbf{e}(k-1))$, with error $\mathbf{e}(k) = \mathbf{T}(k\tau) - \mathbf{r}(k\tau)$, and $\mathbf{T}(-\tau) = \mathbf{0}_{3\times 1}$. PID gains are tuned offline under static fading $h(t) = 1$.
- **Baseline 2** *(LQT Control over Static Channels [11])*: $u(t) = \mathbf{D}\hat{\mathbf{T}}(t)$, with static gain $\mathbf{D} \in \mathbb{R}^{1\times 6}$ computed offline via LQT, assuming $h(t) = 1$.
- **Baseline 3** *(LQT Control over Fading Channels [7])*: $u(t) = \hat{\mathbf{D}}(h(k\tau), k)\hat{\mathbf{T}}(k\tau)$, where $\hat{\mathbf{D}} \in \mathbb{R}^{1\times 6}$ is computed online based on real-time fading $h(k\tau)$ via LQT.
- **Baseline 4** *(NN-based Control via DRL)*: $u(t)$ is given by the online black-box DRL algorithm in Algorithm 1.
- **Baseline 5** *(MPC [12])*: At each k-th timeslot, the thermal dynamics are linearized around $\mathbf{T}(k\tau)$. Based

on the local linear model, the control input sequence $\{u(k\tau), \ldots, u((k+H-1)\tau)\}$ is obtained by solving a finite-horizon quadratic programming (QP) of length $H = 20$, with per-stage cost in Problem 2. Only the first input $u(k\tau)$ is applied, and the optimization is repeated at the next timeslot with updated states.

- **Baseline 6** *(Model-Assisted DDPG [35])*: In the offline pre-training phase, synthetic state–action–reward samples generated from the thermal model ($N_{\text{offline}} = 100$ trajectories of length $L = 200$) are used to pre-train the actor–critic networks. In the online fine-tuning phase, real interaction data from the CPTS are stored in a replay buffer of size $M = 1000$, from which mini-batches of 128 samples are drawn to update the networks. Training continues until the reward averaged over the latest 20 timeslots stabilizes (change $< 1\%$) or until 10^4 steps are reached. After training, the actor directly outputs the control input $u(t)$ from the current temperature $\mathbf{T}(t)$.

Unless otherwise specified, the default parameters of the furnace system are configured as follows. The furnace plant has an effective thermal cavity of dimensions $1\ \text{m} \times 1\ \text{m} \times 1\ \text{m}$, which is a typical scale used in existing furnace studies such as [36]. The reheating slab measures $0.5\ \text{m} \times 0.1\ \text{m} \times 0.5\ \text{m}$, consistent with slab sizes commonly reported in the literature (e.g., see [37], where the height, length, and width are all below 1 m). The slab is characterized by a mass density $\rho = 7800\ \text{kg/m}^3$, a specific heat capacity $c = 460\ \text{J/(kg} \cdot \text{K)}$, and a thermal conductivity $\lambda_s = 48.5\ \text{W/(m} \cdot \text{K)}$, which fall within the typical property range of steel slabs [38]. The heat transfer coefficient between the slab and the heater is set to $10\ \text{W/(m}^2 \cdot \text{K)}$, consistent with values reported in [39], and the emissivity is $\epsilon_0 = 0.4$, a representative value for furnace temperature control [40]. The Stefan–Boltzmann constant is $\sigma = 6.7 \times 10^{-8}\ \text{W/(m}^2 \cdot \text{K}^4\text{)}$, and the boundary heat fluxes are $q_0 = q_1 = 1\ \text{W/m}^2$. The sampling period of the training iterations is $\tau = 1$ ms. The initial and target temperatures are $\mathbf{T}_0 = [100, 100, 100]^T\ ^\circ\text{C}$ and $\mathbf{r}_0 = [500, 500, 500]^T\ ^\circ\text{C}$, respectively. We set $\mathbf{G} = \mathbf{0}_3$, $\mathbf{Q} = \mathbf{I}_3$, and $R = M = 1$.

Simulations were conducted on an Intel i7-9700K CPU. By default, $\bar{f}_s(\hat{\mathbf{T}}(t); \theta_s)$ in our scheme consists of four layers: an input layer (fully connected, $n_1 = 32$), two hidden layers (ReLU, $n_2 = 32$; fully connected, $n_3 = 16$), and an output layer (fully connected, $n_4 = 6$). For comparison, $f_b(\hat{\mathbf{T}}(t); \theta_b)$ in Baseline 4 has five layers: an input layer (fully connected, $n_1 = 32$), two hidden layers (ReLU, $n_2 = 128$; fully connected, $n_3 = 16$), and an output layer (fully connected, $n_4 = 6$). In Baseline 6, both the actor and critic networks consist of three fully connected layers of widths $(64, 64, 1)$. Note that the above NNs are differentiable almost everywhere. At points where the pre-activation of a ReLU unit equals zero and differentiability may fail, standard subgradient methods can be employed by selecting the zero subgradient [41], thereby preserving the validity of the learning algorithms.

A. Temperature Tracking Control Performance Analysis

We evaluate the temperature control performance of the proposed method and the baselines using the top-surface temperature trajectory of the slab, as shown in Fig. 5. Baselines 1, 2, 3, and 5 diverge because they either ignore the nonlinear furnace dynamics or overlook wireless impairments. In contrast, Baselines 4 and 6 and the proposed scheme account for these factors and therefore converge to the target profile. The proposed scheme reaches the target in about 50 minutes, which is nearly twice as fast as Baselines 4 and 6. This improvement results from exploiting the structural properties of the optimal solution in the structured reinforcement learning design, whereas Baselines 4 and 6 rely on generic algorithms that do not use the problem structure.

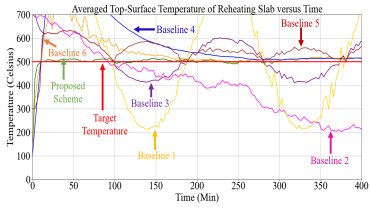


Fig. 5: Time evolution of the reheating slab temperature, averaged over 100 simulation runs. The received SNR at the heater is 0 dB.

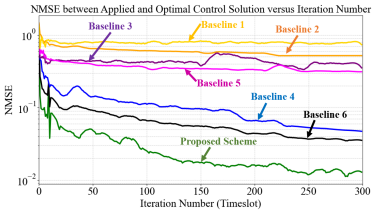


Fig. 6: Convergence behavior of the control solutions via proposed and baseline schemes. The received SNR at the heater is 0 dB.

B. Convergence Performance Analysis

Fig. 6 illustrates the convergence behavior of the control solutions via proposed and baseline schemes by plotting the normalized mean square error (NMSE) between the applied control solution $u(t)$ and the optimal solution $u^*(t)$ as a function of the training iteration number. The expectation is approximated by averaging over 100 independent simulation runs with random initial seeds. Baselines 1–3 and 5 progressively deviate from optimality due to their inability to capture the nonlinear thermal dynamics and wireless signal distortions. In contrast, Baselines 4, 6, and the proposed scheme asymptotically converge to $u^*(t)$. Nevertheless, the proposed approach achieves substantially faster convergence by exploiting the structural properties of the optimal solution to design a structured learning algorithm, whereas Baselines 4 and 6 rely on generic black-box methods without leveraging the problem structure.

Fig. 7 illustrates the convergence behavior of the averaged reward $-\mathbb{E}\big[\|\mathbf{T}(t) - \mathbf{r}(t)\|^2 + u^2(t)\big]$, which corresponds to the negative of the averaged per-stage cost function $\mathbb{E}[r^a(\hat{\mathbf{T}}(t), h(t), u(t))]$ in Problem 2, as a function of the training iteration number. The expectation is approximated by averaging over 100 independent training runs with different random seeds. We examine three configurations of the hidden-layer size of the DNN $\bar{f}_s(\hat{\mathbf{T}}(t); \theta_s)$, where the number of

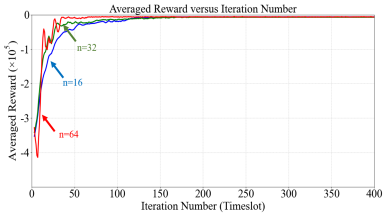


Fig. 7: Convergence behavior of the averaged reward under different NN architecture. The received SNR is 10 dB. $\lambda_s = 58.5$ W/(m·K).

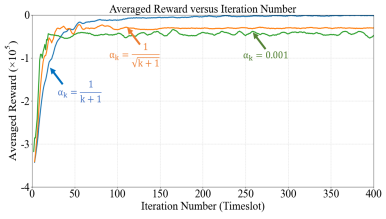


Fig. 8: Convergence behavior of the averaged reward under different learning-rate rules. The received SNR is 10 dB. $\lambda_s = 58.5$ W/(m·K).

neurons is set to $n = 16, 32$, and 64. As observed, the proposed scheme consistently converges across different DNN sizes. A larger hidden-layer size accelerates convergence but also induces larger fluctuations during training.

Fig. 8 illustrates the convergence behavior of the averaged reward as a function of the training iteration number under various learning-rate rules for the proposed Algorithm 2. As shown from the figure, the constant learning rate $\alpha_k = 0.001$ yields the fastest initial convergence but suffers from large oscillations and does not converge exactly to the optimum due to persistent noise. The classical diminishing learning rate $\alpha_k = \frac{1}{k+1}$ guarantees smooth and stable convergence with the smallest variance but exhibits the slowest convergence speed. The intermediate rule $\alpha_k = \frac{1}{\sqrt{k+1}}$ achieves a balance between the two, converging faster than $\frac{1}{k+1}$ while maintaining smaller fluctuations than the constant learning rate.

C. Robustness Performance of the Proposed Scheme

Fig. 9 illustrates the robustness of the proposed scheme under varying thermal and channel conditions by depicting the time evolution of the reheating slab temperature, averaged over 100 simulation runs across four scenarios. Scenario 1 assumes a wired connection ($h(t) = 1$, $v(t) = 0$) with a conductivity of $48.5\,\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$. Scenario 2 introduces a wireless connection with an SNR of 20 dB while maintaining the same conductivity. Scenario 3 retains the 20 dB SNR but increases the

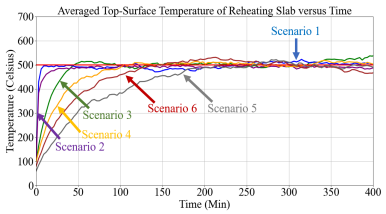


Fig. 9: Robustness performance of the proposed scheme across thermal and channel parameters.

conductivity to $58.5\,\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})$. Scenario 4 further reduces the SNR to 10 dB while keeping the higher conductivity. Scenarios 5 and 6 follow the same setup as Scenario 4 but additionally include sensing noise (additive Gaussian with unit variance) with a 30 ms communication delay, and imperfect CSI (NMSE $= -10$ dB), respectively. The results demonstrate that the proposed scheme consistently converges to the target temperature profile in all cases, highlighting its robustness against thermal variations and wireless channel distortions.

Fig. 10 illustrates the robustness of the proposed scheme under varying furnace sizes by plotting the mean square error (MSE) of temperature tracking against the effective height of the furnace thermal cavity, l_3, and comparing it with the baseline schemes. The expectation is approximated by averaging over 100 independent sample runs and 10^4 timeslots. As shown in the figure, increasing l_3 enlarges the thermal inertia of the reheating slab, making temperature regulation more challenging. Nevertheless, the proposed scheme maintains an MSE on the order of 10^{-2}, representing an improvement of roughly 20 dB over all baselines and demonstrating strong robustness across different furnace sizes. Specifically, Baselines 1–2 neglect wireless fading and thus diverge, while Baselines 3–6 incorporate fading but either linearize the thermal model or ignore structural properties, leading to slow or inaccurate tracking. In contrast, the proposed scheme jointly handles wireless fading and nonlinear thermal dynamics and leverages structural insights in the network design, achieving fast, accurate, and robust tracking across furnace sizes.

Fig. 11 shows the robustness of the proposed scheme under different slab materials by plotting the MSE of temperature tracking versus the thermal conductivity λ_s and comparing it with the baselines. The expectation is estimated over 100 independent runs and 10^4 timeslots. As λ_s increases, the MSE decreases because higher conductivity enables faster heat transfer and a more uniform temperature distribution. Across all conductivity levels, the proposed scheme consistently achieves much lower MSE than the baselines, for the same reasons discussed in Fig. 10, demonstrating robustness to variations in slab material properties.

Fig. 12 shows the robustness of the proposed scheme under different network conditions by plotting the MSE of temperature tracking versus the received SNR at the heater and comparing it with the baselines. The expectation is estimated over 100 independent runs and 10^4 timeslots. As the received SNR increases, the MSE decreases because higher SNR reduces communication errors and improves the reliability of control-signal transmission. Across all SNR levels, the proposed scheme consistently achieves much lower MSE than the baselines, for the same reasons discussed in Fig. 10, demonstrating strong robustness to network variations.

Fig. 13 shows the robustness of the proposed scheme under four channel models, plotting the slab temperature averaged over 100 runs. Scenario 1 uses i.i.d. Gaussian fading $h(t) \sim \mathcal{N}(0,1)$. Scenario 2 adopts Gauss–Markov fading $h(t+1) = \rho h(t) + \sqrt{1-\rho^2}\, w(t)$ with $\rho = 0.5$, $w(t) \sim \mathcal{N}(0,1)$. Scenario 3 adds log-normal shadowing $h(t) = h_s(t) \cdot 10^{X/20}$ with $X \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 6$ dB. Scenario 4 models burst errors via a Gilbert–Elliott channel switching between good ($h(t) = 1$)

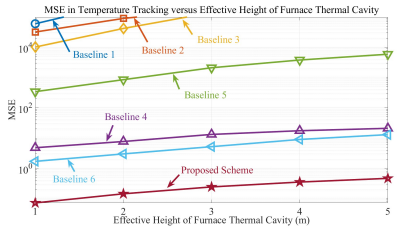


Fig. 10: Robustness performance of the proposed scheme across furnace sizes. $\lambda_s = 48$ W/(m·K). The received SNR at the heater is 0 dB.

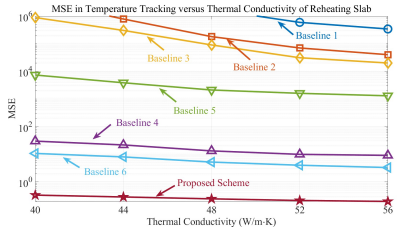


Fig. 11: Robustness performance of the proposed scheme across furnace thermal conductivity. The effective height of the furnace thermal cavity for the reheating slab is $l_3 = 1$m. The received SNR at the heater is 0 dB.

and bad ($h(t) = 0$) states with transition probabilities 0.4 and 0.6. Across all cases, the proposed scheme preserves stable tracking, demonstrating robustness beyond i.i.d. fading by adapting to instantaneous $h(t)$ rather than its distribution, and generalizing effectively under correlation, shadowing, and burst errors.

D. Computational Complexity Analysis of Proposed Scheme

We consider a uniform network width by setting $n_1 = n_2 = n_3 = n$ for $\bar{f}_s(\hat{\mathbf{T}}(t); \theta_s)$. At each timeslot, the computational cost of the proposed CPTS control algorithm is dominated by the forward inference and backward propagation of the NN $\bar{f}_s(\hat{\mathbf{T}}(t); \theta_s)$ for control policy update and execution. Both

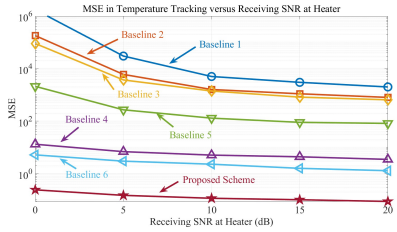


Fig. 12: Robustness performance of the proposed scheme across network conditions. The effective height of the furnace thermal cavity for the reheating slab is $l_3 = 1$m. $\lambda_s = 48$ W/(m·K).

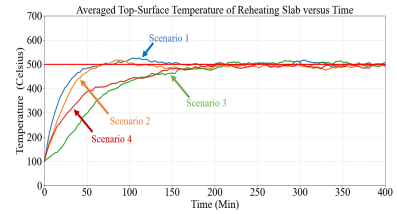


Fig. 13: Robustness performance of the proposed scheme across channel models. The configurations are the same as those in Scenario 4 of Fig. 9, except for the channel models.

Number of Neurons n	CPU Time for Interference of the NN (ms)
16	0.277
32	0.431
64	0.724

TABLE I: CPU Computational Time for NN Inference in the Proposed Algorithm

operations are dominated by fully connected transformations and scale as $\mathcal{O}(n^2)$. Therefore, the per-time-slot computational complexity of the proposed algorithm is $\mathcal{O}(n^2)$. This analysis is supported by the CPU timing results in Table I for $n = 16$, 32, and 64, where the execution time increases with n yet remains below 1 ms in all cases, aligning with the shortest transmission time interval (TTI) in 5G systems. This confirms the real-time feasibility of the proposed scheme.

VI. Conclusions

In this work, we proposed a model-assisted online structured DRL algorithm for temperature control in CPTS with nonlinear dynamics over wireless fading channels. The control problem was formulated by incorporating nonlinear heat transfer in the thermal plant together with signal distortions in the wireless link between the remote controller and the plant. By leveraging the homotopy perturbation method, we exploited structural properties of the optimal solution and developed an efficient structured RL algorithm to approximate it. Analysis of the structured NN parameter updates shows that the proposed scheme achieves fast convergence with low computational complexity. Simulations further demonstrate superior tracking accuracy, computational efficiency, and robustness to wireless impairments compared with existing baselines. Beyond CPTS, the proposed framework applies broadly to CPS settings that require real-time decision-making under stochastic communication constraints, highlighting its potential for robust real-time signal processing in general industrial CPS.

Appendix

A. Derivations of (10) and Proof of $\lambda(\mathbf{A}_1) < 0$, $\lambda(\mathbf{A}_2) = 0$

1) Derivations of (10): We define the Sobolev space $\mathcal{S} := \mathcal{K}^1(-\frac{j_2}{2}, \frac{j_2}{2})$ and the bilinear form $b(s_1, s_2) = \int_{-\frac{j_2}{2}}^{\frac{j_2}{2}} s_1 s_2 \, dy : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. Let $t \in \mathbb{R}_+$. For any trial function $s(y) \in \mathcal{S}$ and scalars $s^-, s^+ \in \mathbb{R}$, the heat transient equation (5) can be equivalently formulated as:

$$
\begin{aligned}
0 = {} & \rho c \, b(s(y), \frac{\partial T^s(y,t)}{\partial t}) + \lambda_s \, b(\frac{\partial s(y)}{\partial y}, \frac{\partial T^s(y,t)}{\partial y}) \\
& + (s^+ - s(\frac{j_2}{2}))\lambda_s \frac{\partial T^s(y,t)}{\partial y}\Big|_{y=\frac{j_2}{2}} - s^+ q^s(\frac{j_2}{2}, t) \\
& - (s^- - s(-\frac{j_2}{2}))\lambda_s \frac{\partial T^s(y,t)}{\partial y}\Big|_{y=-\frac{j_2}{2}} - s^- q^s(-\frac{j_2}{2}, t). \qquad (42)
\end{aligned}
$$

The temperature field of the target object is approximated as $T^s(y,t) = \sum_{i=1}^{3} x_i(t) h_i(y)$, where the Galerkin coefficients are defined as $h_1(y) = 1$, $h_2(y) = \frac{2y}{j_2}$, and $h_3(y) = \left(\frac{2y}{j_2}\right)^2 - \frac{1}{3}$, capturing the mean temperature, asymmetry, and transient

inhomogeneity effects, respectively. Substituting $T^s(y,t)$ into (42), $t \geq 0$, we obtain an equivalent formulation of (5):

$$\dot{\mathbf{x}}(t) = \tilde{\mathbf{A}}_1\mathbf{x}(t) + \tilde{\mathbf{B}}_1^- q^s(-\frac{j_2}{2},t) + \tilde{\mathbf{B}}_1^+ q^s(\frac{j_2}{2},t), \tag{43}$$

where $\mathbf{x}(t) = [x_1(t), x_2(t), x_3(t)]^T \in \mathbb{R}^{3\times 1}$, $\tilde{\mathbf{A}}_1 = -\frac{12\lambda_s}{\rho c j_1^2 j_3^2}\,\mathrm{Diag}(0,1,5) \in \mathbb{R}^{3\times 3}$, $\tilde{\mathbf{B}}_1^+ = \frac{1}{\rho c j_1 j_3}[1,3,15/2]^T \in \mathbb{R}^{3\times 1}$, and $\tilde{\mathbf{B}}_1^- = \frac{1}{\rho c j_1 j_3}[1,-3,15/2]^T \in \mathbb{R}^{3\times 1}$.

This further leads to:

$$\dot{\mathbf{T}}(t) = \bar{\mathbf{A}}_1\mathbf{T}(t) + \bar{\mathbf{B}}_1^- q^s(-\frac{j_2}{2},t) + \bar{\mathbf{B}}_1^+ q^s(\frac{j_2}{2},t), \tag{44}$$

where $\bar{\mathbf{A}}_1 = \begin{bmatrix} \frac{-54\lambda_s}{\rho c j_2^2} & \frac{95\lambda_s}{\rho c j_2^2} & \frac{-42\lambda_s}{\rho c j_2^2} \\ \frac{30\lambda_s}{\rho c j_2^2} & \frac{-60\lambda_s}{\rho c j_2^2} & \frac{30\lambda_s}{\rho c j_2^2} \\ \frac{54\lambda_s}{\rho c j_2^2} & \frac{-96\lambda_s}{\rho c j_2^2} & \frac{42\lambda_s}{\rho c j_2^2} \end{bmatrix} \in \mathbb{R}^{3\times 3}$.

Using the energy balance equation (9), along with the heat convection (1) and radiation equations (3), we derive:

$$\begin{aligned} q^s(\frac{j_2}{2},t) &= \frac{Q(t)}{2j_1j_3} - \frac{(2l_1l_2+2l_2l_3+2l_1l_3)(q_0+q_1)}{2j_1j_3} \\ &+ \frac{h_0}{2}T^s(-\frac{j_2}{2},t) + \frac{\epsilon_0\sigma}{2}(T^s(-\frac{j_2}{2},t))^4 \\ &- \frac{h_0}{2}T^s(\frac{j_2}{2},t) - \frac{\epsilon_0\sigma}{2}(T^s(\frac{j_2}{2},t))^4, \end{aligned} \tag{45}$$

$$\begin{aligned} q^s(-\frac{j_2}{2},t) &= \frac{Q(t)}{2j_1j_3} - \frac{(2l_1l_2+2l_2l_3+2l_1l_3)(q_0+q_1)}{2j_1j_3} \\ &+ \frac{h_0}{2}T^s(\frac{j_2}{2},t) + \frac{\epsilon_0\sigma}{2}(T^s(\frac{j_2}{2},t))^4 \\ &- \frac{h_0}{2}T^s(-\frac{j_2}{2},t) - \frac{\epsilon_0\sigma}{2}(T^s(-\frac{j_2}{2},t))^4. \end{aligned} \tag{46}$$

Substituting (45) and (46) into (44) yields (10).

2) Proof of $\lambda(\mathbf{A}_1) < 0$ and $\lambda(\mathbf{A}_2) = 0$: We first examine the conduction term $\mathbf{A}_1$, which can be written as $\mathbf{A}_1 = \frac{\lambda_s}{\rho c j_2^2}\mathbf{L}$, where $\mathbf{L} = \begin{bmatrix} -54 & 95 & -42 \\ 30-3h_0j_2/\lambda_s & -60 & 30+3h_0j_2/\lambda_s \\ 54 & -96 & 42 \end{bmatrix}$. Let $\alpha = 3h_0j_2/\lambda_s \geq 0$. The characteristic polynomial of $\mathbf{L}$ is $p(s) = \det(sI-\mathbf{L}) = s^3+72s^2+(750+191\alpha)s+(360+96\alpha)$. Matching the Routh–Hurwitz form $P(s) = a_3s^3+a_2s^2+a_1s+a_0$, we identify $a_3 = 1, a_2 = 72, a_1 = 750+191\alpha, a_0 = 360+96\alpha > 0$, and compute $a_2a_1 - a_3a_0 = 72(750+191\alpha) - (360+96\alpha) = 53640+13656\alpha > 0$. Hence all roots of $p(s)$ have strictly negative real parts and $\lambda(\mathbf{L}) < 0$. Since $\lambda_s/(\rho c j_2^2) > 0$, we obtain $\lambda(\mathbf{A}_1) < 0$. Next, note that $\mathbf{A}_2$ can written as $\mathbf{A}_2 = \begin{bmatrix} 0 & 0 & 0 \\ -\beta & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix}$ where $\beta = \frac{3\epsilon_0\sigma}{\rho c j_2} > 0$. we compute $\mathbf{A}_2^2 = \begin{bmatrix} 0 & 0 & 0 \\ -\beta & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 \\ -\beta & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{0}$, showing that $\mathbf{A}_2$ is nilpotent of degree two. Thus, $\lambda(\mathbf{A}_2) = 0$.

B. Proof of Lemma 1

The Bellman optimality principle states that the optimal solution satisfies the Bellman equation:

$$\begin{aligned} \bar{V}(\hat{\mathbf{T}}(t),h(t)) = \min_{u(k)} \int_{k=t}^{T} \big(&\hat{\mathbf{T}}^T(k)\hat{\mathbf{Q}}\hat{\mathbf{T}}(k) + (R+ \\ &Mh^2(k))u^2(k)\big)dk + \mathbb{E}[V(\hat{\mathbf{T}}(T))|\hat{\mathbf{T}}(t),h(t),u(t)], \end{aligned} \tag{47}$$

where $\bar{V}(\hat{\mathbf{T}}(t),h(t)) \in \mathbb{R}$ is an extended value function satisfying $\mathbb{E}[\bar{V}(\hat{\mathbf{T}}(t),h(t))|\hat{\mathbf{T}}(t)] = V(\hat{\mathbf{T}}(t))$.

Taking the expectation over the $h(t)$ at both sides of (47), it gives that

$$\begin{aligned} V(\hat{\mathbf{T}}(t)) = \mathbb{E}[\min_{u(k)} &\int_{k=t}^{T} (\hat{\mathbf{T}}^T(k)\hat{\mathbf{Q}}\hat{\mathbf{T}}(k) + (R+Mh^2(k)) \\ &u^2(k))dk + \mathbb{E}[V(\hat{\mathbf{T}}(T))|\hat{\mathbf{T}}(t),h(t),u(t)]]. \end{aligned} \tag{48}$$

Note that

$$\begin{aligned} &\lim_{T-t\to 0}\int_{k=t}^{T}\big(\hat{\mathbf{T}}^T(k)\hat{\mathbf{Q}}\hat{\mathbf{T}}(k) + (R+Mh^2(k))u^2(k)\big)dk \\ &= \lim_{T-t\to 0}\big(\hat{\mathbf{T}}^T(t)\hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + (R+Mh^2(t))u^2(t)\big)(T-t). \end{aligned} \tag{49}$$

Similarly, we have

$$\begin{aligned} &\lim_{T-t\to 0}\mathbb{E}[V(\hat{\mathbf{T}}(T))|\hat{\mathbf{T}}(t),h(t),u(t)] = \lim_{T-t\to 0}V(\hat{\mathbf{T}}(t)) + \\ &\mathbb{E}\Big[\Big\langle\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)},\hat{\mathbf{T}}(T)-\hat{\mathbf{T}}(t)\Big\rangle + (T-t)\Big|\hat{\mathbf{T}}(t),h(t),u(t)\Big]. \end{aligned} \tag{50}$$

Thus, we obtain

$$\begin{aligned} V(\hat{\mathbf{T}}(t)) = V(\hat{\mathbf{T}}(t)) + \lim_{T-t\to 0}\mathbb{E}\Big[\min_{u(t)}&\big(\hat{\mathbf{T}}^T(t)\hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + (R+Mh^2(t)) \\ \times u^2(t)\big)(T-t) + \Big\langle\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)},\hat{\mathbf{T}}(T)-\hat{\mathbf{T}}(t)\Big\rangle& \\ + \Big\langle\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial t},T-t\Big\rangle\Big|\hat{\mathbf{T}}(t),h(t),u(t)\Big]&. \end{aligned} \tag{51}$$

Dividing both sides of (51) by $T-t$ and rearranging terms, we derive:

$$\begin{aligned} 0 = \mathbb{E}\Big[\min_{u(t)}&\Big(\hat{\mathbf{T}}^T(t)\hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + (R+Mh^2(t))u^2(t) \\ &+ \Big\langle\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)},\dot{\hat{\mathbf{T}}}(t)\Big\rangle + \frac{\partial V(\hat{\mathbf{T}}(t))}{\partial t}\Big)\Big|\hat{\mathbf{T}}(t),h(t),u(t)\Big], \end{aligned} \tag{52}$$

which corresponds to the HJB equation in (19).

Furthermore, the optimal control solution is given by the minimizer of the R.H.S. in (52): $u^*(t) = -\bar{R}(t)\hat{\mathbf{B}}^T(t)\lambda(\hat{\mathbf{T}}(t))$, where $\lambda(\hat{\mathbf{T}}(t)) = \frac{\partial V(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}$ is the co-function. This completes the proof of Lemma 1.

C. Proof of Theorem 1

Let: $l(\hat{\mathbf{T}}(t),u^*(t)) = \hat{\mathbf{T}}^T(t)\hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + (R+Mh^2(t))(u^*(t))^2$, and $f(\hat{\mathbf{T}}(t),u^*(t),t) = \hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) + \hat{\mathbf{B}}(t)u^*(t) + \hat{\mathbf{c}}$. The HJB equation in (19) can be rewritten as:

$$0 = \mathbb{E}_{h(t)}\left[\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial t} + l(\hat{\mathbf{T}}(t),u^*(t)) + f^T(\hat{\mathbf{T}}(t),u^*(t),t)\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\right]. \tag{53}$$

Taking the total derivative and applying Pontryagin's minimum principle, we obtain:

$$\begin{aligned} &\mathbb{E}_{h(t)}\left[\frac{\partial^2V(\hat{\mathbf{T}}(t))}{\partial t\partial\hat{\mathbf{T}}(t)} + \frac{\partial l(\hat{\mathbf{T}}(t),u^*(t))}{\partial\hat{\mathbf{T}}(t)} + \left(\frac{\partial u^*(t)}{\partial\hat{\mathbf{T}}(t)}\right)^T\frac{\partial l(\hat{\mathbf{T}}(t),u^*(t))}{\partial u^*(t)}\right. \\ &+ \left(\frac{\partial f^T(\hat{\mathbf{T}}(t),u^*(t),t)}{\partial\hat{\mathbf{T}}(t)} + \frac{\partial u^*(t)}{\partial\hat{\mathbf{T}}(t)}\frac{\partial f^T(\hat{\mathbf{T}}(t),u^*(t),t)}{\partial u^*(t)}\right) \\ &\left.\times\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)} + \frac{\partial^2V(\hat{\mathbf{T}}(t))}{\partial^2\hat{\mathbf{T}}(t)}f(\hat{\mathbf{T}}(t),u^*(t),t)\right] \\ &= \mathbb{E}_{h(t)}\left[\frac{\partial\lambda(\hat{\mathbf{T}}(t))}{\partial t} + \frac{\partial l(\hat{\mathbf{T}}(t),u^*(t))}{\partial\hat{\mathbf{T}}(t)}\right. \\ &\left.+ \frac{\partial f^T(\hat{\mathbf{T}}(t),u^*(t),t)}{\partial\hat{\mathbf{T}}(t)}\frac{\partial V(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\right] = 0. \end{aligned} \tag{54}$$

By substituting the explicit forms of $l(\cdot)$ and $f(\cdot)$ into (54), we derive the structured optimality condition w.r.t. $\lambda(\hat{\mathbf{T}}(t))$ as given in (20), which completes the proof of Theorem 1.

D. Proof of Theorem 2

The optimality condition (20) can be rewritten as:

$$\begin{aligned}F(\lambda(\hat{\mathbf{T}}(t))) &\coloneqq \mathbb{E}_{h(t),v(t)}[\dot{\lambda}(\hat{\mathbf{T}}(t))] + \hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T\lambda(\hat{\mathbf{T}}(t)) \\ &+ 4\hat{\mathbf{A}}_2^T \odot \hat{\mathbf{T}}^3(t)\mathbf{I}_{1\times 6}\lambda(\hat{\mathbf{T}}(t)) \\ &= \frac{\partial\lambda(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\Big(\hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) - \mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)]\lambda(\hat{\mathbf{T}}(t)) \\ &+ \hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) + \hat{\mathbf{c}}\Big) + \hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T\lambda(\hat{\mathbf{T}}(t)) \\ &+ (4\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t)\mathbf{I}_{1\times 6}))\lambda(\hat{\mathbf{T}}(t)) = \mathbf{0}_{6\times 1}. \end{aligned} \tag{55}$$

We decompose $F(\lambda(\hat{\mathbf{T}}(t)))$ into its linear and nonlinear components: $F(\lambda(\hat{\mathbf{T}}(t))) = L(\lambda(\hat{\mathbf{T}}(t))) + N(\lambda(\hat{\mathbf{T}}(t)))$, where:

$$\begin{aligned}L(\lambda(\hat{\mathbf{T}}(t))) &= \frac{\partial\lambda(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\big(\hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) - \mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)] \\ &\times \lambda(\hat{\mathbf{T}}(t))\big) + \hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T\lambda(\hat{\mathbf{T}}(t)), \end{aligned} \tag{56}$$

$$\begin{aligned}N(\lambda(\hat{\mathbf{T}}(t))) &= \frac{\partial\lambda(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) \\ &+ (4\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t)\mathbf{I}_{1\times 6}))\lambda(\hat{\mathbf{T}}(t)) + \hat{\mathbf{c}}. \end{aligned} \tag{57}$$

1) Homotopy Construction: We define a homotopy $\tilde{\lambda}(\hat{\mathbf{T}}(t), p)$ satisfying:

$$(1-p)L(\tilde{\lambda}(\hat{\mathbf{T}}(t),p)) + pF(\tilde{\lambda}(\hat{\mathbf{T}}(t),p)) = 0, \tag{58}$$

with boundary condition $\tilde{\lambda}(\mathbf{0}_{6\times 1}, p) = \mathbf{0}_{6\times 1}$.

Expanding $\tilde{\lambda}(\hat{\mathbf{T}}(t), p)$ as a Maclaurin series: $\tilde{\lambda}(\hat{\mathbf{T}}(t),p) = \sum_{n=0}^{\infty} p^n\lambda_n(\hat{\mathbf{T}}(t))$, where $\lambda_n(\hat{\mathbf{T}}(t)) = \frac{1}{n!}\frac{\partial^n\tilde{\lambda}(\hat{\mathbf{T}}(t),p)}{\partial^n p}\Big|_{p=0}$. Rearranging terms in the homotopy equation by powers of p and equating coefficients, we obtain:

2) Zeroth-Order (p^0):

$$\begin{aligned}&\frac{\partial\lambda_0(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\Big(\hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) - \mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)]\lambda_0(t)\Big) \\ &+ \hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T\lambda_0(\hat{\mathbf{T}}(t)) = \mathbf{0}_{6\times 1}, \end{aligned} \tag{59}$$

with initial condition $\lambda_0(\mathbf{0}_{6\times 1}) = \mathbf{0}_{6\times 1}$.

3) First-Order (p^1):

$$\begin{aligned}&\frac{\partial\lambda_1(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\Big(\hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) - \mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)]\lambda_1(\hat{\mathbf{T}}(t))\Big) \\ &+ \hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T\lambda_1(\hat{\mathbf{T}}(t)) + \frac{\partial\lambda_0(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) \\ &+ (4\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t)\mathbf{I}_{1\times 6}))\lambda_0(\hat{\mathbf{T}}(t)) + \hat{\mathbf{c}} = \mathbf{0}_{6\times 1}, \end{aligned} \tag{60}$$

with initial condition $\lambda_1(\mathbf{0}_{6\times 1}) = \mathbf{0}_{6\times 1}$.

4) Higher-Order Terms (p^n, $n > 1$):

$$\begin{aligned}&\frac{\partial\lambda_n(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)}\Big(\hat{\mathbf{A}}_1\hat{\mathbf{T}}(t) - \mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)]\lambda_n(t)\Big) \\ &+ \hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T\lambda_n(t) + \frac{\partial\lambda_{n-1}(t)}{\partial\hat{\mathbf{T}}(t)}\hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) \\ &+ (4\hat{\mathbf{A}}_2^T \odot (\hat{\mathbf{T}}^3(t)\mathbf{I}_{1\times 6}))\lambda_{n-1}(\hat{\mathbf{T}}(t)) = \mathbf{0}_{6\times 1}. \end{aligned} \tag{61}$$

When $p = 1$, (59)-(61) recover (20), leading to (21) in Theorem 2 and completing the proof.

E. Proof of Theorem 3

We analyze the structure of $\lambda_n(\hat{\mathbf{T}}(t))$ and derive the structured form of $\lambda(\hat{\mathbf{T}}(t))$ and $u^*(t)$. We begin with $\lambda_0(\hat{\mathbf{T}}(t))$, which satisfies the optimality condition (60). Using Taylor's theorem, $\lambda_0(\hat{\mathbf{T}}(t))$ can be expressed as:

$$\begin{aligned}\lambda_0(\hat{\mathbf{T}}(t)) &= \mathbf{C}_{0,0} + \sum_{m=1}^{\infty}\sum_{\{a,b,c,d,e,f\}\in\mathcal{C}_m}\mathbf{C}_{0,m}(a,b,c,d,e,f) \\ &\times [[\hat{\mathbf{T}}(t)]_1^a, [\hat{\mathbf{T}}(t)]_2^b, ..., [\hat{\mathbf{T}}(t)]_6^f]^T, \end{aligned} \tag{62}$$

where $\mathbf{C}_{0,0} \in \mathbb{R}^{6\times 1}$ and $\mathcal{C}_m = \{a,b,c,d,e,f \mid a+b+c+d+e+f = m\}$. Substituting (62) into (59), we obtain:

$$\lambda_0(\hat{\mathbf{T}}(t)) = \mathbf{C}_{0,1}\hat{\mathbf{T}}(t), \quad \mathbf{C}_{0,1} \in \mathbb{R}^{6\times 6}. \tag{63}$$

Similarly, $\lambda_1(\hat{\mathbf{T}}(t))$ follows the form:

$$\begin{aligned}\lambda_1(\hat{\mathbf{T}}(t)) &= \mathbf{C}_{1,0} + \sum_{m=1}^{\infty}\sum_{\{a,b,c,d,e,f\}\in\mathcal{C}_m}\mathbf{C}_{1,m}(a,b,c,d,e,f) \\ &\times [[\hat{\mathbf{T}}(t)]_1^a, [\hat{\mathbf{T}}(t)]_2^b, ..., [\hat{\mathbf{T}}(t)]_6^f]^T, \end{aligned} \tag{64}$$

where $\mathbf{C}_{1,0} \in \mathbb{R}^{6\times 1}$. Substituting (63) into (60) and simplifying, we obtain:

$$\lambda_1(\hat{\mathbf{T}}(t)) = \mathbf{C}_{1,0} + \mathbf{C}_{1,1}\hat{\mathbf{T}}(t) + \sum_{m=4}^{\infty}\sum_{\{a,...,f\}\in\mathcal{C}_m}\mathbf{C}_{1,m}(\cdot)\hat{\mathbf{T}}(t). \tag{65}$$

By induction, considering the order of high-order terms and their decay as $\mathcal{O}(\frac{1}{n!})$, we obtain:

$$\lambda_n(\hat{\mathbf{T}}(t)) = \mathbf{C}_{n,0} + \mathbf{C}_{n,1}\hat{\mathbf{T}}(t) + \sum_{m=3}^{\infty}\sum_{\{a,...,f\}\in\mathcal{C}_m}\mathbf{C}_{n,m}(\cdot)\hat{\mathbf{T}}(t), \tag{66}$$

where $\mathbf{C}_{n,0} \neq 0$ only for $n = 1$, and $\mathbf{C}_{n,1} \neq 0$ for all n. The limit property ensures $\lim_{n\to\infty}\mathbf{C}_{n,m}(\cdot) = 0$.

Thus, summing over all n:

$$\lambda(\hat{\mathbf{T}}(t)) = \sum_{n=0}^{\infty}\lambda_n(\hat{\mathbf{T}}(t)) = \mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + g(\hat{\mathbf{T}}(t)). \tag{67}$$

Substituting (67) into (20), we obtain:

$$\begin{aligned}&(\mathbf{C}_1 + \frac{\partial g(\hat{\mathbf{T}}(t))}{\partial\hat{\mathbf{T}}(t)})(\mathbf{A}_4\hat{\mathbf{T}}(t) - \mathbb{E}[\bar{R}(t)\hat{\mathbf{B}}(t)\hat{\mathbf{B}}^T(t)](\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) \\ &+ g(\hat{\mathbf{T}}(t))) + \hat{\mathbf{A}}_2\hat{\mathbf{T}}^4(t) + \hat{\mathbf{c}}) + \hat{\mathbf{Q}}\hat{\mathbf{T}}(t) + \hat{\mathbf{A}}_1^T(\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) \\ &+ g(\hat{\mathbf{T}}(t))) + (4\mathbf{A}_5^T \odot (\hat{\mathbf{T}}^3(t)\mathbf{I}_{1\times 6}))(\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t) + g(\hat{\mathbf{T}}(t))) \\ &= \mathbf{0}_{6\times 1}. \end{aligned} \tag{68}$$

This establishes (28)-(30) and completes the proof.

F. Proof of Theorem 4 and Corollary 1

We first analyze the convergence of $\mathcal{A}_i, i \in \{b, s\}$, then characterize their differences in convergence speed.

1) Convergence of Algorithms: Note that $L_i(\theta_{i,k})$ is Lipschitz continuous w.r.t. $\theta_{i,k}$ under Condition 1 of Theorem 4. Applying the NN update rule in (33), we obtain:

$$\begin{aligned}
&L_i(\theta_{i,k+1}) = L_i(\theta_k - \alpha_k \nabla_{\theta_k} e_i(k\tau;\theta_{i,k}))\\
&= L_i(\theta_{i,k}) - \alpha_k \mathbb{E}[(\nabla_{\theta_{i,k}} e_b(k\tau;\theta_{i,k}))^T \nabla_{\theta_k} L_b(\theta_{i,k})] + \alpha_k^2 \mathbb{E}[(\\
&\nabla_{\theta_k} e_i(k\tau;\theta_{i,k}))^T \nabla^2_{\theta_k} L_b(\theta_{i,k})(\nabla_{\theta_{i,k}} e_i(k\tau;\theta_{i,k}))]\\
&\leq L_i(\theta_k) - \alpha_k \mathbb{E}[\nabla_{\theta_{i,k}} (e_i(k\tau;\theta_{i,k}))^T \nabla_{\theta_k} L_b(\theta_{i,k})] +\\
&\mathbb{E}[\alpha_k^2 L \nabla_{\theta_{i,k}} (e_i(k\tau;\theta_{i,k}))]\\
&\leq L_i(\theta_{i,k}) - \alpha_k c \mathbb{E}[\|\nabla_{\theta_k} L_i(\theta_{i,k})\|^2] +\\
&\alpha_k^2 L_i \mathbb{E}[\nabla_{\theta_{i,k}} (e_i(k\tau;\theta_{i,k}))],
\end{aligned} \tag{69}$$

where $L_i > 0$ and $c > 0$ are positive constants.

Note that α_k is the Lipschitz step size, $\mathbb{E}[\|\nabla^2_{\theta_k} L_i(\theta_{i,k})\|] < \infty$, and $L_i(\theta_{i,k})$ is an L_i-smooth function w.r.t. $\theta_{i,k}$. Thus, by Lemma 2.1 of [42] and Theorem 1 of [43], we have $\limsup_{k\to\infty} \mathbb{E}[\|\nabla_{\theta_{i,k}} L_i(\theta_{i,k})\|^2] = 0$, ensuring that the algorithm converges to a stationary point of the loss function $L_i(\theta_i)$ w.p.1.

Furthermore, when $L_i(\theta_i)$ is strongly convex w.r.t. θ_b, the condition $\limsup_{k\to\infty} \mathbb{E}[\|\nabla_{\theta_{i,k}} L_i(\theta_{i,k})\|^2] = 0$ ensures that the algorithm converges to the optimal point $\theta^* = \arg\min_\theta L_i(\theta_i)$. If $L_i(\theta_i^*) = 0$, the control solution obtained via $\mathcal{A}_i$ satisfies: $\lim_{k\to\infty} u(k\tau) = \lim_{k\to\infty} -\bar{R}(k\tau)\hat{\mathbf{B}}^T(k\tau) f(\hat{\mathbf{T}}(k\tau);\theta_{i,k}) = u^*(k\tau)$, implying that the control solution converges to the optimal control solution w.p.1.

2) Convergence Speed Comparisons: Note that (69) indicates that

$$\begin{aligned}
&\inf_{0\leq k\leq K} \mathbb{E}[\|\nabla_{\theta_k} L_i(\theta_{i,k})\|^2]\\
&\leq \frac{L_i(\theta_{i,0}) - L_i(\theta_b^*)}{c\sum_{k=1}^K \alpha_k} + \frac{\sum_{k=1}^K \mathbb{E}[\alpha_k^2 L \nabla_{\theta_{i,k}} (e_i(k\tau;\theta_k))]}{c\sum_{k=1}^K \alpha_k}.
\end{aligned} \tag{70}$$

As a result, the convergence speeds of $\mathcal{A}_i$ is characterized by

$$\mathbb{E}[\|\nabla_{\theta_{i,k}} L_i(\theta_{i,k})\|^2] \sim \mathcal{O}(\frac{\sum_{t=0}^k \mathbb{E}[\alpha_t^2 \nabla_{\theta_{i,t}} (\|e_i(t\tau;\theta_{i,t})\|^2)}{\sum_{t=0}^k \alpha_t}). \tag{71}$$

When the NNs $f_i(\hat{\mathbf{T}}(t);\theta_i), i \in \{b, s\}$ are constructed with $W_i \times L$, the the structure of the black-box NN is complicated at least by the NN approximation for $\mathbf{C}_0 + \mathbf{C}_1\hat{\mathbf{T}}(t)$ compared to $f_s(\hat{\mathbf{T}}(t);\theta_s)$, and can be denoted by

$$f_b(\hat{\mathbf{T}}(t);\theta_b) = \bar{\mathbf{C}}_0 + \bar{\mathbf{C}}_1\hat{\mathbf{T}}(t) + \bar{f}_s(\hat{\mathbf{T}}(t);\theta_s) + \Delta_b(\hat{\mathbf{T}}(t);\tilde{\theta}), \tag{72}$$

where $\left\{\bar{\mathbf{C}}_0 \in \mathbb{R}^{6\times 1}, \bar{\mathbf{C}}_1 \in \mathbb{R}^{6\times 6}, \theta_s \in \mathbb{R}^{l_s\times 1}, \tilde{\theta} \in \mathbb{R}^{l_e\times 1}\right\} \in \theta_b$ is the NN parameter. $\Delta_b(\cdot)$ is an arbitrary biased structure.

This gives (42), and completes the proof.

G. Proof of Lemma 2

According to Theorem 9.5.3 of [44], there is a NN $\bar{f}_s(\hat{\mathbf{T}}(t);\theta_s)$ with a width of $W_s \in [W_s^*, \infty]$ and depth of $L = 2(\lfloor \log_2 6 \rfloor + 2)$ layers such that $\sup_{\hat{\mathbf{T}}(t)\in\Theta} \|f_s(\hat{\mathbf{T}}(t);\theta_s) - \lambda(\hat{\mathbf{T}}(t))\|_2^2 \leq \epsilon$. Similarly, there is a NN $f_b(\hat{\mathbf{T}}(t);\theta_b)$ with a width $W_2 \in [W_2^*, \infty)$ and and a depth of L layers such that $\sup_{\hat{\mathbf{T}}(t)\in\Theta} \|f_b(\hat{\mathbf{T}}(t);\theta_b) - \lambda(\hat{\mathbf{T}}(t))\|_2^2 \leq \epsilon$.

Note that W_s^* is the minimum width for the structured NN that guarantees above ϵ-dependent inequalities. Given the structure of the co-function $\lambda(\hat{\mathbf{T}}(t))$ in corollary 1, it follows that when W_1^* and W_2^* are applied, then following the similar analysis, the structure of f_b can be denoted as (72). This gives that $l_b + 42 \geq l_s$ and $W_b^* = \lceil \frac{W_s^* \times L + 42}{L} \rceil$. This completes the proof for Lemma 2.

H. Proof of Corollary 2

From Lemma 2 and [45], the computational complexity of $\mathcal{A}_b$ is given by $O((W_s L + g)^\gamma)$, and the complexity of $\mathcal{A}_s$ is $O(W_s^2 L^2)$. Here, $\gamma \geq 4$ is a constant reflecting optimization difficulty, and $g > 0$ is a constant. Thus, the ratio of computational complexities, according to [46], can be expressed as $\frac{O((W_s L+g)^\gamma)}{O(W_s^2 L^2)}$. Substituting $L = 2\lfloor \log_2 d \rfloor + 2$, the ratio simplifies to $O\big((W_s(2\lfloor \log_2 d \rfloor + 2))^{\gamma-2}\big)$, which completes the proof of the corollary.

REFERENCES

[1] Z. Ji, C. Chen, and X. Guan, "Observability guaranteed distributed intelligent sensing for industrial cyber-physical system," *IEEE Trans. Signal Process.*, 2024.

[2] J. Du, X. Luo, L. Jin, and F. Gao, "Robust tensor-based algorithm for UAV-assisted IoT communication systems via nested parafac analysis," *IEEE Trans. Signal Process.*, vol. 70, pp. 5117–5132, 2022.

[3] F. Souza, T. Badings, G. Postma, and J. Jansen, "Integrating expert and physics knowledge for modeling heat load in district heating systems," *IEEE Trans. Ind. Informat.*, 2025.

[4] C.-J. Chen, F.-I. Chou, and J.-H. Chou, "Temperature prediction for reheating furnace by gated recurrent unit approach," *IEEE Access*, vol. 10, pp. 33 362–33 369, 2022.

[5] Y. Shen, X.-S. Chen, Y.-C. Hua, H.-L. Li, L. Wei, and B.-Y. Cao, "Bias dependence of non-Fourier heat spreading in GaN HEMTs," *IEEE Trans. Electron Devices*, vol. 70, no. 2, pp. 409–417, 2022.

[6] C. T. Krasopoulos, A. S. Ioannidis, A. F. Kremmydas, I. A. Karafyllakis, and A. G. Kladas, "Convection heat transfer coefficient regression models construction for fast high-speed motor thermal analysis," *IEEE Trans. Magn.*, vol. 58, no. 11, pp. 1–5, 2022.

[7] M. Tang and V. K. Lau, "Online identification and temperature tracking control for furnace system with a single slab and a single heater over the wirelessly-connected IoT controller," *IEEE Internet Things J.*, 2023.

[8] T. Fujita, M. Mae, H. Fujimoto, M. Nakagawa, Y. Yasuda, and A. Yamagiwa, "Dynamic decoupling current control of boost and buck multiple converters," *IEEE Trans. Power Electron.*, 2025.

[9] Y. Tang, H. Su, T. Jin, and R. C. C. Flesch, "Adaptive PID control approach considering simulated annealing algorithm for thermal damage of brain tumor during magnetic hyperthermia," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–8, 2023.

[10] N. Wang, Z. X. Liu, C. Ding, J.-N. Zhang, G.-R. Sui, H.-Z. Jia, and X.-M. Gao, "High efficiency thermoelectric temperature control system with improved proportional integral differential algorithm using energy feedback technique," *IEEE Trans. Ind. Electron.*, vol. 69, no. 5, pp. 5225–5234, 2021.

[11] B. P. K. Sistla, S. Chatterjee, and V. Natarajan, "Stabilization of cascade interconnection of a 1D heat equation in polar coordinates and an ODE using LQR design," in *Proc. IEEE Conf. Decis. Control (CDC)*, 2022, pp. 7364–7369.

[12] X. Jin, Q. Wu, H. Jia, and N. D. Hatziargyriou, "Optimal integration of building heating loads in integrated heating/electricity community energy systems: A bi-level MPC approach," *IEEE Trans. Sustain. Energy*, vol. 12, no. 3, pp. 1741–1754, 2021.

[13] Y. Yu, W. Yang, W. Ding, and J. Zhou, "Reinforcement learning solution for cyber-physical systems security against replay attacks," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 2583–2595, 2023.

[14] X. Jiang, X. Kong, and Z. Ge, "Augmented industrial data-driven modeling under the curse of dimensionality," *IEEE/CAA J. Autom. Sin.*, vol. 10, no. 6, pp. 1445–1461, 2023.

[15] J. Biswas, A. Goyal, B. Selvanathan, S. H. Nistala, and V. Runkana, "Application of reinforcement learning for real-time optimal control of the pellet induration process," *Trans. Indian Inst. Met.*, vol. 75, no. 10, pp. 2539–2546, 2022.

[16] S. Evmorfos, A. P. Petropulu, and H. V. Poor, "Actor-critic methods for irs design in correlated channel environments: A closer look into the neural tangent kernel of the critic," *IEEE Trans. Signal Process.*, vol. 71, pp. 4029–4044, 2023.

[17] G. Wang, P. Cheng, Z. Chen, B. Vucetic, and Y. Li, "Green cell-free massive MIMO: An optimization-embedded deep reinforcement learning approach," *IEEE Trans. Signal Process.*, vol. 72, pp. 2751–2766, 2024.

[18] S. Evmorfos, K. I. Diamantaras, and A. P. Petropulu, "Reinforcement learning for motion policies in mobile relaying networks," *IEEE Trans. Signal Process.*, vol. 70, pp. 850–861, 2022.

[19] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, "Application of deep reinforcement learning to intrusion detection for supervised problems," *Expert Syst. Appl.*, vol. 141, p. 112963, 2020.

[20] G. Pang, A. Van Den Hengel, C. Shen, and L. Cao, "Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2021, pp. 1298–1308.

[21] S.-G. Choi and S.-B. Cho, "Adaptive database intrusion detection using evolutionary reinforcement learning," in *Proc. Int. Workshop Soft Comput. Models Ind. Environ. Appl.* Springer, 2017, pp. 547–556.

[22] Z. Zhao, W. Liu, D. E. Quevedo, Y. Li, and B. Vucetic, "Deep learning for wireless-networked systems: A joint estimation–control–scheduling approach," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 4535–4550, 2023.

[23] J. Yan, W. Cao, X. Yang, C. Chen, and X. Guan, "Communication-efficient and collision-free motion planning of underwater vehicles via integral reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 8306–8320, 2022.

[24] V. Boussange, S. Becker, A. Jentzen, B. Kuckuck, and L. Pellissier, "Deep learning approximations for non-local nonlinear PDEs with Neumann boundary conditions," *Partial Differ. Equ. Appl.*, vol. 4, no. 6, p. 51, 2023.

[25] J.-L. Wang and H.-F. Li, "Memory-dependent derivative versus fractional derivative (ii): Remodelling diffusion process," *Appl. Math. Comput.*, vol. 391, p. 125627, 2021.

[26] G. Zhang, D. Lu, and H. Liu, "IoT-based positive emotional contagion for crowd evacuation," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1057–1070, 2020.

[27] T. Roy and D. K. Maiti, "An optimal and modified homotopy perturbation method for strongly nonlinear differential equations," *Nonlinear Dyn.*, vol. 111, no. 16, pp. 15 215–15 231, 2023.

[28] H. Ren, Y. Wang, M. Liu, and H. Li, "An optimal estimation framework of multi-agent systems with random transport protocol," *IEEE Trans. Signal Process.*, vol. 70, pp. 2548–2559, 2022.

[29] H. Shen, K. Zhang, M. Hong, and T. Chen, "Towards understanding asynchronous advantage actor-critic: Convergence and linear speedup," *IEEE Trans. Signal Process.*, vol. 71, pp. 2579–2594, 2023.

[30] Z. Wu, H. Shen, T. Chen, and Q. Ling, "Byzantine-resilient decentralized policy evaluation with linear function approximation," *IEEE Trans. Signal Process.*, vol. 69, pp. 3839–3853, 2021.

[31] G. Zhou, C. Pan, H. Ren, P. Popovski, and A. L. Swindlehurst, "Channel estimation for RIS-aided multiuser millimeter-wave systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 1478–1492, 2022.

[32] S. Yu, W. Chen, and H. V. Poor, "Real-time monitoring of chaotic systems with known dynamical equations," *IEEE Trans. Signal Process.*, vol. 72, pp. 1251–1268, 2024.

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[34] X. Zheng and V. K. N. Lau, "Online deep neural networks for mmwave massive MIMO channel estimation with arbitrary array geometry," *IEEE Trans. Signal Process.*, vol. 69, pp. 2010–2025, 2021.

[35] R. Rafailov, K. B. Hatch, V. Kolev, J. D. Martin, M. Phielipp, and C. Finn, "Moto: Offline pre-training to online fine-tuning for model-based robot learning," in *Proc. Conf. Robot Learning (CoRL)*, 2023.

[36] M. C. Paul, "On the effects of high-order scattering in 3d cubical and rectangular furnaces," *Heat Mass Transf.*, vol. 44, no. 11, pp. 1337–1344, 2008.

[37] S. H. Han, S. W. Baek, S. H. Kang, and C. Y. Kim, "Numerical analysis of heating characteristics of a slab in a bench scale reheating furnace," *Int. J. Heat Mass Transf.*, vol. 50, no. 9-10, pp. 2019–2023, 2007.

[38] N. A. Holmberg, "Distribution of metal onto the belt of a horizontal strip caster," *Steel Res.*, vol. 69, no. 1, pp. 22–27, 1998.

[39] S. E. G. Jayamaha, N. E. Wijeysundera, and S. K. Chou, "Measurement of the heat transfer coefficient for walls," *Build. Environ.*, vol. 31, no. 5, pp. 399–407, 1996.

[40] T. L. Bergman, A. S. Lavine, F. P. Incropera, and D. P. DeWitt, *Introduction to Heat Transfer*. John Wiley & Sons, 2011.

[41] E. Boursier, L. Pillaud-Vivien, and N. Flammarion, "Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 20 105–20 118, 2022.

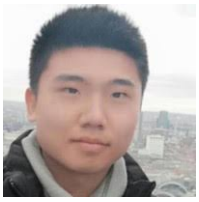
[42] O. Sebbouh, R. M. Gower, and A. Defazio, "Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball," in *Proc. Conf. Learn. Theory (COLT)*. PMLR, 2021, pp. 3935–3971.

[43] J. Liu and Y. Yuan, "On almost sure convergence rates of stochastic gradient methods," in *Proc. Conf. Learn. Theory (COLT)*. PMLR, 2022, pp. 2963–2983.

[44] O. Calin, *Deep Learning Architectures*. Springer, 2020.

[45] V. Froese and C. Hertrich, "Training neural networks is NP-hard in fixed dimension," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 44 039–44 049, 2023.

[46] P. Freire, S. Srivallapanondh, B. Spinnler, A. Napoli, N. Costa, J. E. Prilepsky, and S. K. Turitsyn, "Computational complexity optimization of neural network-based equalizers in digital signal processing: a comprehensive approach," *J. Light. Technol.*, 2024.



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 Post-Doctoral Research Fellow with the Communication Systems Department, EURECOM, France. His current research interests include semantic communications for control, reinforcement learning, networked control systems, and the industrial IoT.

Songfu Cai (Member, IEEE) received the Ph.D. degree in electronic and computer engineering (ECE) from The Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2019. From 2019 to 2024, he was a Postdoctoral Research Fellow with the Department of ECE, HKUST. Since January 2025, he has been a ZJU100 Young Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, supported by the NSFC's Excellent Young Scientists Fund (Overseas). His research interests include stochastic optimal control, wireless networked control, cyber-physical and multi-agent systems, stochastic approximation and optimization, and dynamic games.

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 is currently a Chair Professor with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology. His research interests include massive MIMO for 6G systems, content-centric wireless networking, mission-critical control, integrated AI and PHY, and Bayesian signal processing for machine learning. He is a fellow of Hong Kong Academy of Engineering Sciences and IET, a Changjiang Chair Professor, and the Croucher Senior Research Fellow. He served as an Area Editor for IEEE Transactions on Wireless Communications and IEEE Signal Processing Letters.