Reinforcement Learning Based Optimal Tracking Control Under Unmeasurable Disturbances With Application to HVAC Systems

Syed Ali Asad Rizvi, Amanda J. Pertzborn, and Zongli Lin, *Fellow, IEEE*

***Abstract*— This paper presents the design of an optimal controller for solving tracking problems subject to unmeasurable disturbances and unknown system dynamics using reinforcement learning (RL). Many existing RL control methods take disturbance into account by directly measuring it and manipulating it for exploration during the learning process, thereby preventing any disturbance induced bias in the control estimates. However, in most practical scenarios, disturbance is neither measurable nor manipulable. The main contribution of this article is the introduction of a combination of a bias compensation mechanism and the integral action in the Q-learning framework to remove the need to measure or manipulate the disturbance, while preventing disturbance induced bias in the optimal control estimates. A bias compensated Q-learning scheme is presented that learns the disturbance induced bias terms separately from the optimal control parameters and ensures the convergence of the control parameters to the optimal solution even in the presence of unmeasurable disturbances. Both state feedback and output feedback algorithms are developed based on policy iteration (PI) and value iteration (VI) that guarantee the convergence of the tracking error to zero. The feasibility of the design is validated on a practical optimal control application of a heating, ventilating, and air conditioning (HVAC) zone controller.**

***Index Terms*— Heating, ventilating, and air conditioning (HVAC) control, optimal tracking, Q-learning, reinforcement learning (RL).**

NOMENCLATURE

Symbol	Description and Values
$A_{w_{ew}}$	Area of East/West walls = 9 m^2.
$A_{w_{ns}}$	Area of North/South walls = 12 m^2.
C_{pa}	Specific heat of air = 1.005 kJ/kg-C.
$C_{w_{ew}}$	Thermal capacitance of East/West walls = 70 kJ/C.
$C_{w_{ns}}$	Thermal capacitance of North/South walls = 60 kJ/C.
C_z	Thermal capacitance of the zone = 60 kJ/C.
f_{sa}	Volume flow rate of the supply air = 0.192 m^3.
q	Heat gain from occupants, lights, doors in Watts (W).
T_o	Outside temperature in degrees Celsius (C).
T_{sa}	Supply air temperature in degrees Celsius (C).
$T_{w_{ew}}$	Temperature of the East/West walls in degrees Celsius (C).
$T_{w_{ns}}$	Temperature of the North/West walls in degrees Celsius (C).
T_z	Temperature of the zone in degrees Celsius (C).
$U_{w_{ew}}$	Heat transfer coefficient of East/West walls = 2 W/m^2-C.
$U_{w_{ns}}$	Heat transfer coefficient of North/West walls = 2 W/m^2-C.
K_o	Thermal transfer coefficient between the zone and the outside = 9 W/C.
ρ_a	Density of air = 1.25 kg/m^3.

Manuscript received October 14, 2020; revised March 15, 2021; accepted May 19, 2021. *(Corresponding author: Syed Ali Asad Rizvi.)*

Syed Ali Asad Rizvi and Amanda J. Pertzborn are with the Mechanical Systems and Controls Group, Energy and Environment Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899 USA (e-mail: syedaliasad.rizvi@nist.gov; amanda.pertzborn@nist.gov).

Zongli Lin is with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: zl5y@virginia.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3085358.

Digital Object Identifier 10.1109/TNNLS.2021.3085358

I. INTRODUCTION

REINFORCEMENT learning (RL) is a class of artificial intelligence algorithms, which has gained significant attention in the control community for its potential use in designing intelligent controllers that learn the optimal actions without needing prior knowledge of the system model. This model-free design is desirable because system models are generally hard to obtain and modeling uncertainties can significantly affect the closed-loop stability and the optimality of the controller. One of the important control applications of RL is in solving the optimal tracking problem, which involves designing a controller that forces the system to follow a prescribed reference signal. Tracking control finds application in diverse areas such as robotics, autonomous vehicles, aerospace, building controls, and multiagent systems [1]–[5]. The presence of external disturbances, however, makes the tracking problem more challenging.

One of the pioneering developments in RL-based optimal tracking control involves the idea of state augmentation [6]–[9]. Disturbance rejection capabilities have recently been incorporated in RL by adapting ideas from game theory [10]. The presence of parametric and nonparametric uncertainties [11] and extensions to nonlinear systems [12] have also been considered. Disturbance rejection controllers based on the H_∞ design have been presented to solve the optimal tracking problem [13]–[15]. Extensions of these learning approaches employing policy iteration methods have also

been presented recently [16]. In all the works discussed so far, the disturbance is treated as a decision maker with the disturbance signal being a measurable signal whose L_2 norm is bounded. However, in a practical setting a disturbance is not an intelligent decision maker, and it cannot be measured or influenced by the controller. Consequently, ignoring the disturbance in the learning equation leads to control estimates that may become biased because the learning equation does not hold true as a result of the missing disturbance terms. An analysis of the bias terms arising in the closed-form value function as a result of non-disturbance sources has been carried out in [17], [18]. Different from the state augmentation approach, output regulation based on the internal model principle [19] has also recently been considered in the learning control literature [20], [21]. In these approaches the reference and the disturbance are assumed to be generated by an internal model.

To address the above difficulties, instead of directly measuring the disturbance, we introduce bias terms in the Q-function. The Q-learning algorithm is then designed to learn this modified Q-function, which includes the estimates of the bias incurred by the disturbance. Explicitly including the estimates of the bias terms prevents the crucial control parameters from being affected. To achieve disturbance rejection while tracking, we augment the dynamics with the integral of the tracking error. The Q-learning scheme learns the control parameters for this augmented system while also countering the disturbance induced bias to prevent the estimates from drifting away during the learning phase. As will be shown later in the article, although the integral action helps in rejecting the disturbance to ensure asymptotic tracking, it alone is insufficient to prevent the biasing effect of the disturbance. To relax the exploration condition, we employ off-policy learning in a way similar to [22], in which the behavioral policy that is used to generate system data does not follow the intermediate policies being learned.

The contributions of this present work are summarized as follows. In recent work [13], where the disturbance was assumed to be measurable during the data collection and learning phase, an off-policy RL technique was proposed to solve the optimal tracking problem subject to an L_2 disturbance. This approach also required a discounted cost function. As acknowledged in [13], discounted cost functions may not guarantee closed-loop stability and a system dependent bound on the discounting factor needs to be satisfied (see [23] for discrete-time problems). This work solves the tracking problem in the presence of external disturbances using an integral augmentation approach that does not require a discounted cost function. More importantly, compared to the robust off-policy techniques [13], [15], [24], [25], this work does not require the measurement of the disturbance, which does not need to be an L_2 signal. In another recent work [26], the internal model principle is employed to generate a disturbance, therefore, the disturbance is implicitly measurable. A separate identification process is needed in that approach to solve a set of regulator equations. This work attempts to address the above mentioned difficulties in solving the optimal tracking problem in the presence of unmeasurable disturbances.

The proposed scheme is demonstrated through zone control in a heating, ventilating, and air conditioning (HVAC) application. In this case study, the zone is a room in a commercial office building and the goal is to maintain the zone temperature at the desired set point. The zone temperature is affected by the weather, the airflow rate, the supply air temperature, the thermal mass of the building materials, and the internally generated thermal loads (from equipment, people, etc.). In this scenario the air-handling unit (AHU), which produces the supply air at a given temperature and airflow rate, is the actuator that manipulates the zone temperature. The optimal operation of the system is based on the balance between maintaining the zone at a specified temperature and the cost of the energy required to meet that need.

The remainder of this article is organized as follows: Section II provides a description of the problem. Section III presents the main theoretical development of this paper, where we introduce a bias compensation mechanism and integral action to create a modified Q-function. Then, the design of a Q-learning scheme is presented that learns this Q-function to solve the optimal tracking problem. In particular, we present four Q-learning algorithms based on policy iteration (PI) and value iteration (VI) using state feedback and output feedback. Section IV includes the application of the proposed scheme to the design of an HVAC zone controller. Some concluding remarks are made in Section V.

II. Problem Description

Consider a discrete-time linear time-invariant system in the state space form

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Dd_k, \\ y_k &= Cx_k, \end{aligned} \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the system state, $u_k \in \mathbb{R}^m$ is the control input, $d_k \in \mathbb{R}^q$ is the external disturbance, and $y_k \in \mathbb{R}^p$ is the system output. We define the tracking error as

$$e_k = y_k - r_k,$$

where $r_k \in \mathbb{R}^p$ is the reference trajectory. We assume that $m \geq p$. The control problem is to find the optimal control sequence u_k^* with feedback gain K^* that guarantees asymptotic output tracking, i.e., $\lim_{k\to\infty} e_k = 0$, while minimizing a quadratic cost function of the form

$$J = \sum_{i=0}^{\infty} \left(e_i^{\mathrm{T}} Q_e e_i + \tilde{u}_i^{\mathrm{T}} R \tilde{u}_i \right), \tag{2}$$

where $\left(A, (Q_e)^{1/2} C\right)$ is observable, and $Q_e \geq 0$ and $R > 0$ are the cost weighting matrices that penalize the performance and control energy in terms of the tracking error and the relative control $\tilde{u}_k = u_k - u_{\mathrm{ss}}$, respectively, with the subscript $_{\mathrm{ss}}$ indicating the steady-state values.

III. Design Methodology

A. State Augmentation With Integral Action

In this section, we introduce the integral action to compensate for external disturbances and to guarantee tracking error convergence. To this end, we introduce a new state ω_k

that accumulates the tracking error, which is the discrete-time equivalent of the integral action in the continuous-time setting. Based on this new state, we form the following augmented system:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Dd_k, \\ w_{k+1} &= w_k + e_k, \end{aligned}$$

which can be represented compactly in terms of the augmented state vector $X_k = \begin{bmatrix} x_k^{\mathrm{T}} & w_k^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ as

$$\begin{aligned} X_{k+1} &= \begin{bmatrix} A & 0 \\ C & I_p \end{bmatrix} X_k + \begin{bmatrix} B \\ 0 \end{bmatrix} u_k + \begin{bmatrix} D \\ 0 \end{bmatrix} d_k - \begin{bmatrix} 0 \\ I_p \end{bmatrix} r_k \\ &\triangleq \bar{A} X_k + \bar{B} u_k + \bar{D} d_k + \bar{R} r_k, \\ Y_k &= \begin{bmatrix} C & 0 \\ 0 & I_p \end{bmatrix} X_k \triangleq \bar{C} X_k. \end{aligned} \tag{3}$$

Assuming that the tracking problem is solvable with a steady-state state X_{ss}, and a steady-state control u_{ss} that balances the effect of the disturbance and the reference by means of the integral action, we can obtain the following error dynamics:

$$\begin{aligned} \tilde{X}_{k+1} &= \bar{A} \tilde{X}_k + \bar{B} \tilde{u}_k, \\ \tilde{Y}_k &= \bar{C} \tilde{X}_k, \end{aligned} \tag{4}$$

where $\tilde{X}_k = X_k - X_{\mathrm{ss}}$. For this augmented system, we define the augmented cost function in terms of (2) as

$$\begin{aligned} J &= \sum_{i=0}^{\infty} \left(\begin{bmatrix} \tilde{x}_i \\ \tilde{w}_i \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} C^{\mathrm{T}} Q_e C & 0 \\ 0 & Q_w \end{bmatrix} \begin{bmatrix} \tilde{x}_i \\ \tilde{w}_i \end{bmatrix} + \tilde{u}_i^{\mathrm{T}} R \tilde{u}_i \right) \\ &\triangleq \sum_{i=0}^{\infty} \left(\tilde{X}_i Q \tilde{X}_i + \tilde{u}_i^{\mathrm{T}} R \tilde{u}_i \right). \end{aligned} \tag{5}$$

To design an optimal controller we first establish the controllability conditions for the augmented dynamics.

Lemma 1: The augmented system (3) is controllable if the original system (1) is controllable and has no invariant zeros at $z = 1$, where z is the z-transform variable.

Proof: By the Popov-Belevitch-Hautus (PBH) test, the augmented system is controllable if and only if $\begin{bmatrix} \bar{A} - \lambda I_{n+p} & \bar{B} \end{bmatrix}$ has full row rank of $n + p$. In view of the definition of $\bar{A}$ and $\bar{B}$ in (3), the rank of $\begin{bmatrix} \bar{A} - \lambda I_{n+p} & \bar{B} \end{bmatrix}$ is evaluated as

$$\rho \begin{bmatrix} \bar{A} - \lambda I_{n+p} & \bar{B} \end{bmatrix} = \rho \begin{bmatrix} A - \lambda I_n & 0 & B \\ C & (1-\lambda) I_p & 0 \end{bmatrix}.$$

For $\lambda \neq 1$, we can cancel out the entries of C using the columns of I_p to result in

$$\begin{aligned} \rho \begin{bmatrix} \bar{A} - \lambda I_{n+p} & \bar{B} \end{bmatrix} &= \rho \begin{bmatrix} A - \lambda I_n & B & 0 \\ 0 & 0 & (1-\lambda) I_p \end{bmatrix} \\ &= \rho \begin{bmatrix} A - \lambda I_n & B \end{bmatrix} + p. \end{aligned}$$

Because the original system (1) is controllable, we have $\rho \begin{bmatrix} A - \lambda I_n & B \end{bmatrix} = n$, and therefore,

$$\rho \begin{bmatrix} \bar{A} - \lambda I_{n+p} & \bar{B} \end{bmatrix} = n + p.$$

For $\lambda = 1$, we have

$$\rho \begin{bmatrix} \bar{A} - \lambda I_{n+p} & \bar{B} \end{bmatrix} = \rho \begin{bmatrix} A - \lambda I_n & B \\ C & 0 \end{bmatrix}.$$

Recall that the system (A, B, C) has a zero at $z = 1$ if and only if

$$\rho \begin{bmatrix} A - \lambda I_n & B \\ C & 0 \end{bmatrix} < n + \min\{p, m\} = n + p$$

and, as a result, $\rho \begin{bmatrix} \bar{A} - \lambda I_{n+p} & \bar{B} \end{bmatrix} = n + p$ if the system (A, B, C) has no invariant zeros at $z = 1$. This completes the proof. ∎

Under the conditions of controllability of $(\bar{A}, \bar{B})$ and observability of $\left(\bar{A}, (Q)^{1/2}\right)$, where $(Q^{1/2})^{\mathrm{T}}(Q)^{1/2} = Q$, there exists a unique optimal control given by

$$\tilde{u}_k^* = -(R + \bar{B}^{\mathrm{T}} P^* \bar{B})^{-1} \bar{B}^{\mathrm{T}} P^* \bar{A} \tilde{X}_k = -K^* \tilde{X}_k, \tag{6}$$

where P^* is the unique positive definite solution to the algebraic Riccati equation (ARE) [27]

$$\bar{A}^{\mathrm{T}} \bar{P} \bar{A} - P + Q - \bar{A}^{\mathrm{T}} P \bar{B} (R + \bar{B}^{\mathrm{T}} P \bar{B})^{-1} \bar{B}^{\mathrm{T}} P \bar{A} = 0. \tag{7}$$

Moreover, (6) and (7) suggest that the optimal feedback gain K^* for the error dynamics (4) is identical to that of the augmented dynamics (3). As such, K^* can be obtained independent of the disturbance, reference, and steady-state offsets, which are handled by the integral action, as will be seen.

The design procedure discussed above is an offline approach that assumes the availability of a perfectly known model of the system. That is, the system dynamics matrices are available so that K^* can be obtained by solving the ARE (7). In this work, we are interested in learning K^* by employing the framework of RL. In particular, we present the design of a completely model-free Q-learning method that enables us to learn K^* online. The existing RL control literature identifies a difficulty in applying RL control to the system dynamics (3) that stems from the presence of the extra term corresponding to the external disturbances [28], which are generally not available for measurement in an online setting. The disturbance, if not accounted for, results in bias in the Q-learning estimates, causing them to be suboptimal and, more importantly, may render the closed-loop system unstable. Therefore, in the following, we present a Q-learning scheme that accounts for the biasing effect of the disturbances.

B. Bias Compensated Q-function

In this section, we first seek to derive a Q-function for the augmented system dynamics while accounting for the bias effect of the disturbances. For the design of an online algorithm, we consider the dynamics (3) rather than the error dynamics (4), which is for analysis only and involves the steady-state values that are not available *a priori*. Nevertheless, the resulting optimal feedback control matrix K^* is the same for both (3) and (4), as mentioned in Section III-A. For a stabilizing control $u_k = -K X_k$ with policy K, the total cost incurred when starting from any state X_k is quadratic in the state as given by

$$V_K(X_k) = X_k^{\mathrm{T}} P X_k, \quad P = P^{\mathrm{T}} > 0. \tag{8}$$

The Q-function associated with K is [29]

$$Q_K = X_k^{\mathrm{T}} Q X_k + u_k^{\mathrm{T}} R u_k + V_K(X_{k+1}), \tag{9}$$

which is the sum of the one-step cost of taking an arbitrary action u_k from the state at time k, X_k, plus the total cost of using policy K from time $k+1$ onward. The reference r_k and the disturbance d_k are neither the state nor the decision makers and are, therefore, considered external signals that influence the dynamics. Substituting the dynamics (3) in (9), we have

$$\begin{aligned} Q_K &= \begin{bmatrix} X_k \\ u_k \\ r_k \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} Q+\bar{A}^{\mathrm{T}}P\bar{A} & \bar{A}^{\mathrm{T}}P\bar{B} & \bar{A}^{\mathrm{T}}P\bar{R} \\ \bar{B}^{\mathrm{T}}P\bar{A} & R+\bar{B}^{\mathrm{T}}P\bar{B} & \bar{B}^{\mathrm{T}}P\bar{R} \\ \bar{R}^{\mathrm{T}}P\bar{A} & \bar{R}^{\mathrm{T}}P\bar{B} & \bar{R}^{\mathrm{T}}P\bar{R} \end{bmatrix} \begin{bmatrix} X_k \\ u_k \\ r_k \end{bmatrix} \\ &\quad + 2X_k^{\mathrm{T}}\bar{A}^{\mathrm{T}}\bar{P}\bar{D}d_k + 2u_k^{\mathrm{T}}\bar{B}^{\mathrm{T}}\bar{P}\bar{D}d_k + 2r_k^{\mathrm{T}}\bar{B}^{\mathrm{T}}\bar{P}\bar{D}d_k \\ &\quad + d_k^{\mathrm{T}}\bar{D}^{\mathrm{T}}\bar{P}\bar{D}d_k \\ &\triangleq (z'_k)^{\mathrm{T}}H'z'_k + 2X_k^{\mathrm{T}}\bar{A}^{\mathrm{T}}\bar{P}\bar{D}d_k + 2u_k^{\mathrm{T}}\bar{B}^{\mathrm{T}}\bar{P}\bar{D}d_k \\ &\quad + 2r_k^{\mathrm{T}}\bar{B}^{\mathrm{T}}\bar{P}\bar{D}d_k + d_k^{\mathrm{T}}\bar{D}^{\mathrm{T}}\bar{P}\bar{D}d_k, \end{aligned} \tag{10}$$

where the last four terms involving the unmeasurable signal d_k result in an estimation bias. As d_k is not known, we can lump it together with the unknown system dynamics matrices to write the Q-function more compactly as

$$\begin{aligned} Q_K &= \begin{bmatrix} X_k \\ u_k \\ r_k \\ c \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} Q+\bar{A}^{\mathrm{T}}P\bar{A} & \bar{A}^{\mathrm{T}}\bar{P}\bar{B} & \bar{A}^{\mathrm{T}}P\bar{R} & b_1 \\ \bar{B}^{\mathrm{T}}P\bar{A} & R+\bar{B}^{\mathrm{T}}P\bar{B} & \bar{B}^{\mathrm{T}}P\bar{R} & b_2 \\ \bar{R}^{\mathrm{T}}P\bar{A} & \bar{R}^{\mathrm{T}}P\bar{B} & \bar{R}^{\mathrm{T}}P\bar{R} & b_3 \\ b_1^{\mathrm{T}} & b_2^{\mathrm{T}} & b_3^{\mathrm{T}} & b_4 \end{bmatrix} \begin{bmatrix} X_k \\ u_k \\ r_k \\ c \end{bmatrix} \\ &\triangleq \begin{bmatrix} X_k \\ u_k \\ r_k \\ c \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} H_{XX} & H_{Xu} & H_{Xr} & b_1 \\ H_{uX} & H_{uu} & H_{ur} & b_2 \\ H_{rX} & H_{ru} & H_{rr} & b_3 \\ b_1^{\mathrm{T}} & b_2^{\mathrm{T}} & b_3^{\mathrm{T}} & b_4 \end{bmatrix} \begin{bmatrix} X_k \\ u_k \\ r_k \\ c \end{bmatrix} \\ &\triangleq z_k^{\mathrm{T}}Hz_k, \end{aligned} \tag{11}$$

where c is an arbitrary bias scaling factor. It is worth pointing out that (11) is an extension of the LQR Q-function that incorporates the biasing effect of the disturbance and the b_i's and c are dependent on the disturbance. The optimal Q-function Q_K^* and its corresponding matrix H^* are obtained using the above expression by substituting $P = P^*$. The optimal feedback term can then be obtained as

$$K^* = \left(H_{uu}^*\right)^{-1}\left(H_{uX}^*\right).$$

The result is the same as the feedback gain K^* defined in (6). This suggests that learning the optimal Q-function amounts to learning the optimal feedback controller. In the Section III-C, we will present iterative Q-learning algorithms that provide estimates of this optimal Q-function.

C. Full State Feedback Q-learning Algorithms

In this section, we will present a state feedback Q-learning scheme incorporating the integral action toward solving the optimal tracking problem. Before introducing the bias compensated algorithms, we will present an uncompensated Q-learning algorithm for the augmented system (3). Let $Q'_K = (z'_k)^{\mathrm{T}}H'z'_k$ be the uncompensated Q-function that does not fully take into account the effect of the disturbance. The Q-learning Bellman equation corresponding to this Q-function is obtained as [30]

$$Q'_K(X_k, u_k) = X_k^{\mathrm{T}}QX_k + u_k^{\mathrm{T}}Ru_k + Q'_K(X_{k+1}, -KX_{k+1}),$$

Algorithm 0 State Feedback Q-learning Policy Iteration Algorithm for Tracking Control

input: input-state data
output: H^*

1: **initialize.** Select a stabilizing initial policy $u_k^0 = -K^0X_k + \nu_k$ with ν_k being an exploration signal. Set $j \leftarrow 0$.
2: **acquire data.** Apply input u_k^0 to collect $L \geq l(l+1)/2$ datasets of (X_k, u_k, r_k).
3: **repeat**
4: **policy evaluation.** Determine the least-squares solution of

$$(z'_k)^{\mathrm{T}}H'^j z'_k = X_k^{\mathrm{T}}QX_k + u_k^{\mathrm{T}}Ru_k + (z'_{k+1})^{\mathrm{T}}H'^j z'_{k+1}.$$

5: **policy improvement.** Determine an improved policy as

$$K^{j+1} = (H_{uu}'^j)^{-1}\left(H_{uX}'^j\right).$$

6: $j \leftarrow j+1$.
7: **until** $\left\|K^j - K^{j-1}\right\| < \varepsilon$ for some small $\varepsilon > 0$.

or equivalently

$$\left(z'_k\right)^{\mathrm{T}}H'z'_k = X_k^{\mathrm{T}}QX_k + u_k^{\mathrm{T}}Ru_k + \left(z'_{k+1}\right)^{\mathrm{T}}H'z'_{k+1}. \tag{12}$$

Algorithm 0, the uncompensated Q-learning algorithm, is based on this Q-learning equation and includes an integral feedback term to compensate for the steady-state tracking error resulting from the disturbance. However, as will be shown, the integral action alone will not prevent the disturbance from incurring bias in the Q-learning estimates during learning.

We now proceed to present the bias compensated Q-learning algorithms. For the compensated Q-function in (11), we have the following Q-learning Bellman equation

$$z_k^{\mathrm{T}}Hz_k = X_k^{\mathrm{T}}QX_k + u_k^{\mathrm{T}}Ru_k + z_{k+1}^{\mathrm{T}}Hz_{k+1}, \tag{13}$$

or equivalently

$$\bar{H}^{\mathrm{T}}\bar{z}_k = X_k^{\mathrm{T}}QX_k + u_k^{\mathrm{T}}Ru_k + \bar{H}^{\mathrm{T}}\bar{z}_{k+1}, \tag{14}$$

where

$$\begin{aligned} \bar{H} &= \mathrm{vec}(H) \\ &\triangleq [h_{11}\ 2h_{12}\ \ldots\ 2h_{1l}\ h_{22}\ 2h_{23}\ \ldots \\ &\qquad 2h_{2l}\ \ldots h_{ll}]^{\mathrm{T}} \in \mathbb{R}^{l(l+1)/2}, \quad l = n+m+2p+1, \\ \bar{z}_k &= \left[z_{k1}^2\ z_{k1}z_{k2}\ \ldots\ z_{k1}z_{kl}\ z_{k2}^2\ z_{k2}z_{k3}\ \ldots\ z_{k2}z_{kl}\ \ldots z_{kl}^2\right]^{\mathrm{T}} \end{aligned}$$

with z_{ki} being the components of z_k. Based on (14), both PI and VI algorithms are considered next to learn the Q-function and the optimal feedback controller. Algorithm 1 presents a PI Q-learning algorithm for the linear quadratic tracking problem. This is essentially a two-step procedure. In the policy evaluation step, we use the key equation (14) to solve for the unknown vector $\bar{H}$ in the least-squares sense by collecting $L \geq l(l+1)/2$ data samples of (X_k, u_k, r_k) to form the data

Algorithm 1 Bias Compensated State Feedback Q-learning Policy Iteration Algorithm for Tracking Control

input: input-state data
output: H^*

1: **initialize.** Select a stabilizing initial policy $u_k^0 = -K^0 X_k + \nu_k$ with ν_k being an exploration signal. Set $j \leftarrow 0$.
2: **acquire data.** Apply input u_k^0 to collect $L \geq l(l+1)/2$ datasets of (X_k, u_k, r_k).
3: **repeat**
4: **policy evaluation.** Determine the least-squares solution of

$$\left(\bar{H}^j\right)^{\mathrm{T}}(\bar{z}_k - \bar{z}_{k+1}) = X_k^{\mathrm{T}} Q X_k + u_k^{\mathrm{T}} R u_k.$$

5: **policy improvement.** Determine an improved policy as

$$K^{j+1} = (H_{uu}^j)^{-1}\left(H_{uX}^j\right).$$

6: $j \leftarrow j+1$
7: **until** $\left\|K^j - K^{j-1}\right\| < \varepsilon$ for some small $\varepsilon > 0$.

matrices $\Phi \in \mathbb{R}^{l(l+1)/2 \times L}$ and $\Upsilon \in \mathbb{R}^{L \times 1}$, defined by

$$\Phi = \left[\bar{z}_{k-L+1} - \bar{z}_{k-L+2} \;\; \bar{z}_{k-L+2} - \bar{z}_{k-L+3} \;\; \cdots \;\; \bar{z}_k - \bar{z}_{k+1}\right],$$

$$\Upsilon = \begin{bmatrix} X_{k-L+1}^{\mathrm{T}} Q X_{k-L+1} + u_{k-L+1}^{\mathrm{T}} R u_{k-L+1} \\ X_{k-L+2}^{\mathrm{T}} Q X_{k-L+2} + u_{k-L+2}^{\mathrm{T}} R u_{k-L+2} \\ \vdots \\ X_k^{\mathrm{T}} Q X_k + u_k^{\mathrm{T}} R u_k \end{bmatrix}.$$

Then, the least-squares solution of (14) is given by

$$\bar{H}^j = \left(\Phi \Phi^{\mathrm{T}}\right)^{-1} \Phi \Upsilon, \tag{15}$$

where $\bar{H}^j$ is the jth estimate of the unknown vector $\bar{H}$. Because $u_k = -K X_k$, which is linearly dependent on X_k, (15) will not have a unique solution, which is needed for convergence to the optimal parameters. To overcome this issue, an excitation noise is added in u_k to guarantee a unique solution to (15). Note, however, that the exploration noise is unable to excite the bias term c in the vector z_k, and therefore, there will be a zero entry corresponding to the quadratic term in c in the regression vector $\bar{z}_k - \bar{z}_{k+1}$. Nevertheless, because the bias scaling factor itself is arbitrarily selected, this issue can be tackled by separately adding an arbitrary offset in the corresponding entry of $\bar{z}_{k+1}$. The above measures ensure that the following condition can be satisfied:

$$\rho(\Phi) = l(l+1)/2. \tag{16}$$

Clearly, the rank condition (16) is necessary to obtain the optimal solution, which is a unique solution to the least-squares problem (15). This rank condition is crucial to exploration in off-policy RL control algorithms [22]. Interested readers can refer to [31] for the stochastic version of this condition. Algorithm 1 requires a stabilizing (not necessarily optimal) policy at initialization. However, this requirement could be restrictive in certain applications and as such a policy may be difficult to obtain because the system dynamics are not known in advance or the dynamics are nonlinear [32].

Algorithm 2 Bias Compensated State Feedback Q-learning Value Iteration Algorithm for Tracking Control

input: input-state data
output: H^*

1: **initialize.** Select an arbitrary policy $u_k^0 = -K^0 X_k + \nu_k$ with ν_k being an exploration signal. Set $j \leftarrow 0$ and $H^0 \geq 0$.
2: **acquire data.** Apply input u_k^0 to collect $L \geq l(l+1)/2$ datasets of (X_k, u_k, r_k).
3: **repeat**
4: **value update.** Determine the least-squares solution of

$$\left(\bar{H}^{j+1}\right)^{\mathrm{T}}(\bar{z}_k) = X_k^{\mathrm{T}} Q X_k + u_k^{\mathrm{T}} R u_k + \bar{H}^j \bar{z}_{k+1}.$$

5: **policy improvement.** Determine an improved policy as

$$K^{j+1} = (H_{uu}^{j+1})^{-1}\left(H_{uX}^{j+1}\right)$$

6: $j \leftarrow j+1$.
7: **until** $\left\|K^j - K^{j-1}\right\| < \varepsilon$ for some small $\varepsilon > 0$.

To address this difficulty, we can refer to a slightly different iterative technique, VI, which does not impose this restriction. A bias compensated Q-learning VI algorithm is presented in Algorithm 2.

The data matrices $\Phi \in \mathbb{R}^{l(l+1)/2 \times L}$ and $\Upsilon \in \mathbb{R}^{L \times 1}$ for the case of VI are defined by

$$\Phi = \left[\bar{z}_k^1 \;\; \bar{z}_k^2 \;\; \cdots \;\; \bar{z}_k^L\right],$$

$$\Upsilon = \left[r^1(y_k, u_k) + \bar{H}_{j-1}^{\mathrm{T}} \bar{z}_{k+1}^1 \;\; \cdots \;\; r^L(y_k, u_k) + \bar{H}_{j-1}^{\mathrm{T}} \bar{z}_{k+1}^L\right]^{\mathrm{T}}.$$

Remark 1: Algorithms 1 and 2 are the extension of the standard LQR Q-learning algorithms found in the literature [33]. They learn the disturbance induced bias terms without measuring the disturbance directly, which enables these algorithms to prevent the biasing components from affecting the optimal control parameters.

D. Output Feedback Q-learning Algorithms

Section III-C presented designs that were based on the feedback of the state x_k. However, in practice, only a subset of the state is measurable through the system output. Classical state estimation techniques that estimate x_k through input-output measurements do not apply as the system dynamics are unknown. In our previous work [34], we designed a Q-learning scheme based on a parameterization of the system state in terms of a sequence of delayed measurements of input, output, and disturbance, as follows

$$x_k = M_y \bar{y}_{k-1,k-N} + M_u \bar{u}_{k-1,k-N} + M_d \bar{d}_{k-1,k-N}, \tag{17}$$

where $N \leq n$ is an upper bound on the system's observability index. Interested readers can refer to [34] for the details of this parameterization.

We use state parameterization (17) to describe the Q-function in (9). Based on (17), we obtain the parameterization of the augmented state X_k as follows:

$$X_k = \begin{bmatrix} M_u & M_y & M_d & 0 \\ 0 & 0 & 0 & I_p \end{bmatrix}\left[\bar{u}_k^{\mathrm{T}} \;\; \bar{y}_k^{\mathrm{T}} \;\; \bar{d}_k^{\mathrm{T}} \;\; w_k^{\mathrm{T}}\right]^{\mathrm{T}}$$
$$\triangleq \left[\bar{M}_u \;\; \bar{M}_y \;\; \bar{M}_d \;\; \bar{M}_w\right]\left[\bar{u}_k^{\mathrm{T}} \;\; \bar{y}_k^{\mathrm{T}} \;\; \bar{d}_k^{\mathrm{T}} \;\; w_k^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{18}$$

Next, we derive the expressions for the output feedback Q-function $Q_{\mathcal{K}}$ and the associated output feedback policy $\mathcal{K}$.

Substitution of the augmented state vector with its parameterized form (18) into the state feedback Q-function (11) results in

$$Q_{\mathcal{K}} = \begin{bmatrix} \bar{u}_k \\ \bar{y}_k \\ w_k \\ u_k \\ r_k \\ c \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathcal{H}_{\bar{u}\bar{u}} & \mathcal{H}_{\bar{u}\bar{y}} & \mathcal{H}_{\bar{u}w} & \mathcal{H}_{\bar{u}u} & \mathcal{H}_{\bar{u}r} & b_1 \\ \mathcal{H}_{\bar{y}\bar{u}} & \mathcal{H}_{\bar{y}\bar{y}} & \mathcal{H}_{\bar{y}w} & \mathcal{H}_{\bar{y}u} & \mathcal{H}_{\bar{y}r} & b_2 \\ \mathcal{H}_{w\bar{u}} & \mathcal{H}_{w\bar{y}} & \mathcal{H}_{ww} & \mathcal{H}_{wu} & \mathcal{H}_{wr} & b_3 \\ \mathcal{H}_{u\bar{u}} & \mathcal{H}_{u\bar{y}} & \mathcal{H}_{uw} & \mathcal{H}_{uu} & \mathcal{H}_{ur} & b_4 \\ \mathcal{H}_{r\bar{u}} & \mathcal{H}_{r\bar{y}} & \mathcal{H}_{rw} & \mathcal{H}_{wu} & \mathcal{H}_{rr} & b_5 \\ b_1^{\mathrm{T}} & b_2^{\mathrm{T}} & b_3^{\mathrm{T}} & b_4^{\mathrm{T}} & b_5^{\mathrm{T}} & b_6^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \bar{u}_k \\ \bar{y}_k \\ w_k \\ u_k \\ r_k \\ c \end{bmatrix}$$
$$\triangleq \zeta_k^T \mathcal{H} \zeta_k \tag{19}$$

where $\mathcal{H} = \mathcal{H}^T \in \mathbb{R}^{l\times l}$ with $l = mN + pN + m + 2p + 1$ and the submatrices are defined in an obvious way as in [34]. Note that $b_i's$ and c are again the disturbance dependent terms as seen in the case of the state feedback Q-function (11). Notice that the delayed disturbance dependent term $\bar{d}_k$ introduced as a result of the state parameterization has been lumped together with the biasing term c.

The optimal output feedback policy $\mathcal{K}^*$ can be obtained when the optimal output feedback function $Q_{\mathcal{K}}^*$ is minimized with respect to u_k. This results in

$$\mathcal{K}^* = \left(\mathcal{H}_{uu}^*\right)^{-1}\begin{bmatrix}\mathcal{H}_{u\bar{u}}^* & \mathcal{H}_{u\bar{y}}^* & \mathcal{H}_{uw}^*\end{bmatrix}.$$

Finally, we have the following feedback control law:

$$u_k^* = -\left(\mathcal{H}_{uu}^*\right)^{-1}\left(\mathcal{H}_{u\bar{u}}^*\bar{u}_{k-1,k-N} + \mathcal{H}_{u\bar{y}}^*\bar{y}_{k-1,k-N} + \mathcal{H}_{uw}^* w_k\right)$$
$$\triangleq -\mathcal{K}^*\begin{bmatrix}\bar{u}_{k-1,k-N}^{\mathrm{T}} & \bar{y}_{k-1,k-N}^{\mathrm{T}} & w_k^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}. \tag{20}$$

It will be shown in the proof of Theorem 1 that the integral action w_k is able to compensate for the unmeasurable disturbances $\bar{d}_k$ and d_k in a way similar to the state feedback case.

Having formulated the output feedback Q-function, we need to develop a Q-learning scheme that can learn this function. In view of the output feedback Q-function (19), the output feedback Q-learning Bellman equation follows from (13) as

$$\bar{\mathcal{H}}^{\mathrm{T}}\bar{\zeta}_k = Y_k^{\mathrm{T}} Q_y Y_k + u_k^{\mathrm{T}} R u_k + \bar{\mathcal{H}}^{\mathrm{T}}\bar{\zeta}_{k+1} \tag{21}$$

where

$$\bar{\mathcal{H}} = \mathrm{vec}(\mathcal{H}) \in \mathbb{R}^{l(l+1)/2}, \quad l = mN + pN + m + 2p + 1.$$

The regression vector $\bar{\zeta}_k \in \mathbb{R}^{l(l+1)/2}$ is defined as

$$\bar{\zeta} = \begin{bmatrix}\zeta_{k1}^2 & \zeta_{k1}\zeta_{k2} & \cdots & \zeta_{k1}\zeta_{kl} & \zeta_{k2}^2 & \zeta_{k2}\zeta_{k3} & \cdots & \zeta_{k2}\zeta_l & \cdots \zeta_{kl}^2\end{bmatrix}^{\mathrm{T}}$$

where $\zeta_k = [\zeta_{k1}\ \zeta_{k2}\ \cdots\ \zeta_{kl}]$.

In comparison with (14), the output feedback learning equation (21) involves more parameters, as the internal state information is not readily available, but rather is embedded in the sufficiently long sequence of input–output data $\bar{u}_{k-1,k-N}$ and $\bar{y}_{k-1,k-N}$.

Equation (21) is utilized in the output feedback PI and VI algorithms. The policy iteration Algorithm 3 is the output feedback counterpart of the policy iteration algorithm, Algorithm 1. The two key differences between these two algorithms can be seen in the policy evaluation and the policy update steps, where we observe that the learning and the control update equations do not involve the state information. Algorithm 4 operates in the same way as the state feedback VI Algorithm 2, but without requiring the measurement of the internal state. Furthermore, it also relaxes the condition of a stabilizing initial gain $\mathcal{K}^0$. It is worth noting at this point that both Algorithms 3 and 4 must also satisfy the rank condition (16), with more exploration compared to the state feedback algorithms because the number of unknown parameters is larger.

Algorithm 3 Bias Compensated Output Feedback Q-learning Policy Iteration Algorithm for Tracking Control

input: input-output data
output: $\mathcal{H}^*$

1: **initialize.** Select a stabilizing initial policy $u_k^0 = -\mathcal{K}^0\begin{bmatrix}\bar{u}_{k-1,k-N} & \bar{y}_{k-1,k-N} & w_k\end{bmatrix} + \nu_k$ with ν_k being an exploration signal. Set $j \leftarrow 0$.
2: **acquire data.** Apply input u_k^0 to collect $L \geq l(l+1)/2$ datasets of $(\bar{u}_{k-1,k-N}, \bar{y}_{k-1,k-N}, w_k, u_k, r_k)$.
3: **repeat**
4: **policy evaluation.** Determine the least-squares solution of

$$\left(\bar{\mathcal{H}}^j\right)^{\mathrm{T}}(\bar{\zeta}_k - \bar{\zeta}_{k+1}) = Y_k^{\mathrm{T}} Q_y Y_k + u_k^{\mathrm{T}} R u_k.$$

5: **policy improvement.** Determine an improved policy as

$$\mathcal{K}^{j+1} = \left(\mathcal{H}_{uu}^j\right)^{-1}\begin{bmatrix}\mathcal{H}_{u\bar{u}}^j & \mathcal{H}_{u\bar{y}}^j & \mathcal{H}_{uw}^j\end{bmatrix}.$$

6: $j \leftarrow j + 1$.
7: **until** $\left\|\bar{\mathcal{K}}^j - \bar{\mathcal{K}}^{j-1}\right\| < \varepsilon$ for some small $\varepsilon > 0$.

Algorithm 4 Bias Compensated Output Feedback Q-learning Value Iteration Algorithm for Tracking Control

input: input-output data
output: H^*

1: **initialize.** Select an arbitrary policy $u_k^0 = -\mathcal{K}^0\begin{bmatrix}\bar{u}_{k-1,k-N} & \bar{y}_{k-1,k-N} & w_k\end{bmatrix} + \nu_k$ with ν_k being an exploration signal. Set $j \leftarrow 0$ and $\mathcal{H}^0 \geq 0$.
2: **acquire data.** Apply input u_k^0 to collect $L \geq l(l+1)/2$ datasets of $(\bar{u}_{k-1,k-N}, \bar{y}_{k-1,k-N}, w_k, u_k, r_k)$.
3: **repeat**
4: **value update.** Determine the least-squares solution of

$$\left(\bar{\mathcal{H}}^{j+1}\right)^{\mathrm{T}}(\bar{\zeta}_k) = Y_k^{\mathrm{T}} Q_y Y_k + u_k^{\mathrm{T}} R u_k + \bar{\mathcal{H}}^j\bar{\zeta}_{k+1}.$$

5: **policy improvement.** Determine an improved policy as

$$\mathcal{K}^{j+1} = \left(\mathcal{H}_{uu}^{j+1}\right)^{-1}\begin{bmatrix}\mathcal{H}_{u\bar{u}}^{j+1} & \mathcal{H}_{u\bar{y}}^{j+1} & \mathcal{H}_{uw}^{j+1}\end{bmatrix}.$$

6: $j \leftarrow j + 1$.
7: **until** $\left\|\bar{\mathcal{K}}^j - \bar{\mathcal{K}}^{j-1}\right\| < \varepsilon$ for some small $\varepsilon > 0$.

We now establish the convergence of the proposed scheme toward achieving optimal tracking in Theorem 1:

Theorem 1: Assume that the controllability conditions in Lemma 1 hold and $\left(\bar{A}, (Q)^{1/2}\right)$ (for state feedback) or $\left(\bar{A}, (Q_y)^{1/2}\bar{C}\right)$ (for output feedback) is observable. Then the

proposed scheme generates a sequence of controls $\{u_k^j, j = 1, 2, 3, \ldots\}$ that converges to the optimal feedback controller under the rank condition (16), and the tracking error e_k converges to zero if the disturbance and reference vary infrequently relative to the control dynamics.

Proof: Consider first the state feedback algorithms, Algorithms 1 and 2. The bias compensated Q-function (11) satisfies the state feedback Q-learning equation (13), which forms the basis of Algorithms 1 and 2. This equation has a unique solution if the rank condition (16) holds. Given the controllability and observability assumptions on the pairs $(\bar{A}, \bar{B})$, and $\left(\bar{A}, (Q)^{1/2}\right)$, respectively, the PI and VI Q-learning algorithms, Algorithms 1 and 2, converge to the optimal feedback matrix K^* as shown in [30], [35]. Under K^*, the closed-loop dynamics are given by

$$X_{k+1} = \left(\bar{A} - \bar{B}K^*\right)X_k + \begin{bmatrix}\bar{D} & \bar{R}\end{bmatrix}\begin{bmatrix}d_k^{\mathrm{T}} & r_k^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}.$$

The disturbance and reference can be considered in a steady state if they vary infrequently relative to the dynamics. Because $\bar{A} - \bar{B}K^*$ is Schur stable, under these external steady-state inputs, X_k reaches its steady-state X_{ss} and, therefore, $w_{k+1} = w_k$. Then, because $w_{k+1} = w_k + e_k$, it follows that the tracking error e_k converges to zero asymptotically.

Consider next the output feedback algorithms, Algorithms 3 and 4. The bias compensated output feedback Q-function satisfies the output feedback Q-learning equation (21), which is employed in Algorithms 3 and 4. Because $(\bar{A}, \bar{B})$ and $\left(\bar{A}, (Q_y)^{1/2}\bar{C}\right)$ are controllable and observable, respectively, the output feedback policy iteration and VI algorithms converge to the optimal output feedback gain $\mathcal{K}^*$, as shown in our previous work [36], under the rank condition (16). The closed-loop dynamics under the output feedback control are

$$\begin{aligned}X_{k+1} = {} & \bar{A}X_k - \bar{B}K^*\left[(M_u\bar{u}_{k-1,k-N} + M_y\bar{y}_{k-1,k-N})^{\mathrm{T}}w_k^{\mathrm{T}}\right]^{\mathrm{T}} \\ & + \begin{bmatrix}\bar{D} & \bar{R}\end{bmatrix}\begin{bmatrix}\bar{d}_k^{\mathrm{T}} & r_k^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}.\end{aligned}$$

Adding and subtracting the missing disturbance sequence $M_d\bar{d}_{k-1,k-N}$ to obtain the state feedback form using the parameterization (17) results in

$$\begin{aligned}X_{k+1} = {} & \left(\bar{A} - \bar{B}K^*\right)X_k + \begin{bmatrix}\bar{D} & \bar{R}\end{bmatrix}\begin{bmatrix}d_k^{\mathrm{T}} & r_k^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}} \\ & + \bar{B}K_{1:n}^*M_d\bar{d}_{k-1,k-N}.\end{aligned}$$

Then, similar to the state feedback case, we also have $d_{k-1,k-N}$ in the steady state, which, in view of the fact that $\bar{A} - \bar{B}K^*$ is Schur stable, implies that X_k will reach the steady-state X_{ss} and $w_{k+1} = w_k$. Therefore, the tracking error e_k converges to zero asymptotically. ■

IV. Application to HVAC Zone Control

In this section, we apply the proposed scheme to design an HVAC controller for a zone in a commercial building. This is a multi-objective optimal control problem that requires accounting for both the zone comfort and the energy consumption. To formulate this problem into the presented mathematical framework, the zone comfort is associated with obtaining the desired thermal state (i.e., set point temperature) and the energy cost corresponds to the control energy utilized by the actuators. We consider the AHU as the actuator that supplies the zone with air of an appropriate temperature (i.e., supply air) to manipulate the zone temperature. The dynamic model of an HVAC zone used in this case study is adapted from [37]. The thermal dynamics of a building zone are given by the following set of differential equations:

$$\begin{aligned}\frac{dT_z}{dt} = {} & \frac{f_{sa}\rho_a c_{pa}}{C_z}(T_{sa} - T_z) + 2\frac{U_{w_{ew}}A_{w_{ew}}}{C_z}(T_{w_{ew}} - T_z) \\ & + 2\frac{U_{w_{ns}}A_{w_{ns}}}{C_z}(T_{w_{ns}} - T_z) + \frac{K_o}{C_z}(T_o - T_z) + \frac{q}{C_z}, \\ \frac{dT_{w_{ew}}}{dt} = {} & \frac{U_{w_{ew}}A_{w_{ew}}}{C_{w_{ew}}}(T_z - T_{w_{ew}}) + \frac{U_{w_{ew}}A_{w_{ew}}}{C_{w_{ew}}}(T_o - T_{w_{ew}}), \\ \frac{dT_{w_{ns}}}{dt} = {} & \frac{U_{w_{ns}}A_{w_{ns}}}{C_{w_{ns}}}(T_z - T_{w_{ns}}) + \frac{U_{w_{ns}}A_{w_{ns}}}{C_{w_{ns}}}(T_o - T_{w_{ns}}),\end{aligned}$$

which is discretized with a sampling period of one minute to obtain a state space model of the form (1). The description of the quantities is given in the Nomenclature. The nominal parameters given in the Nomenclature are only used to compute the true values of optimal parameters in order to compare our estimates. In other words, the proposed control scheme itself does not require any knowledge of these parameters and the optimal control parameters are learned online.

In this model the outside temperature and the heat gains from the occupants, lights, *etc.*, are the disturbances, which are all assumed to be unmeasurable. Let the user-defined performance index be specified as $Q_e = 300$, $Q_w = 60$, and $R = 100$. Note that $Q_w \neq 0$ is needed for the observability of the pair $\left(\bar{A}, (Q)^{1/2}\right)$ of the augmented system. The optimal feedback control gain for the augmented dynamics (3) can be found by solving the Riccati (7) and is given by

$$K^* = \begin{bmatrix}1.6864 & 0.1413 & 0.1829 & 0.5906\end{bmatrix}.$$

Before presenting the results of the bias compensated Q-learning algorithms, we will first test the uncompensated Q-learning algorithm, Algorithm 0, to analyze the effect of the unmeasurable disturbances. The parameter estimates experience bias as a result of the external disturbances. The final estimate of the control matrix is

$$\hat{K} = \begin{bmatrix}-0.6861 & 25.4644 & -32.9402 & -1.0840\end{bmatrix}.$$

In this particular case, even though the algorithm was initialized with a stabilizing control matrix, the final control gain estimate is not only biased but also destabilizing. This can be seen from the eigenvalues of the resulting closed-loop dynamics matrix $\bar{A} - \bar{B}\hat{K}$, which are 0.3235, 1.5391, $0.9695 \pm j0.0114$ with $\lambda = 1.5391$ being the unstable eigenvalue. For comparison with existing works, recall the robust off-policy algorithms in [24] and [13]. The off-policy method in [24] employs the knowledge of the input and disturbance matrices, whereas the method in [13] removes this requirement. However, both of these methods require access to the disturbance data during the learning phase (see Step 1 of Algorithm 2 in [13], [24]). The model-free algorithm in [13] is applied to our problem for comparison with our model-free method, first with the disturbance data during the learning phase and then without. A discounting factor of $\alpha = 0.1$ and the disturbance attenuation level of $\gamma = 5$ were selected for the H_∞ cost

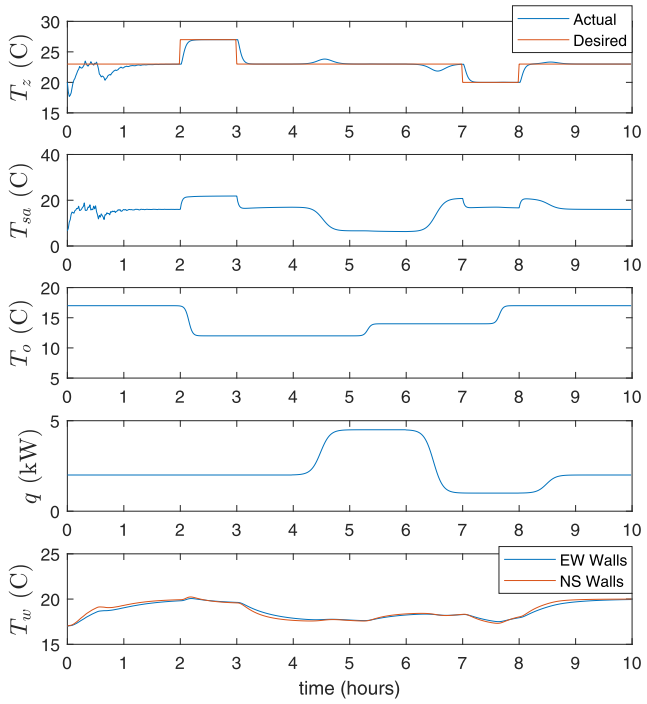

Fig. 1. Evolution of the closed-loop system under Algorithm 1.

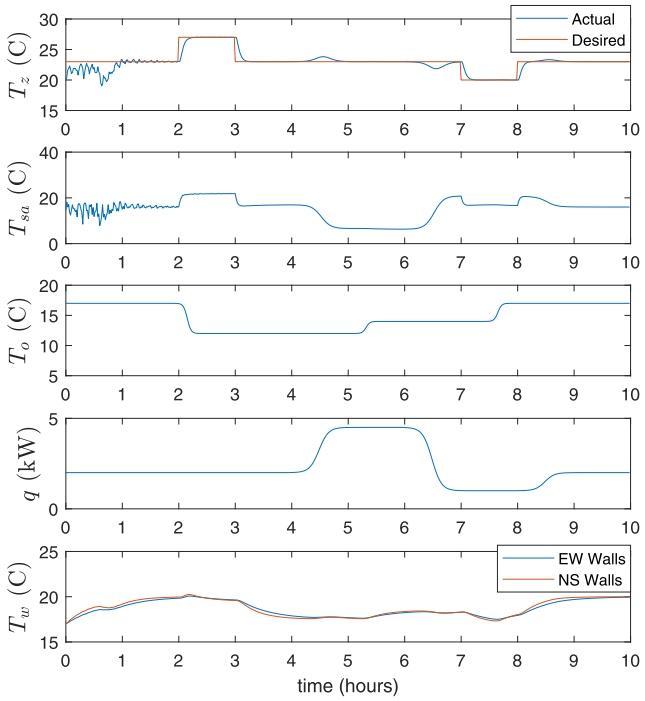

Fig. 2. Evolution of the closed-loop system under Algorithm 2.

function described in [13]. The optimal control gain obtained by this method is

$$K^* = \begin{bmatrix} 44.7092 & 0.1066 & 0.1331 & -45.6562 \end{bmatrix},$$

where the last element corresponds to the reference trajectory, T_{zr}. The learning algorithm in [13] was then applied, with disturbance signals T_o and q known during the learning phase. The final estimate of the control gain is

$$\hat{K} = \begin{bmatrix} 44.7095 & 0.1067 & 0.1329 & -45.6565 \end{bmatrix}$$

which shows convergence to the optimal solution, consistent with the results in [13]. Next, we apply the learning algorithm from [13] without the measurement of the disturbance, although the same disturbances act on the system. In this case, the final estimate of the control gain is

$$\hat{K} = \begin{bmatrix} -72.1094 & -0.6641 & -2.7734 & 52.1094 \end{bmatrix}.$$

The lack of measurement of the disturbances has resulted in a bias in the estimates. Our algorithms address this limitation.

We will now focus on our proposed bias compensated Q-learning algorithms. Let us consider first the policy iteration algorithm, Algorithm 1. We start the algorithm with a stabilizing initial control, which, in this example, is a simple proportional-integral controller. The proportional gain is 0.8432 and the integral gain 0.2953. This corresponds to an initial policy $K^0 = \begin{bmatrix} 0.8432 & 0 & 0 & 0.2953 \end{bmatrix}$. To satisfy the rank condition (16), we add sinusoids of different frequencies and magnitudes to the feedback control policy for the supply air temperature T_{sa}. This initial control is applied during the first 30 minutes to collect online system data. Fig. 1 shows the evaluation of the closed-loop response under Algorithm 1.

During the first 30 minutes we see an exploratory response while the output is still trying to track the reference signal. This is a result of applying an already stabilizing policy that could provide suboptimal tracking in the presence of added exploratory signals. These 30 minutes of online data are then utilized to solve the Bellman equation in the policy iteration step and to update the control parameters in the subsequent iterations $j = 1, 2, \ldots$.

At the beginning of hour 2, a disturbance is introduced as a result of a decrease in the outside temperature, which is almost seamlessly compensated by the controller with an expected increase in the supply air temperature to compensate for the outside temperature drop. At the same time, the desired temperature set point increases. As can be seen, the zone temperature responds to this change and converges to the new set point. Similarly, when the heat gain load changes because of occupancy, lighting, or other sources, the proposed scheme is able to track the desired reference trajectory in the remainder of the period. The proposed Q-learning scheme learns the optimal control parameters for the augmented system while compensating for the external disturbances that would otherwise cause Q-learning to diverge. The final estimates of the optimal control gain are

$$\hat{K} = \begin{bmatrix} 1.6865 & 0.1414 & 0.1831 & 0.5907 \end{bmatrix},$$

which is close to the optimal value despite the presence of the unmeasurable disturbances. This is a result of the bias compensation mechanism introduced in the Q-function and is an advantage of the proposed scheme.

We now proceed to validate the proposed VI algorithm, Algorithm 2. We test this algorithm under the same conditions as for Algorithm 1. Different from the PI algorithm, we initialize the VI algorithm with zero feedback gain, that is, $K^0 = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$. Clearly, this gain is nonstabilizing and we cannot expect tracking during the first 30 minutes of learning,

as can be seen in the zone temperature response in Fig. 2. It is interesting to note that the post learning trajectories, after the first 30 minutes of learning, are the same for both the PI and VI algorithms. This is because both the algorithms eventually converge to the optimal control parameters. The final estimate of the optimal control gain is

$$\hat{K} = \begin{bmatrix} 1.6862 & 0.1410 & 0.1822 & 0.5905 \end{bmatrix}.$$

Note that more iterations are required for Algorithm 2 to converge to the optimal parameters because the search space, which is not limited by a stabilizing initial controller, is larger.

The results presented so far dealt with the full state feedback case, that is, the measurement of the internal state was required for both Algorithms 1 and 2 to learn the optimal control parameters. In the following, we present results for Algorithms 3 and 4 that do not impose this requirement. These algorithms are driven completely by the input-output data instead of requiring the internal state information. For the HVAC application, this means that we no longer need to install sensors on the walls to measure the wall temperature. Instead, only zone temperature measurements are required. This reflects a more realistic HVAC control system.

For the purpose of comparison, the user-defined cost matrices and the rest of the conditions for the output feedback algorithms are kept the same as with the state feedback algorithms. We first validate the output feedback policy iteration algorithm, Algorithm 3. We utilize a proportional-integral controller for initial tracking. The nominal optimal output feedback control parameters for the augmented dynamics (3) can be found by solving the Riccati (7) and the state parameterization (17) and are given by

$$\mathcal{K}^* = [\ 0.4230 \quad -4.5286 \quad 3.1156 \quad 19.2045 \quad -24.2879 \quad 8.1313 \quad 0.5906\ ].$$

Algorithm 3 involves a longer learning phase because there are more unknown parameters to be determined. Specifically, we collected 50 datasets of the input-output data as compared to the 30 datasets for the state feedback algorithms. It can be seen in Fig. 3 that the output feedback algorithm regulates the zone temperature well, similar to the state feedback algorithm but without requiring wall temperature measurements. The final estimate of the output feedback optimal control gain using Algorithm 3 is

$$\hat{\mathcal{K}} = [\ 0.4230 \quad -4.5232 \quad 3.1105 \quad 19.1830 \quad -24.2534 \quad 8.1178 \quad 0.5906\ ]$$

which is close to the optimal output feedback parameters $\mathcal{K}^*$.

If stabilizing initial output feedback parameters are not known, then Algorithm 4 can be applied. In this case the output feedback control parameters are initialized to zero. The system response is shown in Fig. 4. In the absence of a stabilizing initial feedback law, the zone temperature is unable to track the desired reference temperature. In the post learning response, we see that it is able to track the reference signal quite well even in the presence of the unmeasurable disturbances arising from heat gain and outside climate variations. The

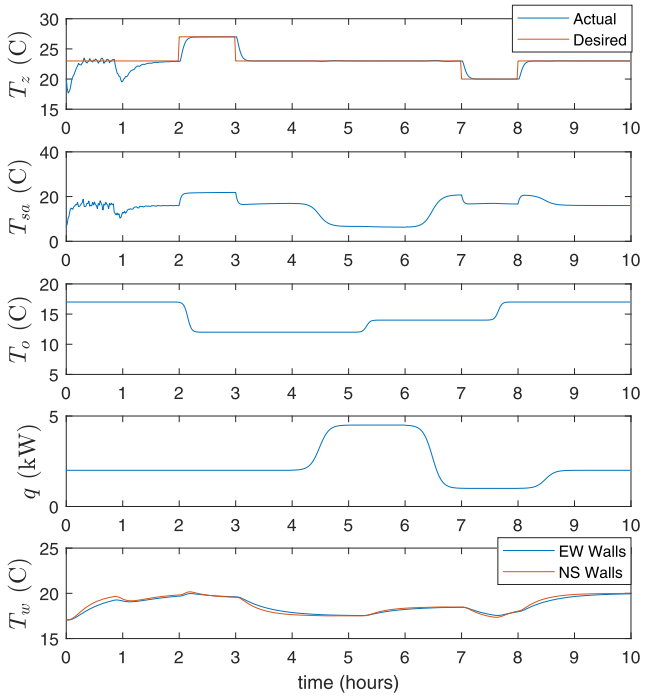

Fig. 3. Evolution of the closed-loop system under Algorithm 3.

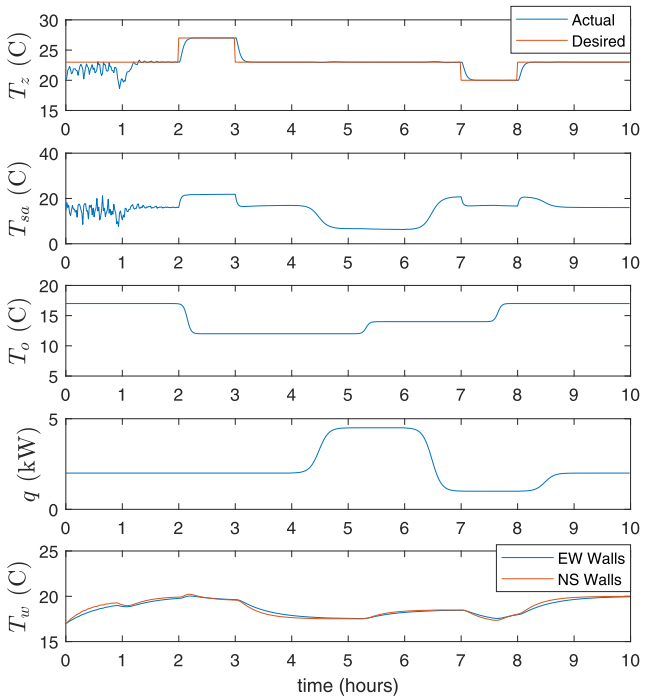

Fig. 4. Evolution of the closed-loop system under Algorithm 4.

final estimate of the output feedback optimal control gain using Algorithm 4 is

$$\hat{\mathcal{K}} = [\ 0.4230\ -4.5350\ 3.1222\ 19.2303\ -24.3312\ 8.1484\ 0.5906\ ]$$

which is also close to the optimal output feedback parameters $\mathcal{K}^*$.

V. Conclusions

This article presented a model-free solution to the optimal tracking problem involving unmeasurable disturbances

based on the framework of reinforcement learning. A new Q-learning based scheme was proposed with a bias compensation mechanism to account for the effect of the disturbance on the learning estimates. An extended Q-function was employed that includes bias compensation terms to prevent the control parameters from drifting away in the presence of the disturbance. Both PI and VI algorithms based on state feedback and output feedback were presented to learn the optimal parameters and to guarantee convergence of the tracking error to zero. Finally, the proposed scheme was validated by designing an optimal set point tracking controller for a practical HVAC zone system in the presence of the unknown disturbances related to outside climate variations and the internal heat gains. In our future work, we will consider extending the design to develop a distributed control scheme for a more complex HVAC system.

REFERENCES

[1] Y. Hong, J. Hu, and L. Gao, "Tracking control for multi-agent consensus with an active leader and variable topology," *Automatica*, vol. 42, no. 7, pp. 1177–1182, Jul. 2006.

[2] F. Liao, J. L. Wang, and G.-H. Yang, "Reliable robust flight tracking control: An LMI approach," *IEEE Trans. Control Syst. Technol.*, vol. 10, no. 1, pp. 76–89, Aug. 2002.

[3] G. Lymperopoulos and P. Ioannou, "Distributed adaptive HVAC control for multi-zone buildings," in *Proc. IEEE 58th Conf. Decis. Control (CDC)*, Dec. 2019, pp. 8142–8147.

[4] C. Mu, Z. Ni, C. Sun, and H. He, "Air-breathing hypersonic vehicle tracking control based on adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 584–598, Mar. 2017.

[5] G. Tao, *Adaptive Control Design and Analysis*. Hoboken, NJ, USA: Wiley, 2003.

[6] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 3051–3056, Nov. 2014.

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

[8] K. G. Vamvoudakis, "Optimal trajectory output tracking control with a Q-learning algorithm," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2016, pp. 5752–5757.

[9] B. Luo, D. Liu, T. Huang, and D. Wang, "Model-free optimal tracking control via critic-only Q-learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2134–2144, Oct. 2016.

[10] K. G. Vamvoudakis, H. Modares, B. Kiumarsi, and F. L. Lewis, "Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online," *IEEE Control Syst.*, vol. 37, no. 1, pp. 33–52, Feb. 2017.

[11] O. Tutsoy, D. E. Barkana, and H. Tugal, "Design of a completely model free adaptive control in the presence of parametric, non-parametric uncertainties and random control signal delay," *ISA Trans.*, vol. 76, pp. 67–77, May 2018.

[12] S. He, H. Fang, M. Zhang, F. Liu, X. Luan, and Z. Ding, "Online policy iterative-based H_∞ optimization algorithm for a class of nonlinear systems," *Inf. Sci.*, vol. 495, pp. 1–13, Aug. 2019.

[13] H. Modares, F. L. Lewis, and Z.-P. Jiang, "H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2550–2562, Jun. 2015.

[14] Y. Liu, Z. Wang, and Z. Shi, "H_∞ tracking control for linear discrete-time systems via reinforcement learning," *Int. J. Robust Nonlinear Control*, vol. 30, no. 1, pp. 282–301, Jan. 2020.

[15] Y. Peng, Q. Chen, and W. Sun, "Reinforcement Q-learning algorithm for H_∞ tracking control of unknown discrete-time linear systems," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 50, no. 11, pp. 4109–4122, Nov. 2020.

[16] B. Luo, Y. Yang, and D. Liu, "Policy iteration Q-learning for data-based two-player zero-sum game of linear discrete-time systems," *IEEE Trans. Cybern.*, early access, Feb. 20, 2020, doi: 10.1109/TCYB.2020.2970969.

[17] O. Tutsoy and M. Brown, "An analysis of value function learning with piecewise linear control," *J. Experim. Theor. Artif. Intell.*, vol. 28, no. 3, pp. 529–545, May 2016.

[18] O. Tutsoy and M. Brown, "Chaotic dynamics and convergence analysis of temporal difference algorithms with bang-bang control," *Optim. Control Appl. Methods*, vol. 37, no. 1, pp. 108–126, Jan. 2016.

[19] J. Huang, *Nonlinear Output Regulation: Theory and Applications*. Philadelphia, PA, USA: SIAM, 2004.

[20] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4164–4169, Dec. 2016.

[21] C. Chen, H. Modares, K. Xie, F. L. Lewis, Y. Wan, and S. Xie, "Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics," *IEEE Trans. Autom. Control*, vol. 64, no. 11, pp. 4423–4438, Nov. 2019.

[22] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, Oct. 2012.

[23] R. Postoyan, L. Busoniu, D. Nešić, and J. Daafouz, "Stability analysis of discrete-time infinite-horizon optimal control with discounted cost," *IEEE Trans. Autom. Control*, vol. 62, no. 6, pp. 2736–2749, Jun. 2017.

[24] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for H_∞ control design," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 65–76, 2014.

[25] Z. Xiao, J. Li, and P. Li, "Output feedback H_∞ control for linear discrete-time multi-player systems with multi-source disturbances using off-policy Q-learning," *IEEE Access*, vol. 8, pp. 208938–208951, 2020.

[26] Y. Jiang, B. Kiumarsi, J. Fan, T. Chai, J. Li, and F. L. Lewis, "Optimal output regulation of linear discrete-time systems with unknown dynamics using reinforcement learning," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3147–3156, Jul. 2020.

[27] F. L. Lewis and V. L. Syrmos, *Optimal Control*. Hoboken, NJ, USA: Wiley, 1995.

[28] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[30] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proc. Amer. Control Conf. (ACC)*, 1994, pp. 3475–3479.

[31] S. He, M. Zhang, H. Fang, F. Liu, X. Luan, and Z. Ding, "Reinforcement learning and adaptive optimization of a class of Markov jump systems with completely unknown dynamic information," *Neural Comput. Appl.*, vol. 32, pp. 14311–14320, 2020.

[32] S. He, H. Fang, M. Zhang, F. Liu, and Z. Ding, "Adaptive optimal control for a class of nonlinear systems: The online policy iteration approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 2, pp. 549–558, Feb. 2020.

[33] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

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

[35] T. Landelius, "Reinforcement learning and distributed local model synthesis," Ph.D. dissertation, Dept. Elect. Eng., Linköping Univ. Electron. Press, Lidingö, Sweden, 1997.

[36] S. A. A. Rizvi and Z. Lin, "Output feedback Q-learning control for the discrete-time linear quadratic regulator problem," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1523–1536, May 2019.

[37] B. Tashtoush, M. Molhim, and M. Al-Rousan, "Dynamic model of an HVAC system for control analysis," *Energy*, vol. 30, no. 10, pp. 1729–1745, Jul. 2005.

Syed Ali Asad Rizvi received the B.E. degree in industrial electronics from the Institute of Industrial Electronics Engineering, NED University of Engineering and Technology, Karachi, Pakistan, in 2012, the M.S. degree in electrical engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2014, and the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2020.

He is currently a Post-Doctoral Fellow with the National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. His current research interests include artificial intelligence, reinforcement learning control, robust control, distributed learning and optimization, and their application in cyber-physical systems.

Amanda J. Pertzborn received the Ph.D. degree from the University of Wisconsin-Madison, Madison, WI, USA.

She is currently the PI of the Intelligent Building Agents Project with the Building Energy and Environment Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. Her research focuses on intelligent control of heating, ventilation, and air conditioning (HVAC) systems.

Zongli Lin (Fellow, IEEE) received the B.S. degree in mathematics and computer science from Xiamen University, Xiamen, China, in 1983, the Master of Engineering degree in automatic control from the Chinese Academy of Space Technology, Beijing, China, in 1989, and the Ph.D. degree in electrical and computer engineering from Washington State University, Pullman, WA, USA, in 1994.

He is currently the Ferman W. Perry Professor with the School of Engineering and Applied Science and a Professor of electrical and computer engineering with the University of Virginia, Charlottesville, VA, USA. His current research interests include nonlinear control, robust control, and control applications.

Dr. Lin is also a fellow of IFAC and the American Association for the Advancement of Science (AAAS). He was the Program Chair of the 2018 American Control Conference and the General Chair of the 13th and 16th International Symposium on Magnetic Bearings, held in 2012 and 2018, respectively. He was an Associate Editor of IEEE TRANSACTIONS ON AUTOMATIC CONTROL from 2001 to 2003, IEEE/ASME TRANSACTIONS ON MECHATRONICS from 2006 to 2009, and *IEEE Control Systems Magazine* from 2005 to 2012. He was elected as a member of the Board of Governors of the IEEE Control Systems Society from 2008 to 2010 and 2019 to 2021 and chaired the IEEE Control Systems Society Technical Committee on Nonlinear Systems and Control from 2013 to 2015. He has served on the operating committees of several conferences. He also serves on the editorial boards of several journals and book series, including *Automatica*, *Systems & Control Letters*, and Birkhauser book series, *Control Engineering*.