Development of a Bias Compensating Q-Learning Controller for a Multi-Zone HVAC Facility

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

***Abstract*—We present the development of a bias compensating reinforcement learning (RL) algorithm that optimizes thermal comfort (by minimizing tracking error) and control utilization (by penalizing setpoint deviations) in a multi-zone heating, ventilation, and air-conditioning (HVAC) lab facility subject to unmeasurable disturbances and unknown dynamics. It is shown that the presence of unmeasurable disturbance results in an inconsistent learning equation in traditional RL controllers leading to parameter estimation bias (even with integral action support), and in the extreme case, the divergence of the learning algorithm. We demonstrate this issue by applying the popular Q-learning algorithm to linear quadratic regulation (LQR) of a multi-zone HVAC environment and showing that, even with integral support, the algorithm exhibits bias issue during the learning phase when the HVAC disturbance is unmeasurable due to unknown heat gains, occupancy variations, light sources, and outside weather changes. To address this difficulty, we present a bias compensating learning equation that learns a lumped bias term as a result of disturbances (and possibly other sources) in conjunction with the optimal control parameters. Experimental results show that the proposed scheme not only recovers the bias-free optimal control parameters but it does so without explicitly learning the dynamic model or estimating the disturbances, demonstrating the effectiveness of the algorithm in addressing the above challenges.**

***Index Terms*—HVAC control, optimal tracking, Q-learning, reinforcement learning (RL).**

I. Introduction

A promising approach to making a building both comfortable for occupants and energy efficient is through optimal control of the heating, ventilation, and air conditioning (HVAC) system [1]. Optimal control of HVAC systems is of practical significance to our society as people spend almost 85% of their time inside buildings [2]. Buildings account for around 76% of electricity consumption in the US [3]. Optimal control theory provides a framework for the design of controllers that meet the prescribed performance and cost objectives. Optimal controllers such as model predictive control (MPC) are popular in HVAC controls even though they involve a complex design process that requires reasonably accurate HVAC models and an estimate of external disturbances [4].

Recently, particularly in the last five years, the HVAC controls literature has shifted to artificial intelligence (AI)-based approaches [5] to designing optimal controllers. The framework of reinforcement learning (RL) is an AI-based optimal decision making paradigm and is considered one of the three pillars of machine learning (ML), with the other two being supervised and unsupervised learning [6]. RL is a promising candidate in HVAC controls because of its ability to learn optimal control policies without solving a traditional model-based optimization problem [7]. Compared to tuning based proportional-integral-derivative (PID) control [8] and ML-based system identification approaches, model-free RL is a direct adaptive control approach that learns the optimal control actions without learning the underlying system models [9]. Thus, model-free RL control is different from the two-step learning based MPC approaches that perform optimization based on the learned model.

RL-based HVAC control literature has been growing recently at different levels of control, ranging from the component-level control [10]–[12] to system-level control [13]–[15]. A comprehensive review of recent studies has been carried out in [16]. While there is significant development in RL control literature to handle nonlinear problems with constraints [17], a drawback with model-free RL is that optimal decision making tends to be a completely blackbox operation and does not take advantage of the physical insights of the HVAC system. As a result, the learning process tends to be long and the resulting control lacks intuition [10], [18]. Early applications of RL in HVAC controls have demonstrated decent control performance at the expense of significantly long learning times [19], [20]. In [21], a comparison of black-box RL based HVAC control with on-off and Fuzzy-PD control showed that while RL outperforms (in thermal comfort), the exhaustive exploration require long training periods and inadvertently limit the performance gains. To address this difficulty, some researchers have explored reduced dimensional approaches to accelerate learning [22]–[24]. Event-based approaches have also been proposed towards reducing learning times [25]. Physical insights of the system can provide

Manuscript received March 6, 2023; revised April 12, 2023; accepted April 18, 2023. This work was supported in part by NIST (70NANB18H161). Recommended by Associate Editor Qing-Long Han. *(Corresponding author: Zongli Lin.)*

Citation: S. A. A. Rizvi, A. J. Pertzborn, and Z. Lin, "Development of a bias compensating Q-learning controller for a multi-zone HVAC facility," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 8, pp. 1704–1715, Aug. 2023.

S. A. A. Rizvi is with the Department of Electrical and Computer Engineering, Tennessee Technological University, Cookeville, TN 38505 USA (e-mail: srizvi@tntech.edu).

A. J. Pertzborn is with the Mechanical Systems and Controls Group in the Building Energy and Environment Division at the National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899 USA (e-mail: amanda.pertzborn@nist.gov).

Z. Lin is with the Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904-4743 USA (e-mail: zl5y@virginia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.123624

valuable information about the structure of the control problem. The idea of incorporating partial knowledge of HVAC dynamics with RL has been considered recently [26], [27]. One of the main reasons for the success of traditional PID control is that the adjustment of control parameters is transparent. However, it does not have a formal design procedure to optimize a prescribed objective function. On the other hand, optimal control methods such as the linear quadratic regulator (LQR) solve an optimization problem involving some user-defined cost function. RL has been successfully used to solve this kind of optimal control problem. Distributed multi-agent techniques are also being incorporated in RL that continue to enhance its innate capabilities [28].

Model-free learning approaches based on RL have been proposed to solve optimal control problems involving disturbances. Two recently popular model-free optimal control approaches to addressing disturbance rejection involve integration of these learning methods with the frameworks of game theory [29] and output regulation [30]. Zero-sum game formulations have been presented in [31], [32]. Inverse RL extensions of these game theory based disturbance rejection designs have also been recently proposed [33], [34]. One of the key difficulties in the treatment of disturbances in the mentioned frameworks is that the disturbance is generally assumed to be measurable to prevent any mismatch in the learning equation with respect to the system dynamics [29]. If not taken into account, the unknown disturbance leads to an inconsistent learning equation, which ultimately leads to estimation bias and could potentially render the closed-loop system unstable.

In this paper, we address the issue of unmeasurable disturbances that are a potential cause of bias (and even instability) in the control estimates for a multi-zone experimental lab facility. It is shown that the RL algorithm learns control actions that meet a prescribed optimality criterion. A range of design considerations have been employed for the practical realization of the proposed scheme. This includes reducing learning times compared to the blackbox RL methods without requiring simulated data sources. Real-world implications of different excitation signals for satisfying RL exploration trade-off and role of numeric solvers in the development of the proposed algorithm are discussed.

The main contributions of this work are as follows. We present the design and implementation aspects of a novel bias compensating reinforcement learning algorithm in a multi-zone commercially equipped HVAC testing facility at the National Institute of Standards and Technology (NIST). Industrial HVAC systems are a great candidate for harnessing the capabilities of such AI algorithms given the significant impact buildings have on our energy footprint. This work is motivated by the need to accelerate the mainstream industrial acceptance of such learning-based optimal control algorithms in commercial settings by presenting various implementation aspects that are critical to the realization of these algorithms. We believe our work is one of the few experimental studies that consider a critical phenomenon of disturbance-induced bias found in RL control literature, which we address by enhancing the learning equation and introducing a bias compensation mechanism. Our study shows that the bias issue, which could lead to suboptimality and instabilities, could be prevented by taking the presented measures, without requiring model identification and disturbance estimation stages.

The rest of the paper is organized as follows. Section II provides a control formulation for the zone-level HVAC control. Section III presents the design of a disturbance compensating Q-learning algorithm. Section IV describes the testing facility and details of the experimental setup. Implementation aspects and results of the study are presented in Section V. Section VI concludes the paper.

II. Problem Formulation

We formulate the zone-level control problem into a mathematical form that is tractable by the structured RL controller. Note that the learning algorithm itself requires the knowledge of neither the zone model parameters nor the unmeasureable load disturbances. Instead, it learns the optimal parameters within the control structure introduced in the following subsections.

A. Zone Dynamics

We consider a zone model that captures the level of complexity required for our design objective, which is to regulate zone thermal comfort based on a prescribed weighted tracking error and setpoint deviations of the equipment. For this purpose, we adopt the following standard dynamic model of an HVAC zone based on ordinary differential equations [35]:

$$\frac{dT_z}{dt} = \frac{f_{sa}\rho_a C_{pa}}{C_z}(T_{sa} - T_z) + 2\frac{U_{w_{ew}}A_{w_{ew}}}{C_z}(T_{w_{ew}} - T_z) + 2\frac{U_{w_{ns}}A_{w_{ns}}}{C_z}(T_{w_{ns}} - T_z) + \frac{q}{C_z}$$

$$\frac{dT_{w_{ew}}}{dt} = \frac{U_{w_{ew}}A_{w_{ew}}}{C_{w_{ew}}}(T_z - T_{w_{ew}}) + \frac{U_{w_{ew}}A_{w_{ew}}}{C_{w_{ew}}}(T_o - T_{w_{ew}})$$

$$\frac{dT_{w_{ns}}}{dt} = \frac{U_{w_{ns}}A_{w_{ns}}}{C_{w_{ns}}}(T_z - T_{w_{ns}}) + \frac{U_{w_{ns}}A_{w_{ns}}}{C_{w_{ns}}}(T_o - T_{w_{ns}}).$$

Table I contains a description of the variables and constant parameters. The above zone model can be represented in a compact matrix form by defining the state vector $x = [T_z \; T_{w_{ew}} \; T_{w_{ns}}]^T$, the control input $u = T_{sa}$, and the disturbance vector $d = [T_o \; q]^T$. The resulting state-space model is

$$\frac{dx(t)}{dt} = \mathcal{A}x(t) + \mathcal{B}u(t) + \mathcal{D}d(t). \tag{1}$$

The state-space model in (1) needs to be discretized (in time) for digital implementation as well as due to the discrete nature of the reinforcement learning algorithm considered in this paper. The standard Euler discretization method can be used to obtain the following discrete-time state-space model:

$$x_{k+1} = Ax_k + Bu_k + Dd_k \tag{2}$$

in which k is the discrete-time index, and the discretized matrices are computed as $A = e^{\mathcal{A}T_s}$, $B = \int_0^{T_s} e^{\mathcal{A}\tau}d\tau\mathcal{B}$ and $D = \int_0^{T_s} e^{\mathcal{A}\tau}d\tau\mathcal{D}$, with T_s being the sampling period. The dimen-

TABLE I
NOMENCLATURE

Symbol	Description
$A_{w_{ew}}$	Area of East/West walls (m^2)
$A_{w_{ns}}$	Area of North/South walls (m^2)
C_{pa}	Specific heat of air (kJ/kgC)
$C_{w_{ew}}$	Thermal capacitance of East/West walls (kJ/C)
$C_{w_{ns}}$	Thermal capacitance of North/South walls (kJ/C)
C_z	Thermal capacitance of the zone (kJ/C)
f_{sa}	Volume flow rate of the supply air (m^3)
q	Heat gain from occupants, lights, doors (Watts) (W)
T_o	Outside temperature (degrees Celsius) (C)
T_{sa}	Supply air temperature (degrees Celsius) (C)
$T_{w_{ew}}$	Temperature of the East/West walls (degrees Celsius) (C)
$T_{w_{ns}}$	Temperature of the North/West walls (degrees Celsius) (C)
T_z	Temperature of the zone (degrees Celsius) (C)
$U_{w_{ew}}$	Heat transfer coefficient of East/West walls (W/m^2C)
$U_{w_{ns}}$	Heat transfer coefficient of North/West walls (W/m^2C)
ρ_a	Density of air (Kg/m^3)

sions of the matrices in (2) are compatible with dimensions of the variable definitions in (1).

B. Control Structure

We are interested in maintaining thermal comfort for which we use the temperature tracking error,

$$e_k = T_{z_k} - T_{r_k}$$

as a metric, where T_{r_k} is the reference temperature. The goal is to find an optimal control setpoint $u_k^* = T_{sa}^*$ for the actuator (air handling unit or the variable air volume (VAV) box–see Section IV) that guarantees asymptotic temperature tracking, i.e., $\lim_{k\to\infty} e_k = 0$, while minimizing a quadratic cost function of the form,

$$J = \sum_{i=0}^{\infty} \left(Q_e e_i^2 + R\tilde{u}_i^2 \right) \tag{3}$$

where $Q_e \geq 0$ and $R > 0$ are the scalar weights that penalize the temperature error as a measure of thermal comfort and the setpoint deviation with respect to steady-state $\tilde{u}_k = u_k - u_{\text{ss}}$ (as a measure of control effort), respectively, with u_{ss} being the steady-state setpoint for the supply air temperature or the VAV box.

We adopt a standard proportional integral control structure that we eventually want the RL algorithm to learn. Note that the choice of such a structure is not a restriction for the reinforcement learning algorithm. For instance, instead of following a proportional integral structure, we can employ other structures such as feedforward control. To this end, we introduce a new state w_k that accumulates the tracking error, which is the discrete-time equivalent of the integral action used in the continuous-time setting. Based on this new state, we form the following augmented system:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Dd_k \\ w_{k+1} &= w_k + e_k \end{aligned} \tag{4}$$

which can be represented compactly in terms of the augmented state vector $X_k = [x_k^T \;\; w_k^T]^T$,

$$X_{k+1} = \bar{A}X_k + \bar{B}u_k + \bar{D}d_k + \bar{R}r_k \tag{5}$$

where $r_k = T_{r_k}$ is the zone reference setpoint, as defined earlier and the matrices $\bar{A}, \bar{B}$, and $\bar{D}$ can be readily obtained from (4). The cost function for this augmented formulation follows from (3) as:

$$J = \sum_{i=0}^{\infty} \left(\tilde{X}_i^T Q \tilde{X}_i + \tilde{u}_i^T R \tilde{u}_i \right) \tag{6}$$

where $Q = \begin{bmatrix} C^T Q_e C & 0 \\ 0 & Q_w \end{bmatrix}$, $C = [1 \;\; 0 \;\; 0]$ and $\tilde{X}_i = X_i - X_{\text{ss}}$ for some steady-state X_{ss}. For problems in (5) and (6), there exists a unique optimal controller given as

$$u_k^* = -(R + \bar{B}^T P^* \bar{B})^{-1} \bar{B}^T P^* \bar{A} X_k = -K^* X_k \tag{7}$$

where P^* is the unique positive definite solution to the algebraic Riccati equation (ARE) [36],

$$\bar{A}^T P \bar{A} - P + Q - \bar{A}^T P \bar{B}(R + \bar{B}^T P \bar{B})^{-1} \bar{B}^T P \bar{A} = 0. \tag{8}$$

Note that the weight Q_w is included to weigh the integral (summation) of the error. This weight penalizes the deviation of the integral (summation) state relative to a steady-state. Since the integral (summation) introduces an unstable mode in the closed-loop system, Q_w needs to be positive definite. A detailed derivation of this optimal controller can be found in [37].

III. Bias Compensating Reinforcement Learning Controller

The previous section introduced a control structure for the HVAC zone dynamics. However, the control design involves the knowledge of zone dynamics, which are generally not available. The traditional approach to dealing with such a design problem is to employ system identification or adaptive control techniques. Unlike these approaches, reinforcement learning has the potential to directly learn the control actions or a control structure (as is the case in this work) based on a prescribed optimality criterion. In this section, we are interested in employing reinforcement learning to learn the control structure that we introduced in the previous section. In particular, we employ a disturbance compensated Q-learning algorithm that was developed in our recent work [37].

A. Reward Function

A critical part of the design of the reinforcement learning controller is the suitable choice of a reward (or penalty/utility) function. The long term cost is evaluated based on this function. Since the long term cost cannot be known *a priori* (as the zone dynamics are unknown), the instantaneous reward obtained from the reward function is primarily used to drive a reinforcement learning algorithm towards learning the optimal control policy. The optimal policy learned from a rein-

forcement learning algorithm corresponds to a particular choice of reward function. As a result, the reward function should be able to capture the essence of the control objective we are to address. The reward function used for learning a predefined control structure based on the optimal control formulation can be readily obtained from the cost function J. For our particular integral (summation) based control structure, it is the instantaneous quadratic cost of the thermal performance and the control utilization,

$$\mathcal{R}_k = X_k^T Q X_k + u_k^T R u_k \tag{9}$$

where the thermal comfort is weighted by the augmented matrix Q as defined earlier. Note that the weight R remains the same as before as it is not affected by the augmentation process.

B. Disturbance Compensated Q-Function

In this work we have considered Q-learning as our primary RL algorithm owing to its model-free capability and its success in learning optimal controllers. Most of the learning algorithms in RL control literature that address disturbance rejection involve measurement (and even manipulation) of the disturbance during the learning phase. In our problem setting, the disturbances result from various sources of heat gain and outside temperature variations. As a result, measurement of the disturbances is generally not available. It was shown in our recent work [37] that these disturbances, if not accounted for, can result in bias in the control policy learned from the standard Q-learning algorithm. Employing the dynamics (5) in the following definition of a Q-function [38]:

$$Q_K = X_k^T Q X_k + u_k^T R u_k + V_K(X_{k+1}) \tag{10}$$

where $V_K(X_{k+1}) = X_{k+1}^T P X_{k+1}$ is the quadratic value function. Substituting it, along with (5), in (10) results in

$$\begin{aligned} Q_K &= \begin{bmatrix} X_k \\ u_k \\ r_k \end{bmatrix}^T \begin{bmatrix} Q + \bar{A}^T P \bar{A} & \bar{A}^T P \bar{B} & \bar{A}^T P \bar{R} \\ \bar{B}^T P \bar{A} & R + \bar{B}^T P \bar{B} & \bar{B}^T P \bar{R} \\ \bar{R}^T P \bar{A} & \bar{R}^T P \bar{B} & \bar{R}^T P \bar{R} \end{bmatrix} \begin{bmatrix} X_k \\ u_k \\ r_k \end{bmatrix} \\ &\quad + 2X_k^T \bar{A}^T P \bar{D} d_k + 2u_k^T \bar{B}^T P \bar{D} d_k + 2r_k^T \bar{B}^T P \bar{D} d_k \\ &\quad + d_k^T \bar{D}^T P \bar{D} d_k \\ &\triangleq (z_k')^T H' z_k' + 2X_k^T \bar{A}^T P \bar{D} d_k + 2u_k^T \bar{B}^T P \bar{D} d_k \\ &\quad + 2r_k^T \bar{B}^T P \bar{D} d_k + d_k^T \bar{D}^T P \bar{D} d_k \end{aligned} \tag{11}$$

where the last four terms involving the unmeasurable signal d_k result in an estimation bias. To address this difficulty, we presented a Q-learning algorithm with disturbance compensation that prevents the bias from occurring during the learning phase. The motivation for introducing the bias compensating term follows from the bias effect itself, which is induced by the terms containing the unmeasurable disturbance d_k in the above Q-function. Therefore, augmenting the regular Q-function with a bias factor c compensates for the missing disturbance term and prevents an estimation bias from occurring in the rest of the terms. This involves the design of the following Q-function with a bias compensation term:

$$\begin{aligned} Q_K &= \begin{bmatrix} X_k \\ u_k \\ r_k \\ c \end{bmatrix}^T \begin{bmatrix} Q + \bar{A}^T P \bar{A} & \bar{A}^T P \bar{B} & \bar{A}^T P \bar{R} & b_1 \\ \bar{B}^T P \bar{A} & R + \bar{B}^T P \bar{B} & \bar{B}^T P \bar{R} & b_2 \\ \bar{R}^T P \bar{A} & \bar{R}^T P \bar{B} & \bar{R}^T P \bar{R} & b_3 \\ b_1^T & b_2^T & b_3^T & b_4 \end{bmatrix} \begin{bmatrix} X_k \\ u_k \\ r_k \\ c \end{bmatrix} \\ &\triangleq \begin{bmatrix} X_k \\ u_k \\ r_k \\ c \end{bmatrix}^T \begin{bmatrix} H_{XX} & H_{Xu} & H_{Xr} & b_1 \\ H_{uX} & H_{uu} & H_{ur} & b_2 \\ H_{rX} & H_{ru} & H_{rr} & b_3 \\ b_1^T & b_2^T & b_3^T & b_4 \end{bmatrix} \begin{bmatrix} X_k \\ u_k \\ r_k \\ c \end{bmatrix} \\ &= z_k^T H z_k \end{aligned} \tag{12}$$

where the sub-matrices H_{**} are defined in terms of the system and cost matrices as they appear in the preceding matrix. The factor c is a bias scaling factor, a virtual factor introduced to mimic the bias effect. Interested readers can refer to [37] for the details of this kind of Q-function. It is worth pointing out that the Q-function above has a well defined quadratic structure that corresponds to the choice of control structure we presented in Section II-B. The Q-function provides a measure of the long term cost of following a certain control policy given the zone dynamics. As the zone dynamics are not known, a Q-learning algorithm is used to learn the Q-function and improve the control policy based on the estimated Q-function.

C. Q-Learning Algorithm

In this section, we present an iterative Q-learning algorithm that provides estimates of the quadratic structured Q-function defined in (12). The iterative algorithm is based on a policy iteration algorithm found in the reinforcement learning literature [39]. Unlike the standard policy iteration based Q-learning algorithm [38], this algorithm seeks to learn the optimal policy subject to the unknown load disturbances that act on the zone (see [37] for a comparison). The Q-learning equation is a key equation that is used to learn the Q-function and the optimal policy, and is given as follows:

$$z_k^T H z_k = X_k^T Q X_k + u_k^T R u_k + z_{k+1}^T H z_{k+1}. \tag{13}$$

Algorithm 1 Disturbance Compensating State Feedback Q-Learning Policy Iteration Algorithm

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 the iteration index $j = 0$.

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

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

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

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

6: Increment index $j \rightarrow j+1$

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

Algorithm 1 is a policy iteration Q-learning algorithm for the zone-level HVAC control. This is essentially a two-step procedure. In the policy evaluation step, we use the key equation (13) to solve for the unknown vector $\bar{H}$. In the policy update step, we perform a one step policy improvement that essentially minimizes the value of the Q-function estimate obtained in the policy update step. These two steps are repeated until the convergence criterion ε is met. Interested readers can refer to [37] for the convergence proof of this algorithm.

D. Solvers for the Bellman Equation

A critical step in the Q-learning algorithm is the evaluation of the policy value by estimating the Q-function. In our problem, this involves estimating the matrix H (as vector $\bar{H}$ with repeated entries of H removed) by solving the Bellman equation (13). While the Bellman equation is linear in terms of the unknown, it is underdetermined. Solvers such as batch least-squares (LS) and recursive least-squares (RLS) are commonly employed to solve the Bellman equation. In this study, we compare the LS and RLS solvers to see the variation in the estimates caused by the solvers.

For the batch least-squares solver, we collect $L \geq l(l+1)/2$ data samples of (X_k, u_k, r_k) to form the data matrices $\Phi \in \mathbb{R}^{l(l+1)/2 \times L}$ and $\Upsilon \in \mathbb{R}^{L \times 1}$, defined as

$$\Phi = \begin{bmatrix} \bar{z}_{k-L+1} - \bar{z}_{k-L+2} & \bar{z}_{k-L+2} - \bar{z}_{k-L+3} \cdots \bar{z}_k - \bar{z}_{k+1} \end{bmatrix}$$

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

where $\bar{z}$ is the quadratic basis. Then, the least-squares solution of (13) is given by

$$\bar{H}^j = \left(\Phi \Phi^T\right)^{-1} \Phi \Upsilon \tag{14}$$

where $\bar{H}^j$ is the jth estimate of the unknown vector $\bar{H}$. Since $u_k = -KX_k$, which is linearly dependent on X_k, (14) will not have a unique solution, which is required for convergence to the optimal parameters. To overcome this issue, an excitation noise is added (only during the learning phase) in u_k to guarantee a unique solution to (14), that is, to ensure the satisfaction of the following rank condition:

$$\rho(\Phi) = l(l+1)/2. \tag{15}$$

It is worth noting that the excitation in u_k will not be able to manipulate the bias scaling factor c. However, this factor is readily adjustable, for example, by selecting a different constant at consecutive intervals. This will prevent zero entries from appearing in the regression vector and ensure the above rank condition is satisfied. Clearly, the rank condition (15) is necessary to obtain the optimal solution, which is a unique solution to the least-squares problem (14).

Recursive least-squares is mainly used to facilitate online implementation of the least-squares solver. In this work, we have considered the following recursive least-squares implementation:

$$\mathcal{E}_k(i) = \mathcal{R}_k - \Phi_k^T(i) \bar{H}(i-1) \tag{16}$$

$$\bar{H}(i) = \bar{H}(i-1) + \frac{P_k(i-1)\Phi_k \mathcal{E}_k(i)}{1 + \Phi_k^T P_k(i-1)\Phi_k} \tag{17}$$

$$P_k(i) = P_k(i-1) - \frac{P_k(i-1)\Phi_k \Phi_k^T P_k(i-1)}{1 + \Phi_k^T P_k(i-1)\Phi_k}$$

$$P_k(0) = P_0 \tag{18}$$

where $P_0 = \beta I$ for some very large $\beta > 0$ and $\Phi_k = \bar{z}_k - \bar{z}_{k+1}$. The convergence of the algorithm is established under the persistence of excitation (PE) condition [40], which assures that the data is sufficiently rich [41]. This PE condition is met by injecting exploration noise in the control signal during the learning phase.

IV. Test Facility And Experimental Setup

The Intelligent Building Agents Laboratory (IBAL), as shown in Fig. 1, was designed to emulate the HVAC system and building loads in a small commercial building, but we utilized only a small portion of the facility for this study (see [42]–[45] for details of the IBAL). Table II shows the specifications of the key equipment with labels corresponding to Fig. 1. Fig. 2 shows most of the air system in the IBAL; D is a damper, which helps control how much air flows through each branch of the system. AHU1 and AHU2 are the air handling units that take air from outside and from the zones (recirculation air) and condition it for use in the zones, i.e., cool and dehumidify. Each AHU conditions air for two zones, but the two zones do not necessarily have the same cooling requirements. The VAVs allow for the common AHU air to be modified for the specific zone by modulating the airflow rate and/or reheating the air from the AHU. So, one of the control decisions in a commercial AHU-VAV system is the temperature of the air supplied to the zones. For this study we controlled the temperature entering the VAV (from the AHU) to approximately 12.8°C (55°F) and then used the electric heater in the VAV to generate the specified setpoint temperature of

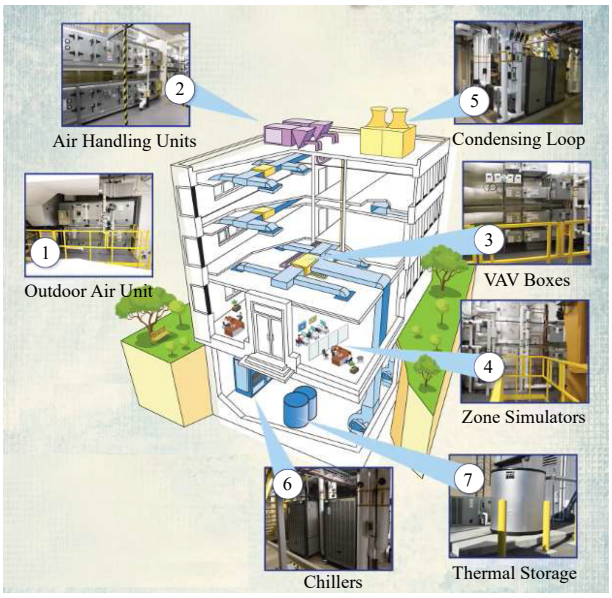

Fig. 1. An overview of the Intelligent Building Agents Laboratory (IBAL).

TABLE II
EQUIPMENT SPECIFICATIONS

1	Outdoor air unit
2	Air Handling Units (2x, 15.5 kW)
3	Variable Air Volume Boxes (4x)
4	Zones Compartments (4x)
5	Condensing Loop
6	Chillers (2x, 26.6 kW and 53 kW)
7	Thermal Storage

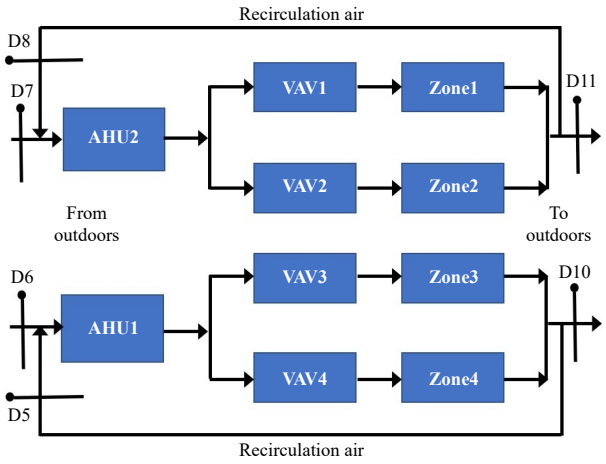

Fig. 2. Overview of the air system in the IBAL.

the air entering the zone. This target setpoint is generated by either a manually tuned PI controller or by the RL algorithm. This is not exactly how the controls work in a real commercial building, but we chose this simplified approach as a proof of concept for the RL control algorithm.

The basic feedback loop involving the RL controller and the VAV-Zone subsystem is shown in Fig. 3. The setpoint temperature downstream of the VAV, $T_{sa,SP}$, is determined based on the error between the zone setpoint temperature, $T_{z,SP}$, and the actual zone temperature, T_z. The VAV heater has its own local PI loop (not shown) running at a faster rate compared to the RL loop that controls the heater to bring the actual temperature, T_{sa}, closer to $T_{sa,SP}$. The RL algorithm learns the gains for the controller that calculates $T_{sa,SP}$.

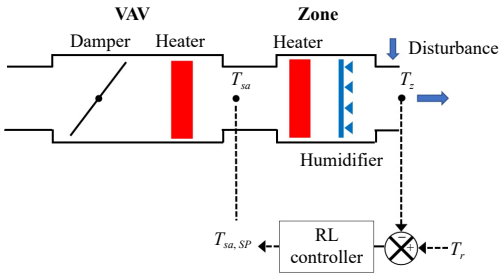

Fig. 3. RL control loop that calculates $T_{sa,SP}$ from the zone temperature and feeds to the VAV heater control signal.

V. IMPLEMENTATION, RESULTS, AND DISCUSSIONS

This section shows the implementation details and the results of the proposed Q-learning scheme. The reward function parameters govern the relative weights of the tracking error and the deviations of the setpoint. In this study, we chose the parameters to be $Q_e = 3000$, $Q_w = 600$ and $R = 100$ for the tracking error, integral (summation) error, and the control effort, respectively. For the discrete-time implementation, a sampling period of one minute was found to be appropriate for our application. In our experimental setup, we have negligible coupling coefficients of the walls due to the insulation in place between the zones. Therefore, the states corresponding to the wall temperatures are not used during the learning (although the proposed RL structure can fully accommodate it, as shown in our simulation study [37]). Thus, we only include the feedback measurements of zone temperature for the state feedback structure.

One important consideration in RL algorithms is that the state-action space needs to be sufficiently explored by means of exploration signals to learn the optimal control actions [6]. For this purpose, a variety of datasets were generated based on the type and magnitude of exploration. We considered the two most commonly employed exploration signals: 1) sinusoids of various frequencies and magnitudes and 2) Gaussian signals of different variances. The final estimates obtained from the batch LS and recursive LS solvers were identical, so we only present plots for the recursive LS solver. Note that while the training datasets were collected, a working stable controller (not necessarily the optimal one) was already running in order to satisfy the initially stabilizing requirement of the PI algorithm and to make sure that the dataset was within the expected operating range of the zones. We refer to this phase of learning as the pre-learning phase, also referred to as the exploration phase, as it involves generation of the training dataset. On the other hand, the phase that employs the learned controller in action is referred to as the post-learning phase. It is during these two phases that the learning iterations in Algorithm 1 are carried out. While the results of all four zones have been included in Table II, plots are shown only for Zones 1 and 2 when the zones exhibit similar behavior.

Before implementing the RL algorithm, we tested a manually tuned PI controller to get an idea of typical gains and responses. This PI controller does not take into account any particular optimality criterion, unlike the RL algorithm that follows a certain reward function to compute the gains. After careful tuning, the values of the proportional and integral gains used were 10 and 1.5. The zone temperature and setpoint variations are similar across all four zones under these manual gains. Fig. 4 shows the responses for Zones 1 and 2.

Starting with the RLS solver, we applied Algorithm 1 to the four zones using two datasets that were generated using sinusoidal exploration of different magnitudes with a number of frequency components. The exploration signal in this case takes the form of

$$\begin{aligned} n_k = A(&0.01\sin(k) + 0.05\sin(3.1k) + 0.03\sin(5.3k) \\ &+ 0.02\cos(2.3k) + 0.07\cos(7.3k) + 0.04\cos(4.3k) \\ &+ 0.01\cos(3.9k) + 0.02\cos(11.9k) + 0.03\cos(13.9k)) \end{aligned} \tag{19}$$

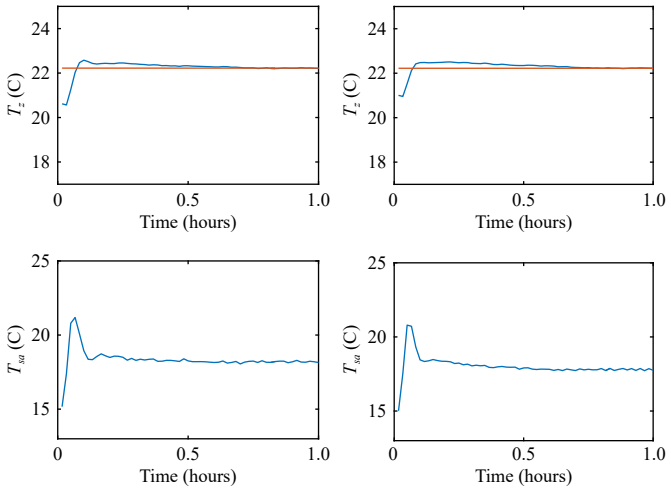

Fig. 4. Response of a manually tuned PI controller for Zone 1 (left) and Zone 2 (right) to a reference temperature T_r.

for $A = 5$ and $A = 10$, which correspond to the sinusoidal exploration datasets S1 and S2, respectively. Such a choice of exploration signal was based on experimental observations that yielded satisfactory response during the learning phase. The sub-harmonic frequencies and amplitudes were arbitrarily selected following these responses. Later discussions explain how a variety of explorations affect the convergence. Note that these explorations are added on top of any initially stable controller, which in our case is just the manually tuned PI controller. We use vector notation to write gains (see Table III), where the first component is the proportional gain and the second is the integral component. Therefore, the initial gain vector is set as $K^0 = [10\ \ 1.5]$. Before applying Algorithm 1, we first use the sinusoidal dataset to learn control gains without bias compensation. The results for Zones 1–4 are shown in the first two rows of Table III, which show non-convergence. In fact, these control gains are not stabilizing as can be seen from their signs. This shows that the regular uncompensated Q-learning LQR algorithm would suffer from bias in the presence of disturbances. We now proceed to apply Algorithm 1 with bias compensation by using the two sinusoidal exploration datasets. The response during the learning phase of Zone 1 using both S1 and S2 datasets is shown in Fig. 5.

The post-learning response and the convergence of the parameter estimates of Zone 1 for both datasets is shown in Figs. 6 and 7, respectively. Upon comparing these figures, we find that convergence of the parameter estimates depends on the magnitude of exploration. This is expected because the convergence of RL algorithms varies with how much the state-action space is explored. We now compare the above convergence results of Zone 1 with those of Zone 2. The pre-learning response of Zone 2 for datasets S1 and S2 is shown in Fig. 8, whereas the post-learning response is shown in Fig. 9. Zone 2 parameters also show convergence subject to different explorations as shown in Fig. 10. The learned control gains for Zone 2 are also quite close to Zone 1, suggesting a nearly identical dynamic environment. Similar observations were made for the remaining zones (only plots for Zones 1 and 2 are shown).

Remark 1: Recently in [46], the decentralized LQ problem was considered and applied to a multi-zone HVAC problem, which also addressed the effect of drifting/bias terms due to unknown disturbance sources (outdoor temperature and heat gains) by learning the bias parameter in the control structure itself. Our approach is in a similar spirit, but instead of learning a control bias term, we learn the bias term arising in the learning equation irrespective of the control. This enables asymptotic temperature tracking (improved thermal comfort) as the error in bias estimates does not explicitly appear in the control. Our multi-zone experimental setup, however, is limited to independent zone control only as it does not take into account the interzone interactions, which would involve a centralized/decentralized formulation.

We continue to analyze the system behavior under Algorithm 1. For both zones and for both datasets S1 and S2, we find that the learned controller tends to give a higher overshoot. However, the tracking response closely follows the reference trajectory under 20 minutes, which is a slight improvement over the manually tuned controller. This is attributed to the particular choice of the relative weights on the thermal comfort versus the setpoint deviations. Clearly, different cost metrics will yield different behavior and can be catered to the users' needs. Similar behavior was seen for the remaining zones, which also utilized the same cost function parameters. The numerical values of the learned control gains are summarized in Table III.

Next, we analyze the effect of Gaussian exploration on parameter convergence. In this study, we considered Gaussian signals of different variances to generate different datasets. Two datasets G1 and G2 (both for each zone) were obtained using zero mean Gaussian noise with standard deviations of 0.25 and 0.5 for Zone 1, and 1.25 and 1.5 for Zone 2. The pre-learning behavior under different Gaussian noise exploration is shown in Fig. 11 for Zone 1 and Fig. 14 for Zone 2. The post-learning response of the learned controller for Zone 1 under both Gaussian datasets G1 and G2 is comparable to the ones obtained using sinusoidal exploration as shown in Fig. 12. As can be seen in Fig. 15, Zone 2 exhibits a slightly higher oscillatory response under dataset G1, however, an improved response is observed under dataset G2. Compared with the parameter convergence results under the sinusoidal datasets S1 and S2, we find that convergence is also achieved under Gaussian datasets G1 and G2 as shown in Fig. 13 for Zone 1 and Fig. 16 for Zone 2. However, compared with the deviations under the sinusoidal datasets, we find the convergence of the iterations to be more sensitive to the standard deviation of the Gaussian noise. Convergence sometimes occurred in fewer iterations, but non-convergence was also encountered under higher noise (the XX entry in Table III refers to non-convergence).

A summary of the learned control gains is shown in Table III. Upon comparing the RL results with the manually tuned PI control (tuned to keep the overshoot small while ensuring fast convergence), we find that the RL results tend to have a larger

TABLE III
SUMMARY OF THE LEARNED CONTROL GAINS UNDER DIFFERENT SOLVERS, EXPLORATIONS, AND DATASETS

Solvers, explorations & datasets	Zone 1	Zone 2	Zone 3	Zone 4
Without Bias Compensation (S1)	[−14.5 −6.1]	[−13.8 −6.5]	[−5.2 −5.1]	[−4.5 −4.9]
Without Bias Compensation (S2)	[−16.8 −11.7]	[−17.0 −12.6]	[−16.6 −12.9]	[−15.9 −12.2]
Batch LS, Sinusoidal (S1)	[15.4 9.5]	[19.7 9.7]	[22.7 13.2]	[24.2 13.6]
Batch LS, Sinusoidal (S2)	[19.4 10.9]	[21.2 10.9]	[14.9 9.7]	[19.1 10.4]
RLS, Sinusoidal (S1)	[15.4 9.5]	[19.7 9.7]	[22.7 13.2]	[24.2 13.6]
RLS, Sinusoidal (S2)	[19.4 10.9]	[21.2 10.9]	[14.9 9.7]	[19.1 10.4]
RLS, Gaussian (G1)	[21.5 11.8]	[33.3 14.9]	[30.5 16.9]	[30.8 14.9]
RLS, Gaussian (G2)	[16.4 10.58]	[20.3 11.0]	[X X]	[10.8 8.4]

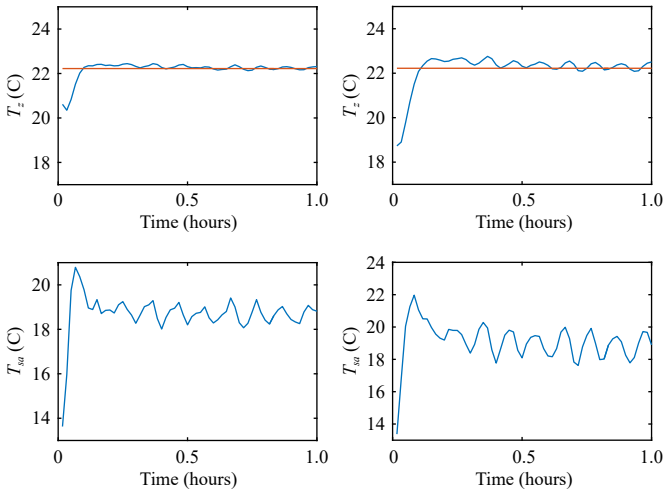

Fig. 5. Pre-learning phase of Zone 1 using dataset S1 (left) and S2 (right).

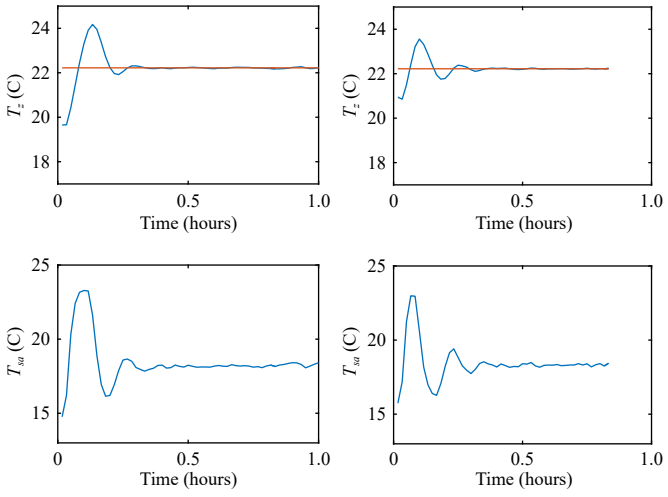

Fig. 6. Post-learning response under RLS rule for Zone 1 using datasets S1 (left) and S2 (right).

overshoot but a smaller settling time. An explanation for this is that the reward function was chosen to put more weight on the thermal tracking compared to the setpoint deviation. This results in higher gains, which correspond to a more aggressive response to minimize the error in a short amount of time.

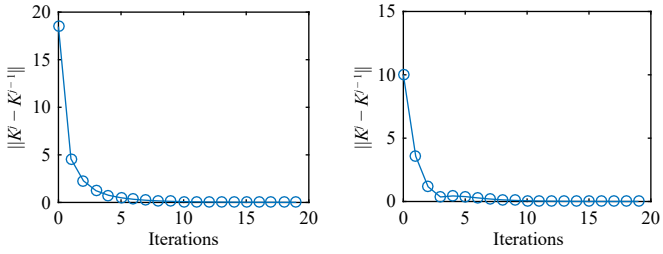

Fig. 7. Parameter convergence under RLS rule for Zone 1 using datasets S1 (left) and S2 (right).

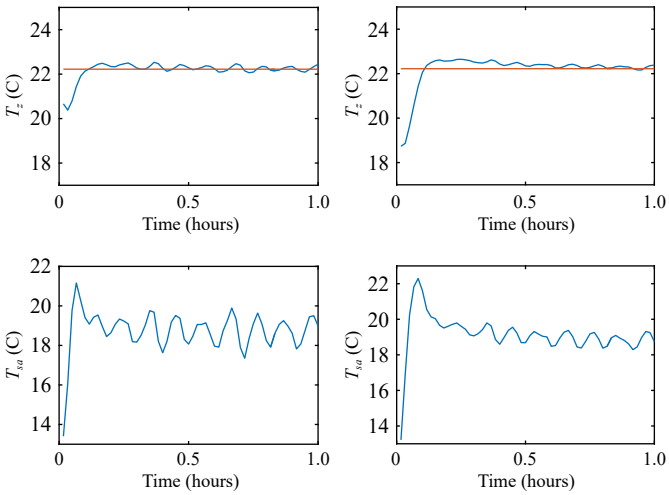

Fig. 8. Pre-learning phase of Zone 2 using datasets S1 (left) and S2 (right).

While the dimensions of all the zones are similar, we found that Zones 3 and 4 tend to have higher gains for the same cost function. This could be due to the details of the laboratory design and other sources of coupling. The type and amplitude of the exploration signals also impacts the gains because the RL algorithms rely heavily on the quality of the dataset and the nature of the exploration signal. Convergence was successful with a few exceptions. In particular, non-convergence occurred when Gaussian noise had a standard deviation of 2.5. When the exploration in the control input becomes more dominant, the system equipment may start to operate in constrained regions or exhibit nonlinear behavior, in which case

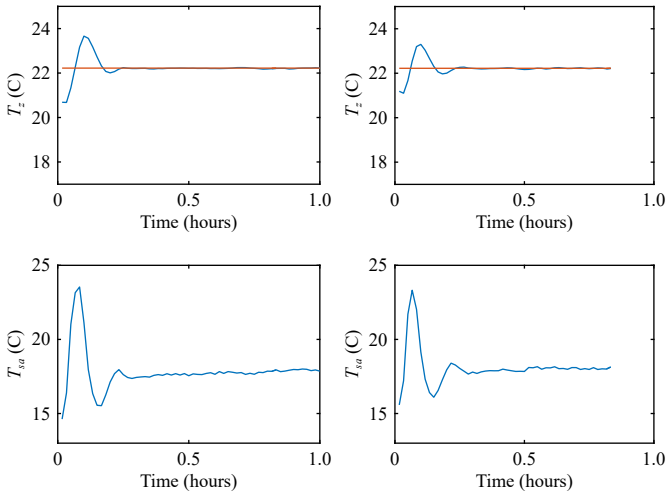

Fig. 9. Post-learning response under RLS rule for Zone 2 using datasets S1 (left) and S2 (right).

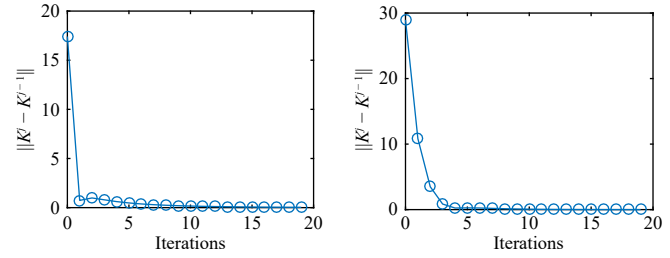

Fig. 10. Parameter convergence with RLS rule for Zone 2 using datasets S1 (left) and S2 (right).

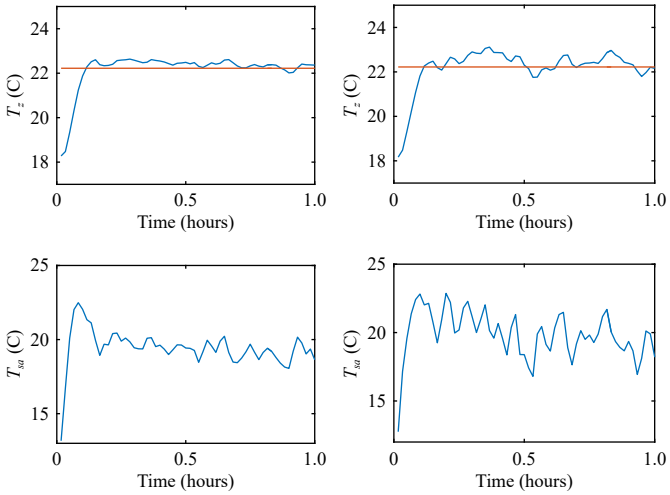

Fig. 11. Pre-learning phase of Zone 1 using dataset G1 (left) and G2 (right).

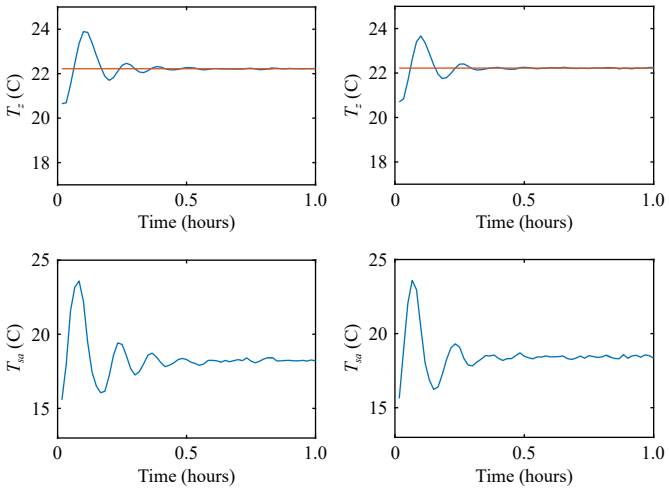

Fig. 12. Post-learning response under RLS rule for Zone 1 using dataset G1 (left) and G2 (right).

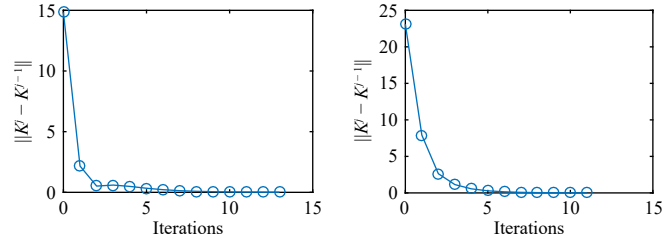

Fig. 13. Parameter convergence with RLS rule for Zone 1 using dataset G1 (left) and G2 (right).

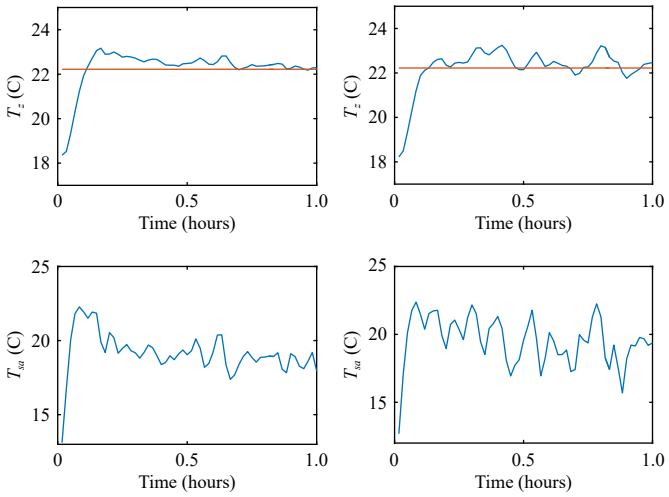

Fig. 14. Pre-learning phase of Zone 2 using dataset G1 (left) and G2 (right).

our prescribed control structure would not hold. Therefore, a careful choice of exploration is essential in RL algorithms, which is an application dependent consideration. In our experiments, we found that sinusoidal explorations resulted in less deviation in parameter estimates when the exploration magnitudes change. On the other hand, variations in Gaussian explorations sometimes lead to larger variations in parameter estimates. Non-convergence was also encountered in one of the zones when the deviation was high.

Table IV shows a summary of mean squared tracking error (MSE) under the nominal carefully tuned PI controller and the RL controller under different learning datasets (sinusoidal, Gaussian) for different zones. In general, the manual PI controller exhibits smaller MSE compared to the learning controller under different exploration conditions. This is because all RL controllers are designed with a high value of the weighting matrix Q, which penalizes based on the duration of

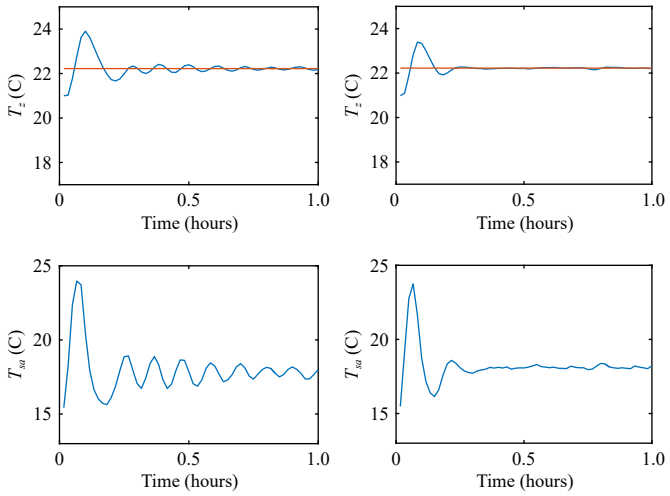

Fig. 15. Post-learning response under RLS rule for Zone 2 using dataset G1 (left) and G2 (right).

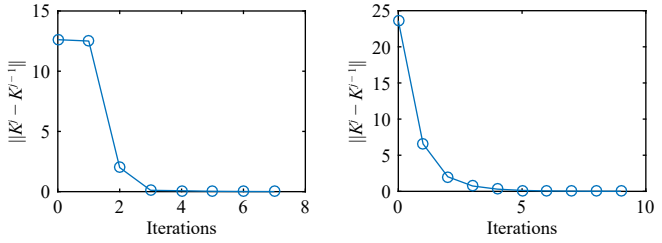

Fig. 16. Parameter convergence with RLS rule for Zone 2 using dataset G1 (left) and G2 (right).

TABLE IV
SUMMARY OF MEAN SQUARED TRACKING ERROR

Controller/Exploration	Zone 1	Zone 2	Zone 3	Zone 4
Baseline PI	0.1214	0.0829	0.1122	0.1032
Without Bias Compensation (S1)	Unstable	Unstable	Unstable	Unstable
Without Bias Compensation (S2)	Unstable	Unstable	Unstable	Unstable
RLS, Sinusoidal (S1)	0.4943	0.1930	0.3747	0.2624
RLS, Sinusoidal (S2)	0.1821	0.1176	0.1374	0.1696
RLS, Gaussian (G1)	0.2448	0.2108	–	–
RLS, Gaussian (G2)	0.1760	0.1150	–	–

transients, and as a result shorter transients are achieved at the expense of overshoots. Such overshoots, however, have a direct impact on the MSE. On the other hand, identical steady-state performance is achievable between the manual PI and automatic RL controller because once the learning transients are over the controller follows the PI structure.

VI. CONCLUSIONS

This paper presented the implementation of a structured reinforcement learning algorithm for the zone level control of a multi-zone HVAC system. A bias compensated Q-learning scheme was developed that takes into account the effect of unmeasurable disturbances. The presented Q-learning algorithm interacts with the HVAC zones instead of the simulated models and employs a quadratic reward function to take into account tracking error and setpoint deviations. By exploiting the control structure and reusing data from the past system interactions, the size of the training datasets and the amount of learning time are lesser owing to a well-defined quadratic parameterization in the learning equation and an *a priori* determined control structure. A comprehensive analysis of various exploration signals demonstrated that the learned control parameters, and consequently the post-learning performance, depend on the quality of the training dataset employed. What constitutes a quality dataset depends on how comprehensively the state and action space of the system are explored. However, sufficient conditions to meet these requirements in practice are currently not understood. In our study, we found that while the convergence of parameter estimates was almost always ensured, sinusoidal explorations resulted in lesser deviations in the final estimates compared to Gaussian explorations. Compared to the traditional PI, the RL algorithm offers the flexibility to tailor the reward function to a prescribed cost. The adaptation capability of the RL method allows us to change the cost requirements, which would cause the RL algorithm to relearn the control for a new specification.

REFERENCES

[1] J. Mei, Z. Y. Lu, J. H. Hu, and Y. L. Fan, "Energy-efficient optimal guaranteed cost intermittent-switch control of a direct expansion air conditioning system," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 11, pp. 1852–1866, Nov. 2020.

[2] N. E. Klepeis, W. C. Nelson, W. R. Ott, J. P. Robinson, A. M. Tsang, P. Switzer, J. V. Behar, S. C. Hern, and W. H. Engelmann, "The national human activity pattern survey (NHAPS): A resource for assessing exposure to environmental pollutants," *J. Exp. Sci. Environ. Epidemiol.*, vol. 11, no. 3, pp. 231–252, May–Jun. 2001.

[3] DOE, "An assessment of energy technologies and research opportunities," in *Proc. AIChE Annu. Meeting*, 2015.

[4] S. Prívara, J. Široký, L. Ferkl, and J. Cigler, "Model predictive control of a building heating system: The first experience," *Energy Build.*, vol. 43, no. 2-3, pp. 564–572, Feb.–Mar. 2011.

[5] Y. Y. Fu, S. C. Xu, Q. Zhu, and Z. O'Neill, "Containerized framework for building control performance comparisons: Model predictive control vs deep reinforcement learning control," in *Proc. 8th ACM Int. Conf. Systems for Energy-Efficient Buildings, Cities, and Transportation*, Coimbra, Portugal, 2021, pp. 276–280.

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

[7] M. M. Ha, D. Wang, and D. R. Liu, "Discounted iterative adaptive critic designs with novel stability analysis for tracking control," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1262–1272, Jul. 2022.

[8] G. Geng and G. M. Geary, "On performance and tuning of PID controllers in HVAC systems," in *Proc. IEEE Int. Conf. Control and Applications*, Vancouver, Canada, 1993, pp. 819–824.

[9] R. S. Sutton, A. G. Barto, and R. J. Williams, "Reinforcement learning is direct adaptive optimal control," *IEEE Control Syst. Mag.*, vol. 12, no. 2, pp. 19–22, Apr. 1992.

[10] M. J. Han, R. May, X. X. Zhang, X. R. Wang, S. Pan, D. Yan, and Y. Jin, "A novel reinforcement learning method for improving occupant

comfort via window opening and closing," *Sustain. Cities Soc.*, vol. 61, p. 102247, Oct. 2020.

[11] P. Fazenda, K. Veeramachaneni, P. Lima, and U.-M. O'Reilly, "Using reinforcement learning to optimize occupant comfort and energy usage in HVAC systems," *J. Ambient Intell. Smart Environ.*, vol. 6, no. 6, pp. 675–690, Jan. 2014.

[12] A. Overgaard, B. K. Nielsen, C. S. Kallesøe, and J. D. Bendtsen, "Reinforcement learning for mixing loop control with flow variable eligibility trace," in *Proc. IEEE Conf. Control Technology and Applications*, Hong Kong, China, 2019, pp. 1043–1048.

[13] T. S. Wei, Y. Z. Wang, and Q. Zhu, "Deep reinforcement learning for building HVAC control," in *Proc. 54th ACM/EDAC/IEEE Design Automation Conf.*, Austin, USA, 2017, pp. 1–6.

[14] E. Barrett and S. Linder, "Autonomous HVAC control, a reinforcement learning approach," in *Proc. Joint European Conf. Machine Learning and Knowledge Discovery in Databases*, Porto, Portugal, 2015, pp. 3–19.

[15] G. T. Costanzo, S. Iacovella, F. Ruelens, T. Leurs, and B. J. Claessens, "Experimental analysis of data-driven control for a building heating system," *Sustain. Energy Grids Netw.*, vol. 6, pp. 81–90, Jun. 2016.

[16] Z. Wang and T. Z. Hong, "Reinforcement learning for building controls: The opportunities and challenges," *Appl. Energy*, vol. 269, p. 115036, Jul. 2020.

[17] J. X. Zhang, K. W. Li, and Y. M. Li, "Output-feedback based simplified optimized backstepping control for strict-feedback systems with input and state constraints," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 6, pp. 1119–1132, Jun. 2021.

[18] Y. J. Chen, L. K. Norford, H. W. Samuelson, and A. Malkawi, "Optimal control of HVAC and window systems for natural ventilation through reinforcement learning," *Energy Build.*, vol. 169, pp. 195–205, Jun. 2018.

[19] G. P. Henze and R. H. Dodier, "Adaptive optimal control of a grid-independent photovoltaic system," *J. Sol. Energy Eng.*, vol. 125, no. 1, pp. 34–42, Feb. 2003.

[20] L. Yang, Z. Nagy, P. Goffin, and A. Schlueter, "Reinforcement learning for optimal control of low exergy buildings," *Appl. Energy*, vol. 156, pp. 577–586, Oct. 2015.

[21] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis, and G. S. Stavrakakis, "Reinforcement learning for energy conservation and comfort in buildings," *Build. Environ.*, vol. 42, no. 7, pp. 2686–2698, Jul. 2007.

[22] C. X. Guan, Y. Z. Wang, X. Lin, S. Nazarian, and M. Pedram, "Reinforcement learning-based control of residential energy storage systems for electric bill minimization," in *Proc. 12th Annu. IEEE Consumer Communications and Networking Conf.*, Las Vegas, USA, 2015, pp. 637–642.

[23] S. Y. Zhou, Z. J. Hu, W. Gu, M. Jiang, and X.-P. Zhang, "Artificial intelligence based smart energy community management: A reinforcement learning approach," *CSEE J. Power Energy Syst.*, vol. 5, no. 1, pp. 1–10, Mar. 2019.

[24] Y. Y. Ran and M. H. Jun, "Performance based thermal comfort control (PTCC) using deep reinforcement learning for space cooling," *Energy Build.*, vol. 203, p. 109420, Nov. 2019.

[25] B. Sun, P. B. Luh, Q.-S. Jia, and B. Yan, "Event-based optimization within the lagrangian relaxation framework for energy savings in HVAC systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 4, pp. 1396–1406, Oct. 2015.

[26] Y. Zhang and M. van der Schaar, "Structure-aware stochastic load management in smart grids," in *Proc. IEEE INFOCOM 2014-IEEE Conf. Computer Communications*, Toronto, Canada, pp. 2643–2651.

[27] B.-G. Kim, Y. Zhang, M. van der Schaar, and J.-W. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2187–2198, Sept. 2016.

[28] D. Bertsekas, "Multiagent reinforcement learning: Rollout and policy iteration," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 2, pp. 249–272, Feb. 2021.

[29] 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. Mag.*, vol. 37, no. 1, pp. 33–52, Feb. 2017.

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

[31] Y. J. Peng, Q. Chen, and W. J. 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.

[32] C. Chen, F. L. Lewis, K. Xie, S. L. Xie, and Y. L. Liu, "Off-policy learning for adaptive optimal output synchronization of heterogeneous multi-agent systems," *Automatica*, vol. 119, p. 109081, Sept. 2020.

[33] B. S. Lian, W. Q. Xue, F. L. Lewis, and T. Y. Chai, "Robust inverse *Q*-learning for continuous-time linear systems in adversarial environments," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13083–13095, Dec. 2022.

[34] B. S. Lian, W. Q. Xue, F. L. Lewis, and T. Y. Chai, "Inverse reinforcement learning for adversarial apprentice games," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021. DOI: 10.1109/TNNLS.2021.3114612

[35] 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.

[36] F. L. Lewis and V. L. Syrmos, *Optimal Control*. 2nd ed. Chichester, UK: John Wiley & Sons, 1995.

[37] S. A. A. Rizvi, A. J. Pertzborn, and Z. L. Lin, "Reinforcement learning based optimal tracking control under unmeasurable disturbances with application to HVAC systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7523–7533, Dec. 2022.

[38] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *Proc. American Control Conf.*, Baltimore, USA, 1994, pp. 3475–3479.

[39] Z. Chen and N. Li, "An optimal control-based distributed reinforcement learning framework for a class of non-convex objective functionals of the multi-agent network," *IEEE/CAA J. Autom. Sinica*, 2022. DOI: 10.1109/JAS.2022.105992

[40] G. Tao, *Adaptive Control Design and Analysis*. Hoboken, USA: Wiley-Interscience, 2003.

[41] C. H. Liu, F. Zhu, Q. Liu, and Y. C. Fu, "Hierarchical reinforcement learning with automatic sub-goal identification," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 10, pp. 1686–1696, Oct. 2021.

[42] IBAL Overview, National Institute of Standards and Technology. [Online]. Available: https://www.nist.gov/el/energy-and-environment-division-73200/intelligent-buildings-agents-project/ibal-overview

[43] A. J. Pertzborn, "Intelligent building agents laboratory: Hydronic system design," U.S. Department of Commerce, National Institute of Standards and Technology, 2016. [Online]. Available: https://doi.org/10.6028/NIST.TN.1933

[44] A. J. Pertzborn and D. A. Veronica, "Intelligent building agents laboratory: Air system design," U.S. Department of Commerce, National Institute of Standards and Technology, 2016. [Online]. Available: https://doi.org/10.6028/NIST.TN.2025

[45] A. J. Pertzborn and D. A. Veronica, "Baseline control systems in the intelligent building agents laboratory," U.S. Department of Commerce, National Institute of Standards and Technology, 2021. [Online]. Available: https://doi.org/10.6028/NIST.TN.2178

[46] Y. Y. Li, Y. J. Tang, R. Y. Zhang, and N. Li, "Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach," *IEEE Trans. Autom. Control*, vol. 67, no. 12, pp. 6429–6444, Dec. 2022.

Syed Ali Asad Rizvi (Member, IEEE) received the B.E. degree in industrial electronics from the Institute of Industrial Electronics Engineering, NED University of Engineering and Technology, Pakistan, in 2012, the M.S. degree in electrical engineering from the National University of Sciences and Technology, Pakistan, in 2014, and the Ph.D. degree in electrical engineering from the University of Virginia, USA, in 2020. He was a Postdoctoral Fellow at the National Institute of Standards and Technology (NIST), USA (2020–2021).

He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Tennessee Technological University, USA. His current research interests include optimal control, reinforcement learning, distributed learning and control, and their application in autonomous systems including autonomous vehicles, building HVAC systems, energy systems, and electric drives.

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

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

Zongli Lin (Fellow, IEEE) is the Ferman W. Perry Professor in the School of Engineering and Applied Science and a Professor of electrical and computer engineering at University of Virginia. He received the B.S. degree in mathematics and computer science from Xiamen University, in 1983, the master of engineering degree in automatic control from Chinese Academy of Space Technology, in 1989, and the Ph.D. degree in electrical and computer engineering from Washington State University, USA, in 1994. His current research interests include nonlinear control, robust control, and control applications.

He was an Associate Editor of *IEEE Transactions on Automatic Control* (2001–2003), *IEEE/ASME Transactions on Mechatronics* (2006–2009) and *IEEE Control Systems Magazine* (2005–2012). He was elected a Member of the Board of Governors of the IEEE Control Systems Society (2008–2010, 2019–2021) and chaired the IEEE Control Systems Society Technical Committee on Nonlinear Systems and Control (2013–2015). He has served on the operating committees of several conferences. He was the program chair of the 2018 American Control Conference and a general chair of the 13th and 16th International Symposium on Magnetic Bearings, held in 2012 and 2018, respectively. He currently serves on the editorial boards of several journals and book series, including *Automatica*, *Systems & Control Letters*, and Springer/Birkhauser book series Control Engineering. He is a Fellow of IEEE, CAA, International Federation of Automatic Control (IFAC), American Association for the Advancement of Science (AAAS).