

# Q-Learning Based LQR Occupant-Centric Control of Non-Residential Buildings

Oumaima Ait-Essi<sup>1</sup>, Joseph J. Yamé<sup>1,\*</sup> <sup>a</sup>, Hicham Jamouli<sup>2</sup> <sup>b</sup> and Frédéric Hamelin<sup>1</sup> <sup>c</sup>

<sup>1</sup>CRAN, CNRS, UMR 7039, Université de Lorraine, Campus Sciences, Vandoeuvre-lès-Nancy, France

<sup>2</sup>LISAD, ENSA, Université Ibn Zohr, Agadir, Morocco

{oumaima.ait-essi, joseph.yame, frederic.hamelin}@univ-lorraine.fr, h.jamouli@uiz.ac.ma


**Keywords:** Reinforcement Learning, Q-Learning, Data-Based Linear Quadratic Control, HVAC-VAV Systems, Building Occupants, Occupant-Centric Control.


**Abstract:** We propose a novel approach to the control of variable-air-volume (VAV)-HVAC systems for the regulation of thermal comfort in rooms of a non-residential building where the number of occupants may vary considerably and randomly during the day. Specifically, we develop a reinforcement learning control algorithm based on model-free optimal linear quadratic control. We leverage the quality function, the so-called  $Q$ -function, derived from Bellman dynamic programming, to develop a learning control algorithm based solely on system-generated data including building dynamics and its occupants. Simulations are carried out on a new HVAC-VAV system installed in a building at the University of Lorraine, demonstrating the potential of the proposed method for maintaining climatic conditions and the comfort of room occupants while optimizing the airflow demand of VAV boxes, which is correlated with the energy consumed per room.


## 1 INTRODUCTION

Energy consumption in buildings accounts for over 36.9% of primary energy consumption, of which 17.2% is accounted for by commercial and non-residential buildings (EIA, 2024). Heating, ventilation, and air conditioning (HVAC) systems account for 40% of total building energy consumption. Ensuring occupant comfort while achieving energy savings is a key objective in optimal building operation, as comfort plays a vital role in human well-being and productivity. Recent contributions regarding human-building interactions highlight the impact of occupant information, such as occupancy and behavior, on building energy consumption. The potential for improving the operation of buildings and their control systems through such human-building interactions is now well recognized, and has led to occupant-centric control (OCC) as an important research topic (Soleimanijavid et al., 2024; Ouf et al., 2021; Yu et al., 2024; Xu et al., 2023; Jia et al., 2017). Al-

though the concept of OCC is not perfectly defined, it can be categorized in two main ways (Ouf et al., 2021) : occupant-centric controls and occupant behavior-centric controls. In the first meaning, OCC deals with the presence/absence of occupants and HVAC control based on occupants counts while in its second meaning OCC focuses on occupant behaviors and preferences from occupant's interactions with building systems, e.g., thermostats setpoints adjusting, windows openings, exercising, etc. With regards to the behavioral aspects of OCC, information and potential characteristics can be extracted from energy consumption data using machine learning methods which are subsequently used to identify and classify typical behavior patterns (Yu et al., 2024). However, occupant behaviors are highly stochastic and unpredictable, with temporal complexities in the process of identifying consistent behavioral strategies (Xu et al., 2023). To meet these challenges, reinforcement learning is increasingly becoming one of the most effective ways of developing control strategies that take occupant behavior into account to ensure thermal comfort and optimize building energy consumption (Han et al., 2020), (Liu and Gou, 2024), (Wang et al., 2023). In the work presented here, OCC is addressed under its first categorization as the presence/absence and the number of occupants in a building have a direct

<sup>a</sup>  <https://orcid.org/0000-0002-4349-6240>

<sup>b</sup>  <https://orcid.org/0000-0002-9064-0372>

<sup>c</sup>  <https://orcid.org/0000-0002-5535-5680>

\*Corresponding author: joseph.yame@univ-lorraine.fr.

impact on its energy consumption, and strongly influence the operation of energy systems throughout the building. Buildings with highly variable occupancy profiles, such as classrooms, computer rooms or university laboratories, and with occupancy varying throughout the day all along the year, raise issues for the design and implementation of control strategies aimed at balancing comfort and energy efficiency. Among these issues, the lack of an accurate model of building dynamics is a barrier to the design of optimal controllers to ensure thermal comfort and optimum energy consumption. Optimal controllers, such as the linear quadratic controller (LQR), are frequently used for systems with known dynamics. In the case of unknown dynamics, a basic approach is to fit a model to the system using input/output observations and then use the fitted model for control purposes. A more direct approach would be to design the optimal controller directly from observations, without the intermediate step of a fitted plant model. Such an approach, driven by advances in machine learning, is currently undergoing significant growth, with LQR control being considered as a standard benchmark for learning-based control of systems with unknown dynamics (Farjadnasab and Babazadeh, 2022). It is in this context that this research project was carried out on thermal comfort and the reduction of energy consumption in a university building where classroom occupancy varies greatly from one day to the next.

## 2 SYSTEM DESCRIPTION AND MODELING

### 2.1 System Description

The system under consideration is an HVAC installation recently commissioned in a central building, called ATELA building, comprising practical electrical, electronics and control engineering laboratories with lectures and computer rooms at the Faculty of Science and Technology of the University of Lorraine in Nancy, France.

The air-conditioning system is a typical VAV-based HVAC system for a multizone building as depicted in figure 1

Each zone is equipped with a VAV terminal box which receives conditioned air from a central air handling unit (AHU) at a constant temperature, called the supply air-temperature. The AHU consists mainly of a thermal wheel, i.e., a rotary heat exchanger, heating and cooling coils and a supply fan. The thermal wheel recovers heat from the exhaust air and transfers it to the fresh air stream coming from outside.

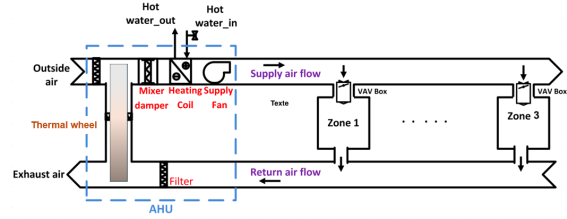


Figure 1: Multizone VAV-based HVAC system.

The heating and cooling coils are water to air heat exchangers which control the temperature of supply air flow by varying hot or chilled water flow of constant temperature supplied by a district heating network or a local chiller for the chilled water. Each VAV box has a damper which regulates the volume of conditioned air delivered to the box according to the thermal needs of the zone. As all VAV boxes regulate their air volume independently, the total air volume delivered by the AHU varies according to the demands of the VAV boxes. Consequently, the fan speed is controlled to meet the overall demand of all zones. Figure 2 shows the closed-loop structure for controlling thermal conditions in each zone. Note that VAV boxes are equipped with embedded electronic air-mass flow controllers. The temperature controller tracks the zone temperature setpoint  $T_{z,sp}$  by providing the required air mass flow rate setpoint  $u = \dot{m}$  to the VAV box depending on the zone temperature measurement  $T_z$ . The VAV's embedded flow controller then modulates the VAV-damper opening, through signal  $u_d$  to deliver the required air-mass flow rate  $\dot{m}_f$  to the zone.

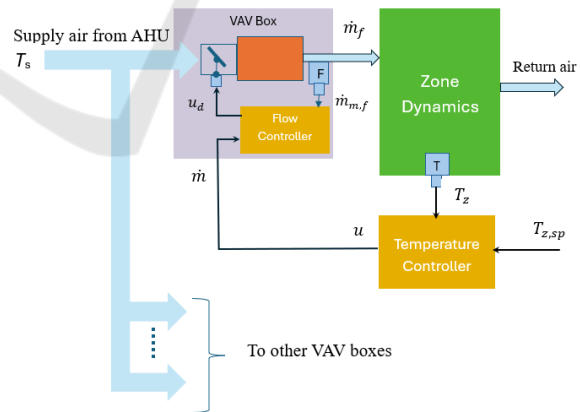


Figure 2: Closed-loop structure at each zone.

### 2.2 Thermal System Modeling

To gain the understanding needed to formulate the data-driven control problem in section 4.1 within the reinforcement learning framework, we now establish a basic model of zone thermal dynamics, in winter

season, based on the principles of heat transfer. Considering a zone as a thermodynamic control volume (Amende et al., 2021; Bergman and Lavine, 2017), the principle of energy conservation is simply expressed by the differential equation

$$\rho c_p v_i \frac{dT_{z_i}}{dt} = \dot{m}_i c_p (T_s - T_{z_i}) - \dot{Q}_{load,i} \quad (1)$$

In this equation,  $\rho$  and  $c_p$  are the air density and the specific heat of air at constant pressure, respectively. The supply air temperature to the zone is constant and equal to  $T_s$ , and all other variables indexed by  $i = 1, \dots, N_b$ , where  $N_b$  is the number of zones, are zone-specific. These variables are the zone volume  $v_i$ , the zone temperature  $T_{z_i}$  and the zone sensible heating load  $\dot{Q}_{load,i}$ . The sensible heating load is the net amount of energy that must be supplied to the zone to maintain a specified thermal state for the zone, and this is nothing more than the net sum of heat losses. The first term of the right-hand side of (1) is the energy supplied by the AHU to the zone to heat it and this is expressed in terms of the mass flowrate, specific heat and temperature difference between the supply air and the zone. In this work, we assume that heat losses are mainly to the environment at ambient temperature  $T_{oa}$  and to the adjacent zones whilst heat gains are mainly due to occupancy and solar gains through the glazings. Therefore, the net sensible heating load of zone  $i$  reads as

$$\dot{Q}_{load,i} = UA_i(T_{z_i} - T_{oa}) + UA_{ij}(T_{z_i} - T_{z_j}) - \eta_i \phi_h - \dot{q}_{sol,i} \quad (2)$$

where  $T_{z_{ij}}$  is the temperature of zones  $j$  adjacent to zone  $i$ ,  $\phi_h$  is the average rate of heat flow from a human (ASHRAE, 2021, Chapter 9),  $\eta_i$  the number of occupants of the zone,  $A_i$  the total area of the envelope of the zone surrounded by the environment,  $A_{ij}$  the area of the wall of zone  $i$  adjacent to zones  $j$ ,  $U$  the overall U-value of the zone,  $T_{oa}$  the outside air temperature, and  $\dot{q}_{sol,i}$  the solar gain of the zone. Let  $C_i = \rho c_p v_i$  be the thermal capacitance of zone  $i$  and define the following constants

$\alpha_i = UA_i/C_i$	$\alpha_{ij} = (U_{ij}A_{ij})/C_i, j \neq i$
$\alpha_{ii} = -\left(\alpha_i + \sum_{j \neq i}^{N_b} \alpha_{ij}\right)$	$\gamma_i = c_p/C_i$
$\beta_i = \gamma_i T_s$	$\phi_i = \phi_h/C_i$

Construct the  $N_b$ -dimensional column vectors  $\underline{\alpha}$ ,  $\underline{\beta}$ ,  $\underline{\gamma}$  whose elements are respectively the  $\alpha'_i$ 's,  $\beta'_i$ 's,  $\gamma'_i$ 's,  $i = 1, \dots, N_b$ . Now, introduce the following matrices

$$\mathbb{A} = [\alpha_{ij}]_{i,j=1}^{N_b}, \quad \mathbb{B}(T_z) = \text{diag}\left(\underline{\beta} - \underline{\gamma} \odot T_z\right)$$

where  $T_z$  is the  $N_b$ -dimensional vector of zone temperatures and the symbol  $\odot$  denotes the Hadamard product. Matrix  $\mathbb{B}(T_z)$  is the diagonal matrix whose

diagonal elements are the components of vector  $\underline{\beta} - \underline{\gamma} \odot T_z$ . The combination of (1) and (2) written for the  $N_b$  zones results into the following state-space equation for the multizone building dynamics

$$\dot{T}_z = \mathbb{A}T_z + \mathbb{B}(T_z)\dot{m} + \Phi\underline{\eta} + \underline{\alpha}T_{oa} + \mathbb{E}\dot{q}_{sol} \quad (3)$$

where  $\Phi$  is the diagonal matrix whose diagonal elements are the  $\phi'_i$ 's and  $\mathbb{E}$  is the thermal elastance matrix which is the inverse of the capacitance matrix  $\text{diag}(C_1, \dots, C_{N_b})$ . Clearly, the dynamics is nonlinear but appears as a linear-type dynamic system with a state-dependent input matrix which is characteristic of bilinear systems.

### 3 RECAP ON THE STANDARD DISCRETE TIME LINEAR QUADRATIC REGULATOR AND THE Q-FUNCTION

The linear quadratic control problem is the following constrained optimization problem

$$\begin{aligned} \min_{u_k} V(x_k) &= \sum_{j=k}^{\infty} \ell(x_j, u_j) \\ \text{s.t.} \quad x_{k+1} &= A_d x_k + B_d u_k \end{aligned} \quad (4)$$

where the constraint is a linear dynamics with  $x_k \in \mathbb{R}^n$  and  $u_k \in \mathbb{R}^m$  its state and control input at time step  $k \geq 0$  and  $A_d \in \mathbb{R}^{n \times n}$  and  $B_d \in \mathbb{R}^{n \times m}$  being the state transition matrix and the control input matrix, respectively, of that system. The objective function of the optimization problem is a long-run cost in which the stage cost at time step  $k$ ,  $\ell(x_k, u_k)$  is given by the quadratic form

$$\ell(x_k, u_k) = x_k^T Q_d x_k + u_k^T R_d u_k \quad (5)$$

in which the parameters  $Q_d \in \mathbb{R}^{n \times n}$  and  $R_d \in \mathbb{R}^{m \times m}$  are weighting matrices that are chosen symmetric and positive semidefinite and positive definite, i.e.,  $Q_d \geq 0$  and  $R_d > 0$ . The cost in (4) can be written in a recursive form, known as the Bellman equation

$$V(x_k) = \ell(x_k, u_k) + V(x_{k+1}) \quad (6)$$

from which, the Bellman optimality principle states that the optimal value function  $V^{\text{opt}}(x_k)$  satisfies the following relationship called the *Bellman optimality equation*

$$V^{\text{opt}}(x_k) = \min_{u_k} (\ell(x_k, u_k) + V^{\text{opt}}(x_{k+1})) \quad (7)$$

For the optimization problem (4), solving the Bellman optimality equation (7) amounts to solving a discrete-time algebraic Riccati equation (DARE)

$$A_d^T P A_d - P + Q_d - A_d^T P B_d (R_d + B_d^T P B_d)^{-1} B_d^T P A_d = 0 \quad (8)$$

whose unique positive definite solution,  $P^{\text{opt}} > 0$ , yields the optimal value function

$$V^{\text{opt}}(x_k) = x_k^T P^{\text{opt}} x_k \quad (9)$$

and the corresponding optimal control policy

$$u_k^{\text{opt}} = -K^{\text{opt}} x_k \quad (10)$$

with the gain  $K^{\text{opt}} = (R_d + B_d^T P^{\text{opt}} B_d)^{-1} B_d^T P^{\text{opt}} A_d$ . At this point, it is interesting to see that the minimizing argument in Bellman's optimality equation (7), i.e.,  $u_k^{\text{opt}}$  written as a function of  $x_k$ ,  $u_k^{\text{opt}} = \pi(x_k)$ , is

$$\pi(x_k) = \underset{u_k}{\operatorname{argmin}} Q^\pi(x_k, u_k) \quad (11)$$

where  $Q^\pi(x_k, u_k)$ , called the  $Q$ -function, is given by

$$Q^\pi(x_k, u_k) = \ell(x_k, u_k) + V^\pi(x_{k+1}) \quad (12)$$

Now, given the optimal value function  $V^\pi$  in (9), it is immediately seen that the optimal  $K^{\text{opt}}$  can be computed through the  $Q$ -function (12) which is actually the cost of executing an arbitrary control  $u_k$  at time  $k$ , and then following the optimal policy  $\pi$  from time  $k+1$  to all future times. The optimal control  $u_k^{\text{opt}}$  which minimizes the cost is obtained by solving the first-order necessary optimality condition (FONC)  $\nabla_{u_k} Q^\pi(x_k, u_k) = 0$ , where  $\nabla_{u_k}$  stands for the gradient w.r.t  $u_k$ . For the LQR policy, straightforward calculations using the dynamics system in (4) and (9) show that the  $Q$ -function (12) is a quadratic form in  $z_k = \operatorname{col}(x_k, u_k)$  denoting the column vector obtained by stacking vectors  $x_k$  and  $u_k$ ,

$$Q^\pi(x_k, u_k) = z_k^T \mathbb{Q} z_k \quad (13)$$

and

$$\mathbb{Q} = \begin{bmatrix} A_d^T P^\pi A_d + Q_d & A_d^T P^\pi B_d \\ B_d^T P^\pi A_d & B_d^T P^\pi B_d + R_d \end{bmatrix} \quad (14)$$

with  $P^\pi$  being the unique positive definite solution of the DARE (8). The FONC yields consequently the optimal control (10).

## 4 Q-LEARNING BASED MODEL-FREE LQR SYNTHESIS

### 4.1 Problem Formulation

The main problem addressed in this paper can be formulated as follows:

**Problem.** Consider the multizone building system (3) with unknown parameters. Given discrete-time measurements of the zone temperatures and the

air mass flows, design an optimal controller that will maintain setpoint temperatures independent of occupancy while minimizing the energy demand of each zone.

As the building dynamics is unknown and strongly influenced by the high variability of occupancy during the day, the objective is to learn the optimal controller directly on the basis of time series from the system and reinforcement learning techniques. Let's formalize the problem by setting  $x(t) = \underline{T}_z(t)$  and  $u(t) = \dot{m}(t)$  in (3) for notation convenience. The building normally operates around an operating point for thermal comfort, and it can be assumed that it has linear dynamics in the region around the set point temperature, say  $\underline{T}_r$ , so that the unknown dynamics (3) reads

$$\begin{aligned} \dot{x} &= Ax + Bu + \Phi \underline{\eta} + \underline{\alpha} T_{oa} + \mathbb{E} \dot{q}_{sol} \\ y &= x \end{aligned} \quad (15)$$

where  $A = \mathbb{A}$ ,  $B$  is the Jacobian matrix of  $\mathbb{B}(x)$  at the operating point, and  $y$  is the output of the system. A key signal for comfort monitoring is the real-time deviation between setpoint and temperature measurement. This signal is available in the building automation system controlling the HVAC plant, and is given by

$$e = \underline{T}_r - y \quad (16)$$

To define the environment to be controlled within the reinforcement learning framework, gather all signals that exist outside the controller, the so-called agent, to be designed. The environment will therefore be everything that exists outside the agent, and constitutes a system with a boundary through which the agent sends actions and receives observations and penalties. The setpoints  $\underline{T}_r$  and the occupancy profile  $\underline{\eta}(t)$ ,  $t \geq 0$ , determined by the number of occupants, are seen as provided by the environment and generated by a signal generator given by

$$\begin{bmatrix} \dot{\underline{T}}_r \\ \dot{\underline{\eta}} \end{bmatrix} = \begin{bmatrix} A_r & 0 \\ 0 & A_\eta \end{bmatrix} \begin{bmatrix} \underline{T}_r \\ \underline{\eta} \end{bmatrix} \quad (17)$$

where matrices  $A_r$  and  $A_\eta$  are non-Hurwitz matrices.

The problem stated above can be reformulated as having the objective of designing a model-free optimal feedback control  $u(t)$  such that

- (i) the closed-loop system is stable
- (ii) for all initial conditions of the signal generator and the building states,  $e(t) \rightarrow 0$  as  $t \rightarrow \infty$
- (iii) properties (i) and (ii) are robust to variations in building dynamics

## 4.2 Method

Towards solving the problem, we select a structure based on the internal model principle (IMP) (Francis and Wonham, 1976; Davison and Goldenberg, 1975), which states that a necessary condition for achieving the above objectives is that the open loop system contains the modes of the dynamic structure of the setpoint  $\underline{T}_r(t)$  and the occupancy profile  $\underline{\eta}(t)$ . Such modes are to be embedded in a filter driven by the error signal (16), and this will enable us to set the environment to be used for reinforcement learning control.

Let  $\delta_r(s)$  and  $\delta_\eta(s)$  be the minimal polynomials of  $A_r$  and  $A_\eta$ , and let  $\delta(s)$  be the least common multiple of  $\delta_r(s)$  and  $\delta_\eta(s)$  given by

$$\delta(s) = s^p + \lambda_{p-1}s^{p-1} + \dots + \lambda_1s + \lambda_0 \quad (18)$$

Define the  $pN_b$ -dimensional IMP-filter by

$$\dot{\xi} = \Lambda\xi + \Gamma e \quad (19)$$

where  $\Lambda$  and  $\Gamma$  are block diagonal matrices comprising  $N_b$  blocks given by

$$\Lambda = \text{block diag} \left[ \underbrace{\Lambda^* \quad \Lambda^* \quad \dots}_{N_b\text{-times}} \right]$$

$$\Gamma = \text{block diag} \left[ \underbrace{\Gamma^* \quad \Gamma^* \quad \dots}_{N_b\text{-times}} \right]$$

with  $\Lambda^*$  a  $p \times p$  matrix and  $\Gamma^*$  a  $p$ -dimensional vector given by

$$\Lambda^* = \left[ \begin{array}{c|c} 0_{p-1,1} & I_{p-1} \\ \hline -\lambda_0 & -\lambda_1 \dots -\lambda_{p-1} \end{array} \right]$$

$$\Gamma^* = [0 \quad 0 \quad \dots \quad 1]^T$$

where  $0_{m,n}$  and  $I_n$  denotes the zero matrix of size  $(m \times n)$  and the identity matrix of dimension  $n$ , respectively. The error-driven IMP-filter is used to augment the original system as shown in figure 3 to create the environment of the problem to be solved. Then, the augmented plant has the following composite dynamics

$$\begin{bmatrix} \dot{x} \\ \dot{\xi} \end{bmatrix} = \begin{bmatrix} A & 0 \\ -\Gamma & \Lambda \end{bmatrix} \begin{bmatrix} x \\ \xi \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u \quad (20)$$

$$+ \begin{bmatrix} 0 & \Phi \\ \Gamma & 0 \end{bmatrix} \begin{bmatrix} \underline{T}_r \\ \underline{\eta} \end{bmatrix} + \begin{bmatrix} \alpha & \mathbb{E} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} T_{oa} \\ \dot{q}_{sol} \end{bmatrix}$$

which is written compactly as

$$\dot{\zeta} = \mathcal{A}\zeta + \mathcal{B}u + \mathcal{D}\omega + \mathcal{E}d \quad (21)$$

where  $\zeta = \text{col}(x, \xi)$ ,  $\omega = \text{col}(\underline{T}_r, \underline{\eta})$ ,  $d = \text{col}(T_{oa}, \dot{q}_{sol})$  and matrices  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{D}$ , and  $\mathcal{E}$

are self-explanatory with regards to equation (20). The following theorem, adapted from (Davison and Goldenberg, 1975), solves the problem stated when a model of the system is available

**Theorem 1.** (Davison's Theorem) *If the composite dynamics (20) is controllable, Then*

- (i) it can be stabilized by a state feedback control  $u = -K\zeta$
- (ii) for such stabilizing state feedback,  $\lim_{t \rightarrow \infty} e(t) = 0$  for any initial conditions of  $x$ ,  $\xi$ ,  $\underline{T}_r$ , and  $\underline{\eta}$
- (iii) properties (ii) is robust for any variations in the building parameters  $(A, B)$ , the controller parameter  $K$  and the IMP-filter parameter  $\Gamma$  provided the closed-loop is stable.

## 4.3 Results

The setup of the problem is shown in fig. 3

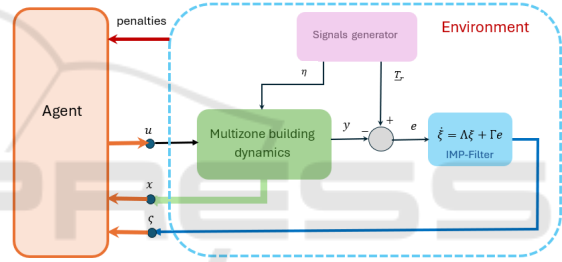


Figure 3: Reinforcement learning control structure.

where the agent, i.e. the controller, is to be designed based on the states  $\zeta_k = \text{col}(x_k, \xi_k)$  of the unknown environment. The optimization problem to be solved reads

$$\min_{u_k} \sum_{j=k}^{\infty} \zeta_j^T Q_d \zeta_j + u_k^T R_d u_k \quad (22)$$

under Unknown environment dynamics

Referring to the  $Q$ -function defined for discrete-time LQR in section 3, it can be written here as

$$Q^\pi(\zeta_k, u_k) = \begin{bmatrix} \zeta_k \\ u_k \end{bmatrix}^T \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} \zeta_k \\ u_k \end{bmatrix} \quad (23)$$

and the FONC  $\nabla_{u_k} Q^\pi(x_k, u_k) = 0$  yields the optimal control

$$u_k^{\text{opt}} = \pi(\zeta_k) = -Q_{22}^{-1} Q_{21} \zeta_k \quad (24)$$

Thus, if the kernel matrix  $Q$  were known, the optimal control could be computed without resorting to the dynamical equations of the environment. To achieve such a model-free system control design, assume that there exists a kernel  $\hat{Q}$  that approximates the kernel  $Q$  of the unknown environment, so that  $Q^\pi(\zeta_k, u_k) \approx z_k^T \hat{Q} z_k$  and  $Q^\pi(\zeta_{k+1}, \pi(\zeta_{k+1})) \approx \tilde{z}_{k+1}^T \hat{Q} \tilde{z}_{k+1}$  with  $z_k =$

$\text{col}(\zeta_k, u_k)$  and  $\tilde{z}_{k+1} = \text{col}(\zeta_{k+1}, \pi(\zeta_{k+1}))$ . From (12), the following approximate equality holds

$$(z_k - \tilde{z}_{k+1})^\top \hat{\mathbb{Q}}(z_k - \tilde{z}_{k+1}) - \ell(\zeta_k, u_k) \approx 0 \quad (25)$$

The left-hand side of the above approximate equation is linear in  $\hat{\mathbb{Q}}$ , and setting it as a residual  $\varepsilon_k$  and using the handy property of vectorization (Graham, 2018) leads to the error equation

$$\Psi_k^\top \text{vec}(\hat{\mathbb{Q}}) - \ell(\zeta_k, u_k) = \varepsilon_k \quad (26)$$

where  $\Psi_k = (z_k - \tilde{z}_{k+1}) \otimes (z_k - \tilde{z}_{k+1})$ ,  $\otimes$  being the Kronecker product, and  $\text{vec}(\hat{\mathbb{Q}})$ , is the column vector of dimension  $(N_b(p+2))^2$  obtained by stacking the columns of the matrix  $\hat{\mathbb{Q}}$  on top of one another. From (26), it is seen that matrix  $\hat{\mathbb{Q}}$  can be estimated by minimizing the sum of squares of  $N$  residuals,

$$\min_{\hat{\mathbb{Q}}} \sum_{j=k}^{k+N-1} \varepsilon_j^2 = \min_{\hat{\mathbb{Q}}} \|\Psi \text{vec}(\hat{\mathbb{Q}}) - \underline{\ell}\|_2^2 \quad (27)$$

with data matrices,

$$\Psi = \begin{bmatrix} -\Psi_j^\top - \\ -\Psi_{j+1}^\top - \\ \vdots \\ -\Psi_{j+N-1}^\top - \end{bmatrix}, \quad \underline{\ell} = \begin{bmatrix} \ell_j \\ \ell_{j+1} \\ \vdots \\ \ell_{j+N-1} \end{bmatrix} \quad (28)$$

Note that the size of matrix  $\Psi$  is  $(N \times n_\Psi)$  with  $n_\Psi = (N_b(p+2))^2$ . For future reference, the conditions under which problem (27) has a unique solution is stated below as a lemma.

**Lemma 2.** *The solution of problem (27) is unique solution if and only if matrix  $\Psi$  is of full column-rank, and this requires at least that  $N \geq n_\Psi$ .*

An important practical issue here is that, due to the IMP-filter, the environment is not stable and consequently the time series  $\{\zeta_k, u_k\}$ ,  $k = 0, 1, \dots$ , generated in an open-loop experimental setting, are unbounded. Therefore, the agent cannot be trained with open-loop data. However, this problem can be circumvented in a closed-loop experimental setting with a stabilizing controller, i.e.,  $u_k = -K_s \zeta_k$  with  $K_s$  being a stabilizing feedback gain. Unfortunately, these closed-loop data are such that  $\zeta_k$  and  $u_k$  are highly dependent on each other due to feedback, and this is likely to prevent a large part of the state space of the environment to be explored. A common technique for relieving this problem is to add a dither (noise) to the feedback control signal to promote exploration of the state space, i.e.,  $u_k = -K_s \zeta_k + \tilde{n}_k$ , with  $\tilde{n}_k$  the additive noise. This is related to persistently exciting inputs in system identification and dual control in stochastic control (Yamé, 1987). A key observation here is that,

despite the additive noise to enforce non-collinearity between the columns of matrix  $\Psi$ , this matrix has a particular rigid structure in this respect revealed by the following proposition.

**Proposition 3.** *Let matrix  $\Psi$  be given by (28) with  $N \geq n_\Psi$ . Then,  $\Psi$  is rank-deficient, i.e.,  $\text{rank}(\Psi) = r_\Psi < n_\Psi$ .*

*Proof.* The proof is omitted for lack of space, see (Yamé, 2024).  $\square$

The rank deficiency of matrix  $\Psi$  makes problem (27) ill-posed, as it does not satisfy Hadamard's criteria for the existence of a solution, the uniqueness of this solution and its continuous dependence on the input data (Vogel, 2002). This ill-posedness is tackled here by augmenting matrix  $\Psi$  with a scalar matrix  $\sqrt{\lambda} I_{n_\Psi}$ ,  $\lambda \geq 0$ , leading to the optimization problem

$$\min_{\hat{\mathbb{Q}}} \left\| \begin{bmatrix} \Psi \\ \sqrt{\lambda} I_{n_\Psi} \end{bmatrix} \text{vec}(\hat{\mathbb{Q}}) - \begin{bmatrix} \underline{\ell} \\ 0_{n_\Psi, 1} \end{bmatrix} \right\|_2^2 \quad (29)$$

Note that the rank of the augmented " $\Psi$ -matrix" is equal to  $n_\Psi$ , and from lemma 2, the minimization problem (29) admits a unique solution given analytically by

$$\text{vec}(\hat{\mathbb{Q}}) = (\Psi_\lambda^\top \Psi_\lambda)^{-1} \Psi_\lambda^\top \underline{\ell}_\lambda \quad (30)$$

with  $\Psi_\lambda$  and  $\underline{\ell}_\lambda$  being the augmented  $\Psi$ -matrix and augmented  $\underline{\ell}$ -vector, respectively, in (29). This solution is also known as Tikhonov's regularized solution to the original problem (27) and  $\lambda$  is the regularized parameter (Vogel, 2002). The problem as formulated in (29) is not merely a bulwark against the rank deficiency of the original problem's data matrix  $\Psi$ , but it allows also to deal with the noise in this data matrix (El Ghaoui and Lebret, 1997), resulting from the noise injection into the control signal as expounded above to facilitate exploration of the state space of the environment. The solution (30) is implemented in a batch-processed form through the following algorithm

## 5 SIMULATION STUDIES

The building automation system (BAS) of the ATELA classrooms has been commissioned using simple PID controllers tuned by trial and error. The main challenge was to tune the zone temperature controllers that controls the VAV damper openings via the set-point of the mass airflow controllers, see fig. 2. It should be noted that the mass airflow controllers, which are internal to the VAV boxes, were factory-set. To improve thermal comfort and minimize the

energy demand of each zone, given the high variability of their occupancy and the number of occupants during the day, the model-free  $Q$ -learning based LQR control developed in the previous section was implemented. In this experimental simulation study, which served as a first test for the implementation of the  $Q$ -learning control algorithm in the BAS, only zone 1 was concerned, see fig. 1. For this zone, a database of time series in closed-loop has been generated consisting of  $\tilde{N} = 1000$  sample points, see fig 4. The room temperature setpoint was  $T_r = 22^\circ\text{C}$  and noise was added to the control input signal to promote exploration of the system dynamics. These time series can also be generated by a building simulator.

---

Algorithm 1:  $Q$ -Learning based model-free LQR.

---

- 1: Build a database of times series  $\{\zeta_0, u_0, \zeta_1, u_1, \dots, \zeta_{\tilde{N}}, u_{\tilde{N}}\}$  obtained in a stable closed-loop with noise added to the control inputs  $u_k, k = 0, \dots, \tilde{N} - 1, \tilde{N}$  sufficiently large
  - 2: Take a sub-series of length  $N \geq n_\Psi$  from step 1.
  - 3: Choose a regularization parameter  $\lambda > 0$
  - 4: **Initialization:** Select any stabilizing policy  $\pi_t(\zeta)$
  - 5: **repeat**
  - 6:     **for**  $k = 1, 2, \dots, N$  **do**
  - 7:         Construct :  $z_k = \text{col}(\zeta_k, u_k)$
  - 8:         Compute :  $\tilde{z}_{k+1} = \text{col}(\zeta_{k+1}, \pi_t(\zeta_k))$
  - 9:         Compute :  $\Psi_k = (z_k - \tilde{z}_{k+1}) \otimes (z_k - \tilde{z}_{k+1})$
  - 10:         Compute :  $\ell_k = z_k^T \text{diag}(Q_d, R_d) z_k$
  - 11:     **end for**
  - 12: Build data matrices  $\Psi$  and  $\underline{\ell}$  from (28)
  - 13: Solve (29) for  $\hat{Q}$
  - 14: Update policy to  $\pi_{t+1}(\zeta)$  from (24)
  - 15:  $\pi_t \leftarrow \pi_{t+1}$
  - 16: **until** Convergence of the policy, and set the optimal policy as  $\pi^{\text{opt}}(\zeta) = \pi_t(\zeta)$
- 

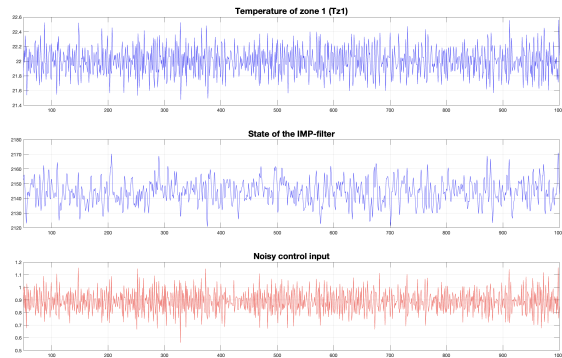


Figure 4: Closed-loop time-series for the  $Q$ -learning process for zone 1.

To apply the method developed in section 4, the temperature set-point and occupancy signals of zone 1 are described as piecewise constant signals, i.e.,

$$\begin{aligned} \dot{T}_r(\tau) &= 0, \tau \in [t_i, t_{i+1}) & \dot{\eta}(\tau) &= 0, \tau \in [t_j, t_{j+1}) \\ T_r(t_i) &= T_{r,i} & \eta(t_j) &= \eta_j \end{aligned} \quad (31)$$

where indices  $i, j$  belong to a subset of  $\mathbb{N}$  and  $T_r(t_i), \eta(t_j)$  are the initial conditions of  $T_r(\tau), \eta(\tau)$  on the intervals  $[t_i, t_{i+1})$  and  $[t_j, t_{j+1})$ , respectively. Clearly, the least common multiple of the minimal polynomials of the state matrices of the two first-order differential equations in (31) is  $\delta(s) = s$ , so that  $p = 1$ . The IMP-filter is therefore a 1-dimensional filter simply given by  $\xi = e$ , and the environment state reads  $\zeta = \text{col}(x, \xi)$  with  $x = T_{z_1}$ , the temperature of zone 1. The LQR parameters are chosen as  $Q_d = I_2, R_d = \rho = 10^4$  and the sampling period equal  $h = 60\text{s}$ . The regularization parameter  $\lambda$  is chosen as the square of the smallest singular value of the rank-deficient data matrix  $\Psi$ . The controller was learned by running algorithm 1 which yields the optimal  $Q$ -learning based LQR feedback gain  $K = \hat{Q}_{22}^{-1} \hat{Q}_{21} = [0.5022, -0.0055]$ .

A test was carried out on the system on January 15, 2024, from 00:00 to 23:00, with a temperature setpoint profile and the measured zone occupancy (number of occupants) as shown in the figure 5. During the typical hours when the room is unoccupied, i.e., between midnight (00:00) and 06:00 and between 20:00 and midnight, the setpoint is set to  $17^\circ\text{C}$  for energy-savings. Between 06:00 and 08:00, the setpoint follows a linear time profile to reach the comfort temperature of  $22^\circ\text{C}$ . During lunchtime (12:00- 14:00), the setpoint is lowered to  $20^\circ\text{C}$ , and then set back to  $22^\circ\text{C}$  in the afternoon, until classes end at 18:00. From 18:00 onwards, the setpoint again follows a linear time profile, being gently reduced to  $17^\circ\text{C}$  from 20:00 onwards.

As it can be seen from figure 5, the learned controller performs very well and allows the temperature setpoint to be correctly maintained regardless of occupancy and outside temperature. More surprisingly, although the setpoint model used for the design is that of piecewise-constant signals, it is found that the controller is capable of perfectly tracking a setpoint that varies linearly with time. It is also seen that the VAV's air mass flow demand remains contained due to the formulation of the optimal control problem, thus ensuring a low energy demand for room heating.

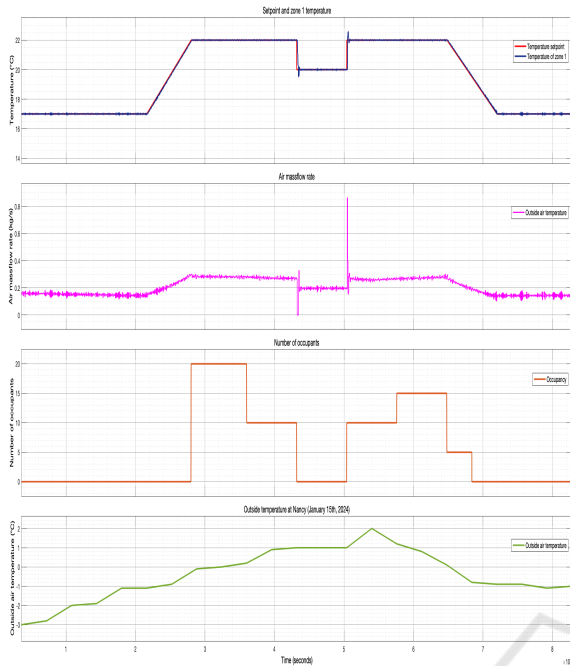


Figure 5: One-day simulation results with the  $Q$ -learning-based LQR controller.

## 6 CONCLUSIONS

In this paper, we have developed a  $Q$ -learning based LQR controller from time series generated by a building-HVAC system. The main objective was to design the control of VAV terminal boxes to guarantee thermal comfort despite climatic variations and large, random variations in building rooms occupancy. The test results obtained on the building with the model-free  $Q$ -learning controller show the excellent behavior of this controller in setpoint tracking, regardless of occupancy and external climatic conditions. From a mathematical viewpoint with regards to the estimation of the parameters of the quality function, interesting questions arise concerning the rank-deficiency of the data matrix. These issues will be addressed and reported elsewhere.

## ACKNOWLEDGEMENTS

This project is funded by the French Ministry of Europe and Foreign Affairs (MEAE) and the Ministry of Higher Education and Research (MESR) and the Moroccan Ministry of Higher Education, Scientific Research and Innovation (MESRSI), under the framework of the Franco-Moroccan bilateral program PHC TOUBKAL TBK/23/165, with Grant number: Cam-

pus N° 48604PM.

## REFERENCES

- Amende, K. L., Keen, A. J., Catlin, L. E., Tosh, M., Sneed, A. M., and Howell, R. H. (2021). *Principles of Heating, Ventilating and Air-Conditioning*. ASHRAE, Peachtree Corners, GA.
- ASHRAE (2021). *ASHRAE Handbook, Fundamentals*, volume 2021.
- Bergman, T. L. and Lavine, A. S. (2017). *Fundamentals of Heat and Mass Transfer*. John Wiley & Sons, Hoboken, NJ, 8th edition.
- Davison, E. J. and Goldenberg, A. (1975). Robust control of a general servomechanism problem: The servo compensator. *Automatica*, 11(5):461–471.
- EIA (2024). How much energy is consumed in U.S. buildings? <https://www.eia.gov/tools/faqs>.
- El Ghaoui, L. and Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064.
- Farjadnasab, M. and Babazadeh, M. (2022). Model-free lqr design by  $q$ -function learning. *Automatica*, 137:110060.
- Francis, B. A. and Wonham, W. M. (1976). The internal model principle of control theory. *Automatica*, 12(5):457–465.
- Graham, A. (2018). *Kronecker Products and Matrix Calculus with Applications*. Dover Publications Inc, Mineola, N. Y.
- Han, M., May, R., Zhang, X., Wang, X., Pan, S., Da, Y., and Jin, Y. (2020). A novel reinforcement learning method for improving occupant comfort via window opening and closing. *Sustainable Cities and Society*, 61:102247.
- Jia, M., Srinivasan, R. S., and Raheem, A. A. (2017). From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency. *Renewable and Sustainable Energy Reviews*, 68:525–540.
- Liu, X. and Gou, Z. (2024). Occupant-centric hvac and window control: A reinforcement learning model for enhancing indoor thermal comfort and energy efficiency. *Building and Environment*, 250:111197.
- Ouf, M. M., Park, J. Y., and Gunay, H. B. (2021). On the simulation of occupant-centric control for building operations. *Journal of Building Performance Simulation*, 14(6):688–691.
- Soleimanijavid, A., Konstantzos, I., and Liu, X. (2024). Challenges and opportunities of occupant-centric building controls in real-world implementation: A critical review. *Energy and Buildings*, 113958.
- Vogel, C. R. (2002). *Computational Methods for Inverse Problems*. SIAM series on Frontiers in Applied Mathematics.



- Wang, Z., Calautit, J., Tien, P. W., Wei, S., Zhang, W., Wu, Y., and Xia, L. (2023). An occupant-centric control strategy for indoor thermal comfort, air quality and energy management. *Energy and Buildings*, 285:112899.
- Xu, X., Yu, H., Sun, Q., and Tam, V. W. (2023). A critical review of occupant energy consumption behavior in buildings: How we got here, where we are, and where we are headed. *Renewable and Sustainable Energy Reviews*, 182:113396.
- Yamé, J. (1987). Dual adaptive control of stochastic systems via information theory. In *26th IEEE Conference on Decision and Control*, pages 316–320. doi:10.1109/CDC.1987.272811.
- Yamé, J. J. (2024). On the rank-deficiency of the data matrix of the  $q$ -learning based lqr problem. Technical Report 02-24, CRAN.
- Yu, H., Tam, V. W., and Xu, X. (2024). A systematic review of reinforcement learning application in building energy-related occupant behavior simulation. *Energy and Buildings*, 114189.

