Stochastic Optimal Control of HVAC system for Energy-efficient Buildings

Yu Yang, *Student Member, IEEE,* Guoqiang Hu, *Senior Member, IEEE,* and Costas J. Spanos, *Fellow, IEEE*

***Abstract*—This paper aims to develop an agile, adaptive and energy-efficient method for HVAC control via Markov decision process (MDP). Our main contributions are outlined *First*, we formulate the problem as a MDP, which incorporates *i)* the multiple uncertainties resulting from the weather and occupancy, *ii)* the elaborate Predicted Mean Vote (PMV) thermal comfort model. *Second*, to cope with the computational challenges, we propose a gradient-based policy iteration (GBPI) method to learn the policies based on the performance gradients. *Thrid*, we theoretically prove that the method can converge to an optimal policy of the formulated MDP. The advantages of the proposed method are that: *i)* it uses off-line computation to learn control policies thus reducing on-line computation burden, and *ii)* it handles the non-convex and nonlinear system dynamics efficiently and can accommodate the non-analytical thermal comfort models (e.g., PMV) in the literature. The favorable performance of the policies yield by the GBPI is demonstrated through comparisons with the optimal solution obtained by assuming all the information is available before the planning in several case studies.**

***Index Terms*—Healting, ventilation and air-conditioning (HVAC) systems, Markov decision process (MDP), off-line, uncertainties, Predicted Mean Vote (PMV).**

NOMENCLATURE

Notations:

α_w	The absorption efficient of the wall;
A_{gs}	The area of glass window [m^2];
A_{wl}/A_{wl}	The area of left/right wall [m^2];
C_p	Air specific heat [$\mathrm{J}/(\mathrm{kg}\cdot\mathrm{K})$)];
C_w	The wall capacity [$\mathrm{J}/(\mathrm{kg}\cdot\mathrm{K})$];
c_t	The electricity price at time t [s\$/kW];
η	The reciprocal of coefficient of performance of chiller;
$G_t^{\mathrm{FAU}}/G_t^{\mathrm{FCU}}$	The supply air flow rate of FAU/FCU at time t [$\mathrm{kg\,s^{-1}}$];
$G^{\mathrm{FCU,Rated}}$	The nominal air flow rate of FCU [$\mathrm{kg\,s^{-1}}$];
$G^{\mathrm{FAU,Rated}}$	The nominal air flow rate of FAU [$\mathrm{kg\,s^{-1}}$];
$\underline{G}^{\mathrm{FAU}}/\underline{G}^{\mathrm{FCU}}$	The lower bound of the supply air flow rate by FAU/FCU [$\mathrm{kg\,s^{-1}}$];
$\overline{G}^{\mathrm{FAU}}/\overline{G}^{\mathrm{FCU}}$	The upper bound of the supply air flow rate by FAU/FCU [$\mathrm{kg\,s^{-1}}$];
H_t^o	Outdoor relative humidity at time t [%];
H_t^a	Indoor relative humidity at time t [%];
H_g	The average humidity generation rate per occupant [$\mathrm{kg\,s^{-1}}$];
H_t^{FAU}	The relative humidity of the supply air by the FAU [%];
h_{gs}	The heat transfer coefficient of glass window [$\mathrm{J}/(\mathrm{m}^2\cdot{}^\circ\mathrm{C})$];
h_w	The heat transfer coefficient of walls [$\mathrm{J}/(\mathrm{m}^2\cdot{}^\circ\mathrm{C})$];
m^a	The mass of indoor air [kg];
m^{wl}/m^{wr}	The mass of the left/right wall [kg];
$P^{\mathrm{FCU,fan,Rated}}$	The nominal fan power of FCU [kW];
$P^{\mathrm{FAU,fan,Rated}}$	The nominal fan power of FAU [kW];
Q_o	The average internal heat generation rate per occupant [$\mathrm{J\,s^{-1}}$];
Q_t^{dev}	The average heat generation rate of devices caused by per occupant at time t [$\mathrm{J\,s^{-1}}$];
Q_t^w	The solar radiation density on the wall at time t [$\mathrm{J}/\mathrm{m}^2\cdot\mathrm{s}$];
T_t^o	Outdoor temperature at time t [°C];
T_t^a	Indoor temperature at time t [°C];
T_t^{wl}/T_t^{wr}	The temperature of the interior left (right) wall [°C];
$T_t^{\mathrm{FAU}}/T_t^{\mathrm{FCU}}$	The set-point temperature of FAU/FCU at time t [°C];
$\underline{T}^{\mathrm{FAU}}/\underline{T}^{\mathrm{FCU}}$	The lower bound of the set-point temperature of FAU/FCU [°C];
$\overline{T}^{\mathrm{FAU}}/\overline{T}^{\mathrm{FCU}}$	The upper bound of the set-point temperature of FAU/FCU [°C].

Acronyms:

$HVAC$	Heating, ventilation and air-conditioning;
MDP	Markov decision process;
MPC	Model predictive control;
$GBPI$	Gradient-based policy iteration.

This work is supported by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore.

Yu Yang is with SinBerBEST, Berkeley Education Alliance for Research in Singapore, Singapore 138602 e-mail: (yu.yang@bears-berkeley.sg).

Guoqiang Hu is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798 e-mail: (gqhu@ntu.edu.sg).

Costas J. Spanos is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, 94720 USA email: (spanos@berkeley.edu).

I. INTRODUCTION

BUILDINGS, especially buildings' heating, ventilation and air-conditioning (HVAC) systems, account for a large proportion of the world's energy consumption [1]. This issue has raised widespread concerns from the governments, stakeholders and research communities with the common target towards a sustainable future.

As the traditional patterns of HVAC control, i.e., fixed set-points and heuristic–based rules, are far from being energy-

efficient, there exists considerable potential to save energy by improve their energy efficiency.

A. Literature

In the literature, there exist various works on investigating HVAC controllers (see [2–4] and the references therein). The available results can generally be categorized by *i)* modeling aspects and *ii)* control methods. The recent decades have seen the progress of softwares and models for HVAC systems including *(i)* physics-based models or softwares (e.g., Dest [5], EnergyPlus [6]), *(ii)* data-driven or black-box models based on machine learning or artifical neutral network (ANN) [7, 8], and *(iii)* gray-box models based on some simplified physical principles, such as energy conservation equations [9, 10] and resistance-capacitance (RC) network [11, 12], etc. Generally, the last two types are mainly used for control purpose. The typical control methods include sequential quadratic progamming (SQP) [13], mixed-integer linear programming (MILP) [11, 14], fuzzy logic or genetic algorithms [7, 15–18], and rule-based strategies [19–21]. However, the complex system behaviors (non-linear and non-convex) still represent a computational challenges for studying HVAC control. Most of the existing optimization-based control methods depend on some approximations or linearization techniques to tackle the non-linear system dynamics (see [13] and the references above), which are not efficient and may not be adaptable. Moreover, the introduction of elaborate thermal comfort models, such as Predicted Mean Vote (PMV) [22] will further contribute to the computational challenges as they are shown to be nonlinear and non-analytical in nature (see [11, 23, 24]). Another remaining challenge that has not been well studied is the various uncertainties, such as the outdoor weather conditions and the indoor occupancy. To address this problem, most of the existing control methods were designed using a MPC framework, i.e., at each time instance, the current control inputs are computed by solving a multi-time step optimization problem based on the current measures and predictions for the predefined planning horizon. The procedure is repeated for the next time instance until the end of optimization horizon is reached. However, MPCs have the following limitations: *i)* the computational burden depends on the modeling complexity and some linearization, approximation and convexity techniques are usually required to make the problem tractable or computable [11, 25]; *ii)* the performance of these methods are usually affected by the available prediction accuracy and prediction periods [26–28]; *(iii)* they are usually conservative as they compute a deterministic control sequence based on the predicted (average) information, which has to cope with all possible disturbance caused by the uncertainties [28]; *(iv)* they are usually computationally inefficient as they requires repeated on-line computation over the stages. For the second issue, some MPC variations, i.e., stochastic model model predictive control (SMPC) [29, 30] and explicit MPC [31, 32] have been exploited to deal with the uncertainties. For SMPC, the performance and conservatism of the control can be balanced by introducing chance constraints, which state the tolerable probability of constraint violations [29]. However, those chance constraints contribute to the computational challenges of the problems. Some of the existing solution methods for SMPCs are restrictive in application due to the Gaussian distribution assumptions imposed on disturbances, which does not hold for the weather and occupancy [33, 34]. Most of the successful implementations of SMPCs [29] or their variations, such as Randomized MPC (RMPC) [35], are scenario-based approximations for the chance constraints. Though demonstrated with satisfactory performance in numeric studies, they generally rely on a large number of the on-line scenarios and require high computation cost. In contrast to the SMPCs, the explicit MPCs attempt to reduce on-line computation by adopting an off-line parametric programming beforehand. However, such kind of methods are usually not applicable due to model complexity and the various uncertainties.

Except for MPCs and their variations, another general framework for sequential decision-making under uncertainties is Markov decision process (MDP) [36]. As MDP has created a general framework to tackle uncertainties and doesn't depend on the problem structures, it has been exxplored for HVAC control in the literature [10, 37–39]. In particular, Sun *et. al.* [10, 37] studied the energy-efficient management of building energy system through the integrated control of HVAC system, lights and natural ventilation based on MDP. However, as clarified in those above works, the applications of such framework to HVAC control still face computational challenges as the computation burden of the existing traditional methods for MDPs generally exponentially increase with the state and action space.

B. Our Contributions

Complementary to the literature, this paper is aimed to develop an agile, adaptive and energy-efficient control method for HVAC systems via MDP. Our main contribution are outlined. *First*, we formulate the problem as a MDP, which incorporates the multiple uncertainties (i.e, weather and occupancy) and the well-known PMV index to describe thermal comfort. *Second*, to tackle the challenging computation issue, we propose a gradient-based policy iteration (GBPI) method based on performance gradients. *Third*, we theoretically prove that the proposed method can approach an optimal policy of the MDP. The main advantages of the proposed method against the literature are that *i)* it uses off-line computations to learn/obtain the control policies, thus reducing on-line computation burden in implementation; and *ii)* it handles the non-linear and non-convex system dynamics efficiently and can accommodate the non-analytical thermal comfort model, such as PMV index. The performance of the policies yield by GBPI is demonstrated through comparisons with the optimal solution attained by assuming all the information is available before the planning.

The remainder is arranged as follows. In Section II, the problem is investigated and formulated as a MDP. In Section III, the GBPI method is proposed. In Section VI, the performance of the proposed method is studied in several case studies. In Section V, we briefly conclude this paper.

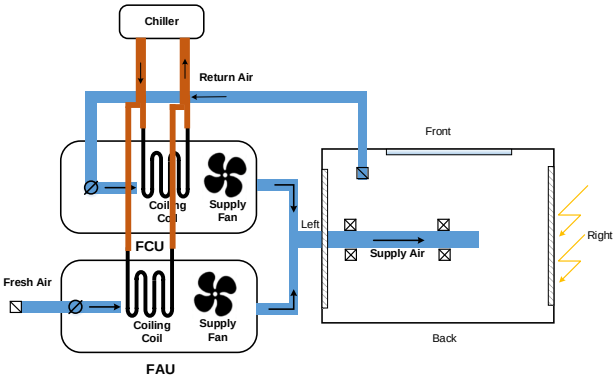


Fig. 1. The schematic of a typical HVAC system for a general room

II. The Problem

This section studies the problem formulation for a general HVAC system.

A. HVAC System

Following [10, 37–39], this paper studies a typical HVAC system as depicted in Fig. 1. The HVAC system is mainly composed of air handling units, i.e., fresh air unit (FAU), fan coil unit (FCU) and chiller. The air handling units are responsible for cooling/heating the air to the set-point temperature and forcing the supply air to the duct network through the fans. The air handling units are responsible for cooling/heating the air to the set-point temperature and forcing the supply air to the duct network through the fans. Besides, the air can be dehumidified while circulating the air handling units if necessary. Generally, the recirculated air of the room and the outdoor fresh air are handled by the FCU and FAU, respectively. The chiller provides chilled water to the cooling coils within the air handling units. Without loss of generality, this paper investigates the cooling mode of the HVAC system. We refer the readers to [10] for more details. This paper uses such type of system to develop a general framework, and the following problem formulation framework and solution method can be extended to other cases, such as the HVAC system discussed in [40].

As shown in Fig. 1, we focus on the control of HVAC system for a general room embraced by four sides , i.e., left (wall), right (wall), front (glass window) and back (door). We assume there only exists heat gain from solar radiation on the right (wall). The indoor thermal condition (i.e., temperature, humidity) depends on the operation of the HVAC system (i.e., FAU and FCU) and the interplay of indoor and outdoor environment. To improve energy efficiency, this paper discusses the control of the supply air flow rates and the set-point temperature of FAU and FCU with the objective to minimize the total energy cost while guaranteeing the desired thermal comfort indicated by the PMV metric.

To simplify discussions, the problem is discussed and formulated in a discrete framework with *i)* the optimization cycle (one day) equally divided into $T = 48$ stages with a decision interval $\Delta t = 30$ mins; and *ii)* the state variables, i.e., the range of outdoor temperature, outdoor relative humidity and indoor occupancy, are equally discretized into L_T, L_H and L_A levels.

B. MDP

1) System State: As the cooling demand of the room depends on the current indoor thermal condition, the dynamics of outdoor weather conditions and the internal thermal loads, we define the system state as

$$S_t = [T_t^o, H_t^o, T_t^a, H_t^a, N_t^a]^T$$

2) Decision Variables: Following [10, 37], the supply air flow rates and the set-point temperature of FAU and FCU are selected as the decision variables at time t:

$$A_t = [G_t^{\text{FAU}}, T_t^{\text{FAU}}, G_t^{\text{FCU}}, T_t^{\text{FCU}}]^T \tag{1}$$

3) System Dynamics: The operation of the HVAC system is subject to the dynamics of indoor thermal condition, the outdoor weather condition, and the occupancy. The gray-box models in [9, 10] based on energy and mass conservation are introduced to describe the system dynamics.

Indoor temperature: the indoor temperature is affected by various factors, i.e., the operation of the HVAC system, the internal heat generation caused by the occupants, and the heat conduction between the indoor and outdoor through the walls and the window. Thus we have

$$\begin{aligned} C_p m^a (T_{t+1}^a - T_t^a) &= N_t^i Q_o \Delta_t + P_t^{\text{dev}} \Delta_t \\ &+ h_{gs} A_{gs,i} (T_t^o - T_t^a) \Delta_t \\ &+ h_w A_{wl} (T_t^{wl} - T_t^a) \Delta_t + h_w A_{wr} (T_t^{wr} - T_t^a) \Delta_t \\ &+ G_t^{\text{FAU}} (T_t^{\text{FAU}} - T_t^i) \Delta_t + G_t^{\text{FCU}} (T_t^{\text{FCU}} - T_t^a) \Delta_t. \end{aligned} \tag{2}$$

As the internal heating generation closely relates to the occupancy, we use $P_t^{\text{dev}} = N_t^a Q_t^{\text{dev}}$ to estimate the internal heat generation caused by the electrical devices at time t. In (2), the first two terms capture the internal heat generation from the occupants and the devices, the third and fourth term calculate the heat transfer through the glass window and the wall, and the last two terms denote the cooling load supplied by the HVAC system. Wherein the temperature of the interior left and right wall can be estimated by

$$\begin{aligned} C_w m^{wl} (T_{t+1}^{wl} - T_t^{wl}) &= h_w A_{wl} (T_t^a - T_t^{wl}) \Delta_t \\ C_w m^{wr} (T_{t+1}^{wr} - T_t^{wr}) &= h_w A_{wr} (T_t^a - T_t^{wr}) \Delta_t \\ &+ \alpha_w A_{wr} Q_t^w \Delta_t \end{aligned} \tag{3}$$

Indoor relative humidity: the dynamics of indoor humidity can be described as

$$\begin{aligned} m^a (H_{t+1}^a - H_t^a) =& N_t^a H_g \Delta_t + G_t^{\text{FAU}} (H_t^{\text{FAU}} - H_t^a) \Delta_t \\ &+ G_t^{\text{FCU}} (H_t^{\text{FCU}} - H_t^a) \Delta_t \end{aligned} \tag{4}$$

where the humidity of the supply air by the FAU H_t^{FAU} can be estimated by

$$H_t^{\text{FAU}} = \min(H_t^o, H^{\text{FAU,Sat}}) \tag{5}$$

with $H^{\text{FAU,Sat}}$ denoting the saturation humidity of FAU.

Indoor occupancy: Based on the results of [41, 42], a Markov chain is used to capture the dynamic patterns of occupancy, i.e.,

$$\Pr(N_{t+1}^a = j | N_t^a = i) = p_t^{ij},\ \forall i, j, \in \{1, \cdots, L_A\}. \tag{6}$$

where p_t^{ij} denotes the transition probability of the occupancy from level i to level j at time t.

Similarly, as the outdoor weather (i.e., temperature and humidity) dynamically changes over the time, two Markov chains are introduced to capture their stochastic patterns, i.e.,

$$\begin{aligned} &\Pr(T_{t+1}^o = j | T_{t+1}^o = i) = p_t^{ij},\ \forall i, j \in \{1, \cdots, L_T\}. \\ &\Pr(H_{t+1}^o = j | H_{t+1}^o = i) = p_t^{ij},\ \forall i, j \in \{1, \cdots, L_H\}. \end{aligned} \tag{7}$$

where p_t^{ij} denotes the transition probability of the temperature or relative humidity from level i to level j at time t. The transition probabilities of the Markov chains for weather are determined based on the historical weather data of Singapore, which will be illustrated in Section IV of this paper.

4) Objective Function: To improve energy efficiency, the expected total HVAC cost is selected as the objective, i.e.,

$$J = \mathbb{E}\Big\{ \sum_{t=0}^{T-1} c_t \big(\eta(C_t^{\text{FCU}} + C_t^{\text{FAU}}) + P_t^{\text{FCU, fan}} + P_t^{\text{FAU, fan}}\big)\Delta t \Big\} \tag{8}$$

As indicated in (8), the power consumption of the HVAC system is mainly composed of *i)* the cooling power C_t^{FCU}, C_t^{FAU}, and *ii)* the fan power $P_t^{\text{FCU, fan}}$, $P_t^{\text{FAU, fan}}$. Wherein each part can be estimated by [10]

$$\begin{aligned} C_t^{\text{FCU}} =& C_p G_t^{\text{FCU}} (T_t^i - T_t^{\text{FCU}}) + C_p G_t^{\text{FCU}} \big(H_t^i (2500 + 1.84 T_t^i) \\ & - H_t^{\text{FCU}} (2500 + 1.84 T_t^{\text{FCU}})\big) \end{aligned} \tag{9a}$$

$$\begin{aligned} C_t^{\text{FAU}} =& C_p G_t^{\text{FAU}} (T_t^o - T_t^{\text{FAU}}) + C_p G_t^{\text{FAU}} \big(H_t^o (2500 + 1.84 T_t^o) \\ & - (2500 + 1.84 T_t^{\text{FAU}})\big) \end{aligned} \tag{9b}$$

$$P_t^{\text{FCU, fan}} = P^{\text{FCU,fan,Rated}} \cdot \Big(\frac{G_t^{\text{FCU}}}{G^{\text{FCU,Rated}}}\Big)^3 \tag{9c}$$

$$P_t^{\text{FAU,fan}} = P^{\text{FAU,fan,Rated}} \cdot \Big(\frac{G_t^{\text{FAU}}}{G^{\text{FAU,Rated}}}\Big)^3 \tag{9d}$$

5) PMV Metric: For providing building energy service, it's required that the indoor thermal comfort be guaranteed. This paper selects the well-known PMV metric [22] to indicate the average thermal comfort of occupants, which is determined by a particular combination of various parameters, i.e.,

$$\text{pmv}_t = PMV(M, W, T_t^a, H_t^a, t_t^r, v_t^a, I_{cl}) \tag{10}$$

where $PMV(\cdot)$ indicates the PMV model consisting of a group of nonlinear equations. We refer the readers to [11] for the details. The PMV model is non-analytical and its calculation relies on an iterative numerical process. The relevant parameters for calculating the PMV metric are shown in the right-hand-side of (10), which contains: *(1)* metabolic rate M (W/m^2), *(2)* the rate of mechanic work W (W/m^2), *(3)* air temperature T_t^a (°C), *(4)* relative humidity H_t^a (%), *(5)* mean radiation temperature t_t^r (°C), *(6)* indoor air velocity v^a (m^2), and *(7)* clothing insulation I_{cl} ($\text{m}^2\text{K}/\text{W}$). The PMV model establish a mapping of the indoor thermal condition to the comfort range $[-3, 3]$, with -3, 0, 3 indicating too cold, ideal, and too hot, respectively.

6) Constraints: The operation of the HVAC system should comply with its operation limits in practice, i.e., *(i)* the (lower and upper) bounds of the supply air flow rates imposed by the dampers within the FAU and FCU (11a) and (11b); *(ii)* the set-point temperature ranges of FAU and FCU determined by chiller capacity (11c) and (11d).

$$\underline{G}^{\text{FAU}} \le G_t^{\text{FAU}} \le \overline{G}^{\text{FAU}} \tag{11a}$$

$$\underline{G}^{\text{FCU}} \le G_t^{\text{FCU}} \le \overline{G}^{\text{FCU}} \tag{11b}$$

$$\underline{T}^{\text{FAU}} \le T_t^{\text{FAU}} \le \overline{T}^{\text{FAU}} \tag{11c}$$

$$\underline{T}^{\text{FCU}} \le T_t^{\text{FCU}} \le \overline{T}^{\text{FCU}} \tag{11d}$$

Besides, we use $\underline{\text{pmv}}$ and $\overline{\text{pmv}}$ to capture the thermal comfort requirements indicated by the PMV metric, i.e.,

$$\underline{\text{pmv}} \le \text{pmv}_t \le \overline{\text{pmv}} \tag{12}$$

7) Optimization Problem: In conclusion, the optimal operation of the HVAC system under the uncertainties can be depicted as the following constrained MDP:

$$\begin{aligned} \min_{\pi}\ & J(S_0, \pi) = \mathbb{E}\Big\{ \sum_{t=0}^{T-1} c_t \big(\eta(C_t^{\text{FCU}} + C_t^{\text{FAU}}) + P_t^{\text{FCU, fan}} + P_t^{\text{FAU, fan}}\big)\Delta t \Big\} \\ \text{subject to}\quad & \text{System dynamics: } (2) - (7), \\ & \text{Operation limits: } (11), \\ & \text{Thermal comfort: } (12),\ \forall t \in \{0, 1, \cdots, T-1\}. \end{aligned} \tag{13}$$

where we use $\pi = (\pi_0, \pi_1, \cdots, \pi_{T-1})$ to denote the policy of HVAC system over the optimization horizon. S_0 is the initial system state. At each time t, the control rule π_t: $\mathcal{S}_t \to \mathcal{A}_t$ establishes a mapping from the sate space $\mathcal{S}_t$ to the action space $\mathcal{A}_t$. $\mathbb{E}^\pi$ denotes the expectation under the policy π.

It's nontrivial to search for an optimal policy for problem (13) concerning that:

- (*i*) the multiple uncertainties make it difficult to analytically evaluate the policies under expectation.
- (*ii*) the various nonlinear constraints imposed by the system dynamics and PMV metric make it difficult to figure out the feasible policies.
- (*iii*) the multi-stage decision problem tends to suffer from curse of dimensionality due to the large state and the action space.

As a consequence, the existing traditional methods for finite-stage MDPs, such as dynamic programming (DP) [43], are not viable as they require to traverse Q-factors for the large state and action spaces.

III. Gradient-based Policy Iteration

To cope with the computational challenges, this section proposes a gradient-based policy iteration (GBPI) method based on performance gradients for problem (13). Generally, the main idea is to iteratively update the policy based on the performance gradients until the stopping criterion is reached. The remainder of this section introduces the main notations involved and the establishment of the method.

Notations: we use the lower cases s_t and a_t (b_t, c_t) to represent a state and action instance at time t. We use the integer spaces $s_t \in \{1, 2, \cdots, |\mathcal{S}_t|\}$ and $a_t, b_t, c_t \in \{1, 2, \cdots, |\mathcal{A}_t|\}$

to represent the state and action spaces at time t. For a (random) policy $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_{T-1})^T$, $\theta_t \in \mathbb{R}^{|\mathcal{S}_t|\times|\mathcal{A}_t|}$ establishes a mapping from the state space $\{1, 2, \cdots, |\mathcal{S}_t|\}$ to the action space $\{1, 2, \cdots, |\mathcal{A}_t|\}$, and we have $\theta_t(s_t, a_t) \in [0, 1]$ denoting the probability to take action a_t at the state s_t under the policy θ. Without specification, the lower cases $p_t(s_{t+1}|s_t, a_t)$ and $p_t^{\boldsymbol{\theta}}(s_{t+1}|s_t)$ indicate the transition probability from state s_t to state s_{t+1} under the action a_t or the policy $\boldsymbol{\theta}$, and we have $\boldsymbol{P}_t^{\theta} = [p_t^{\boldsymbol{\theta}}(s_{t+1}|s_t)]_{|\mathcal{S}_t|\times|\mathcal{S}_{t+1}|}$ indicate the transition probability matrix under the policy $\boldsymbol{\theta}$ at time t. The superscript $k \in \mathbb{N}$ of $\boldsymbol{\theta}^k$ denotes the iteration.

For problem (13), we can describe the stage-cost as

$$r_t(s_t, a_t) = c_t\big(\eta(C_t^{\text{FCU}} + C_t^{\text{FAU}}) + P_t^{\text{FCU, fan}} + P_t^{\text{FAU, fan}}\big)\Delta t$$

For any two policies $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$, we have the following performance difference equation [44]:

$$J(\boldsymbol{\mu}; S_0) - J(\boldsymbol{\sigma}; S_0) = \sum_{t=0}^{T-1} \pi_t^{\boldsymbol{\mu}} \Big[(\boldsymbol{r}_t^{\boldsymbol{\mu}} - \boldsymbol{r}_t^{\boldsymbol{\sigma}}) + (\boldsymbol{P}_t^{\boldsymbol{\mu}} - \boldsymbol{P}_t^{\boldsymbol{\sigma}}) V_{t+1}^{\boldsymbol{\sigma}}\Big] \quad (14)$$

where we use $\boldsymbol{\pi}_t^{\boldsymbol{\mu}} = (\pi_t^{\mu}(1), \pi_t^{\mu}(2), \cdots, \pi_t^{\mu}(|\mathcal{S}_t|))^T$ to denote the state distribution at time t under policy $\boldsymbol{\mu}$ $\boldsymbol{r}_t^{\boldsymbol{\theta}} = (r_t^{\boldsymbol{\theta}}(1), r_t^{\boldsymbol{\theta}}(2), \cdots, r_t^{\boldsymbol{\theta}}(|\mathcal{S}_t|))^T$, $\boldsymbol{P}_t^{\boldsymbol{\theta}}$, and $V_{t+1}^{\boldsymbol{\theta}} = (V_{t+1}^{\boldsymbol{\theta}}(1), V_{t+1}^{\boldsymbol{\theta}}(2), \cdots, V_{t+1}^{\boldsymbol{\theta}}(|\mathcal{S}_{t+1}|))^T$ denote the one-step cost, transition probability matrices, and the performance potentials under a given policy $\boldsymbol{\theta}$ ($\boldsymbol{\theta} \in \{\boldsymbol{\sigma}, \boldsymbol{\mu}\}$), and we have

$$V_{t+1}^{\boldsymbol{\theta}}(s_{t+1}) = \sum_{\tau=t+1}^{T-1} r_\tau(s_\tau, a_\tau), \text{with } a_\tau = \boldsymbol{\theta}(s_\tau). \quad (15)$$

We note that (14) quantifies the performance gap for the two policies ($\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$). Based on the theory of Perturbation Analysis (PA), we view $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ as the base and perturbed policy. One may note that the performance of the perturbed policy $\boldsymbol{\mu}$ can be calculated only if the performance potentials ($V_{t+1}^{\boldsymbol{\sigma}}$) under the base policy $\boldsymbol{\sigma}$ and the state distribution ($\boldsymbol{\pi}^{\boldsymbol{\mu}}$) under the perturbed policy $\boldsymbol{\mu}$ can be figured out (not know as a *a priori*). Thus, it's impractical to update the policy based on the performance difference equation (14). However, we can imply some information from it to achieve policy improvement. To achieve such goal, we suppose a random policy $\boldsymbol{\delta}$, which adopts policy $\boldsymbol{\sigma}$ with probability δ and adopts policy $\boldsymbol{\mu}$ with probability 1–δ. We can figure out the state transition probability matrices $\boldsymbol{P}_t^{\boldsymbol{\delta}} = \boldsymbol{P}_t^{\boldsymbol{\sigma}} + \delta\Delta\boldsymbol{P}_t$ with $\Delta\boldsymbol{P}_t = \boldsymbol{P}_t^{\boldsymbol{\mu}} - \boldsymbol{P}_t^{\boldsymbol{\sigma}}$, and the stage-cost vectors $\boldsymbol{r}_t^{\boldsymbol{\delta}} = \boldsymbol{r}_t^{\boldsymbol{\sigma}} + \delta\Delta\boldsymbol{r}_t$ with $\Delta\boldsymbol{r}_t = \boldsymbol{r}_t^{\boldsymbol{\mu}} - \boldsymbol{r}_t^{\boldsymbol{\sigma}}$ for policy $\boldsymbol{\delta}$. Thus, the performance difference equation (14) is equivalent to

$$J(\boldsymbol{\delta}; S_0) - J(\boldsymbol{\sigma}; S_0) = \sum_{t=0}^{T-1} \pi_t^{\boldsymbol{\delta}} \Big[\Delta\boldsymbol{r}_t + \Delta\boldsymbol{P}_t \boldsymbol{V}_{t+1}^{\boldsymbol{\sigma}}\Big] \quad (16)$$

By letting $\delta \to 0$ in (16), we have the performance differential equation, i.e.,

$$\frac{dJ(\boldsymbol{\delta}; S_0)}{d\delta} = \lim_{\delta\to 0} \sum_{t=0}^{T-1} \pi_t^{\boldsymbol{\delta}} \Big[\Delta\boldsymbol{r}_t + \Delta\boldsymbol{P}_t \boldsymbol{V}_{t+1}^{\boldsymbol{\sigma}}\Big] \quad (17)$$

From (17), we further have

$$\frac{\partial J(\boldsymbol{\sigma}; S_0)}{\partial \sigma_t(s_t, a_t)} = \pi_t^{\boldsymbol{\sigma}}(s_t)\Big[\frac{\partial r_t^{\boldsymbol{\sigma}}(s_t)}{\partial \sigma_t(s_t, a_t)} + \sum_{s_{t+1}\in\{1,2,\cdots,|\mathcal{S}_{t+1}|\}} \frac{\partial p_t^{\boldsymbol{\sigma}}(s_{t+1}|s_t)}{\partial \sigma_t(s_t, a_t)} \boldsymbol{V}_{t+1}^{\boldsymbol{\sigma}}(s_{t+1})\Big] \quad (18)$$

As $r_t^{\boldsymbol{\sigma}}(s_t) = \sum_{a_t=1}^{|\mathcal{A}_t|} p_t^{\sigma}(a_t|s_t) r_t(s_t, a_t)$ and $p_t^{\boldsymbol{\sigma}}(s_{t+1}|s_t) = \sum_{a_t=1}^{|\mathcal{A}_t|} p_t^{\boldsymbol{\sigma}}(a_t|s_t) p(s_{t+1}|s_t, a_t)$, we have

$$\begin{aligned}\frac{\partial J(\boldsymbol{\sigma}; S_0)}{\partial \sigma_t(s_t, a_t)} &= \pi_t^{\boldsymbol{\sigma}}(s_t)\Big[\sum_{b_t=1}^{|\mathcal{A}_t|} \frac{\partial p_t^{\boldsymbol{\sigma}}(b_t|s_t)}{\partial \sigma_t(s_t, a_t}) r_t(s_t, b_t) \\ &\quad + \sum_{b_t=1}^{|\mathcal{A}_t|} \frac{\partial p_t^{\boldsymbol{\sigma}}(b_t|s_t)}{\partial \sigma_t(s_t, a_t)} p_t(s_{t+1}|s_t, b_t) \boldsymbol{V}_{t+1}^{\boldsymbol{\sigma}}(s_{t+1})\Big] \quad (19) \\ &= \pi_t^{\boldsymbol{\sigma}}(s_t)\Big[\sum_{b_t=1}^{|\mathcal{A}_t|} \frac{\partial p_t^{\boldsymbol{\sigma}}(b_t|s_t)}{\partial \sigma_t(s_t, a_t)} \big(r_t(s_t, b_t) + V_t^{\boldsymbol{\sigma}}(s_t, b_t)\big)\Big]\end{aligned}$$

where we have $V_t^{\boldsymbol{\sigma}}(s_t, b_t) = p_t(s_{t+1}|s_t, b_t) V_{t+1}^{\boldsymbol{\sigma}}(s_{t+1})$.

As $p_t^{\boldsymbol{\sigma}}(b_t|s_t) = \frac{\sigma_t(s_t, b_t)}{\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t(s_t, c_t)}$, we have

$$\frac{\partial p_t^{\boldsymbol{\sigma}}(b_t|s_t)}{\partial \sigma_t(s_t, a_t)} = \begin{cases} \dfrac{\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t(s_t, c_t) - \sigma_t(s_t, a_t)}{[\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t(s_t, c_t)]^2}, & b_t = a_t \\ \dfrac{-\sigma(s_t, b_t)}{[\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t(s_t, c_t)]^2}, & b_t \neq a_t \end{cases} \quad (20)$$

(19) can be interpreted as the gradients of the performance at policy $\boldsymbol{\sigma}$. Thus, an improved policy $\boldsymbol{\sigma}^{i+1}$ can be obtained from a base policy $\boldsymbol{\sigma}^k$ according to

$$\boldsymbol{\sigma}^{k+1} = \boldsymbol{\sigma}^k - \boldsymbol{\gamma}^k \cdot \nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0) \quad (21)$$

where we have $\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0) = \big[\frac{\partial J(\boldsymbol{\sigma}^k; S_0)}{\partial \sigma_t^k(s_t, a_t)}\big]_{|\mathcal{S}_t|\times|\mathcal{A}_t|}$. $\boldsymbol{\gamma}^k = [\gamma^k(s_t, a_t)]_{|\mathcal{S}_t|\times|\mathcal{A}_t|}$ denotes the step-size at iteration k.

Observe (21), we note that the remaining problems include *(i)* computing the performance gradients $\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0)$; *(ii)* determining the step-size $\boldsymbol{\gamma}^k$. As there exists a variety of randomness, it is very difficult to analytically estimate the performance gradients of policies under expectation in practice. To overcome this difficulty, the Monte Carlo (MC) method [45] is adopted. Specifically, the performance gradients of a given policy are estimated by averaging (19) under a number of randomly generated sample paths (scenarios). The main procedures to estimate the performance gradients are shown in **Algorithm** 1, and the complete framework to perform GBPI is in **Algorithm** 2. The step-size $\boldsymbol{\gamma}^k$ determines the converge of the method, and we have the main results in ***Theorem*** 1. We define the stopping criterion of **Algorithm** 2 as

$$\|\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0)\|_2 \leq \epsilon \quad (22)$$

where ϵ denotes a positive threshold.

One may note that one advantage of the proposed method is that it can be implemented off-line based on the historical data, thus greatly reducing the on-line computations in implementation. More specifically, as the control policies have been learned off-line, the main computation for the on-line implementation is to look up the policy table and figure out

the right action based on the measured system state (i.e., indoor/outdoor temperature, humidity and occupancy).

Algorithm 1 Performance Gradients Estimation Based on MC

1: **Input**: a given policy $\boldsymbol{\theta}$.
2: Generate S feasible sample paths [1] by performing policy $\boldsymbol{\theta}$, and index the state, action, stage-cost sequences as

$$\{s_0^\omega, a_0^\omega, r_t^\omega, s_1^\omega, a_1^\omega, r_1^\omega \cdots, s_{T-1}^\omega, a_{T-1}^\omega, r_{T-1}^\omega\}, \quad \forall \omega \in \{1, 2, \cdots, S\}.$$

3: **For** $t = 0, 1, \cdots, T$ **do**
4: Record the state s_t^ω, the actions a_t and the sample path indexes according to

$$\Omega_t(S_t) = \{s_t^\omega | \omega \in \{1, 2, \cdots, S\}\}.$$
$$\Omega_t(A_t|s_t) = \{a_t^\omega | S_t = s_t, \omega \in \{1, 2, \cdots, S\}\} \quad \forall s_t \in \Omega_t(S_t).$$
$$\mathcal{I}(s_t) = \{\omega | \omega \in \{1, 2, \cdots, S\}, s_t^\omega = s_t\}, \quad \forall s_t \in \{1, 2, \cdots, |\mathcal{S}_t|\}.$$
$$\mathcal{I}(s_t, a_t) = \{\omega | \omega \in \{1, 2, \cdots, S\}, s_t^\omega = s_t, a_t^\omega = a_t\}, \quad \forall s_t \in \Omega_t(S_t), a_t \in \Omega_t(A_t|s_t).$$

5: **For** $s_t \in \Omega_t$ **do**
6: **For** $a_t \in \Omega_t(A_t|s_t)$ **do**
7: According to (19) and (20), estimate $\frac{\partial J(\boldsymbol{\sigma}; S_0)}{\partial \sigma_t(s_t, a_t)}$ by

$$\frac{\partial J(\boldsymbol{\sigma}; S_0)}{\partial \sigma_t(s_t, a_t)} = \pi_t^{\boldsymbol{\sigma}}(s_t)\Big[\sum_{b_t \in \Omega_t(A_t|s_t)} \frac{\partial p_t^{\boldsymbol{\sigma}}(b_t|s_t)}{\partial \sigma_t(s_t, a_t)}\big(r_t(s_t, b_t) + V_t^{\boldsymbol{\sigma}}(s_t, b_t)\big)\Big], \text{with}$$
$$\pi_t^{\boldsymbol{\sigma}}(s_t) \approx |\mathcal{I}(s_t)|/S$$
$$r_t(s_t, b_t) \approx \frac{1}{|\mathcal{I}(s_t, b_t)|} \sum_{\omega \in \mathcal{I}(s_t, b_t)} r_t^\omega$$
$$V_t^{\boldsymbol{\sigma}}(s_t, b_t) \approx \frac{1}{|\mathcal{I}(s_t, b_t)|} \sum_{\omega \in \mathcal{I}(s_t, b_t)} \sum_{\tau=1}^{T-1} r_\tau^\omega$$

8: **EndFor**
9: **EndFor**
10: **EndFor**

Algorithm 2 Gradient-based Policy Iteration (GBPI)

1: **Initialization:** $k \to 0$, $\boldsymbol{\sigma}^0$.
2: **Iteration:**
3: Estimate the gradients $\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0)$ at policy $\boldsymbol{\sigma}^k$ according to **Algorithm** (1).
4: **Policy Update:**

$$\boldsymbol{\sigma}^{k+1} = \boldsymbol{\sigma}^k - \boldsymbol{\gamma}^k \cdot \nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0) \tag{23}$$

5: Stop if the stopping criterion is reached, otherwise go to **Step** 3.

***Theorem* 1.** *For any given feasible initial policy $\boldsymbol{\sigma}^0$, **Algorithm** 2 can converge to an optimal policy of problem (13) with the selected step-size $\gamma^k(s_t, a_t) = \frac{\sigma_t^k(s_t, a_t)}{\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t^k(s_t, c_t)}$ ($\forall s_t \in \{1, 2, \cdots, |\mathcal{S}_t|\}$, $a_t \in \{1, 2, \cdots, |\mathcal{A}_t|\}$) and the performance gradients $\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0)$ estimated accurately enough.*

The detailed proof can refer to **Appendix** A .

Remark 1. *As illustrated in **Theorem** 1, the convergence of the method relies on the estimation accuracy of the performance gradients for the state-action pairs. Generally, an quite accurate estimation can be obtained by increasing the number of sample paths that used. However, it's not required to guarantee performance in practice. Instead, it will be sufficient if the (performance) order of the (action) candidates can be distinguished from the estimations. Give a simple example, we only require the estimation $\tilde{J}(a) \leq \tilde{J}(b)$ if $J(a) \leq J(b)$ for any two actions a and b, where $J(\cdot)$ and $\tilde{J}(\cdot)$ denote the real and predicted value for the actions.*

IV. Case Studies

The section illustrates the performance of GBPI on HVAC control through comparison with the optimal onces obtained under a MPC approach. This section is mainly composed of the following part. *First*, the stochastic characteristics of weather in Singapore is analyzed based on historical data, and two Markov chains are used to capture the stochastic patterns of outdoor temperature and humidity, respectively. *Second*, two strategies that can be used to reduce computation are discussed based on the data analysis. *Third*, three case studies under different discretization step-size for the states and decision variables and the different configurations for HVAC systems are studied.

A. Data Analysis

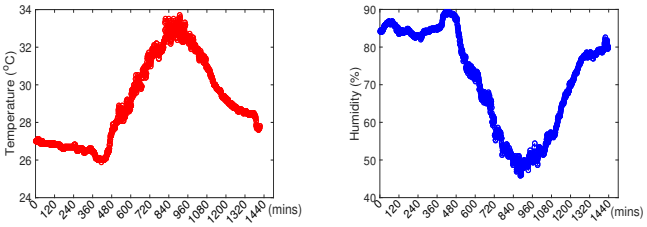
In order to perform the GBPI method, this part illustrates the establishment of the Markov chains to capture the weather patterns based on the real weather data in Singapore (from 2019/09/01 to 2019/10/13, 43 days). According to the data, the temperature and humidity curve for a typical day (2019/09/24) are plotted in Fig. 2. The figures imply some characteristics for the weather patterns of the county. Most visibly, one may observe that the peak temperature and valley humidity are most likely to appear between 12:00 and 16:00 in Singapore (see Fig. 2 (a)-(b)). However, for the early morning and late night, the outdoor temperature tends to drop but the humidity keeps at a relative high level during those periods.

To build the Markov chains, we first equally discretize the temperature and humidity with a resolution of 1°C and 5 %, respectively. After that the transition probability of the two Markov chains are estimated by

$$\text{Temperature: } p_t^{ij} \approx \frac{\sum_{\omega=1}^{Day} \mathbb{I}(T_t^{o,\omega} = i, T_{t+1}^{o,\omega} = j)}{\sum_{i=1}^{Day} \mathbb{I}(T_t^{o,\omega} = i)}, \quad \forall i, j \in \{1, 2, \cdots, L_T\},\ t \in \{1, \cdots, T-1\}.$$
$$\text{Relative humidity: } p_t^{ij} \approx \frac{\sum_{\omega=1}^{Day} \mathbb{I}(H_t^{o,\omega} = i, H_{t+1}^{o,\omega} = j)}{\sum_{i=1}^{Day} \mathbb{I}(H_t^{o,\omega} = i)}, \quad \forall i, j \in \{1, 2, \cdots, L_H\},\ t \in \{1, \cdots, T-1\}.$$

[1] While generating the sample paths, we repeatedly check the thermal comfort constraints (12). If the thermal comfort is not satisfied at state s_t while taking action a_t, we set $\sigma_t(s_t, a_t) = 0$ and regenerate a new action based on the updated policy until the thermal comfort is satisfied.

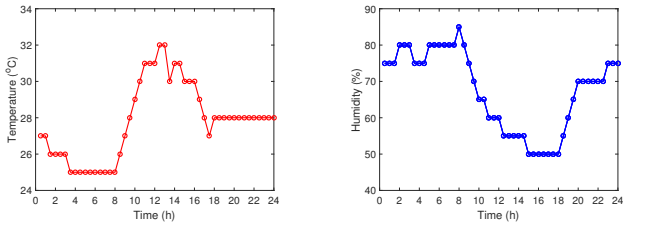
where $\mathbb{I}(\cdot)$ is the indicator function. ω denotes the index of the day, and we have $Day = 43$.



(a) (b)

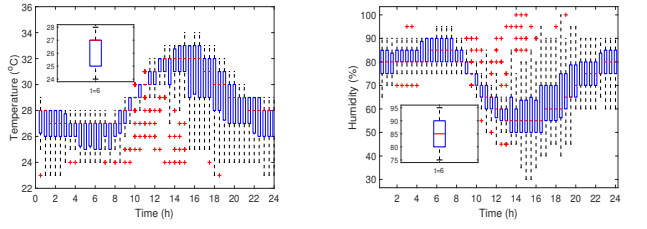
Fig. 2. (a) Outdoor temperature for a typical day. (b) Outdoor relative humidity for a typical day.

To preliminarily validate the Markov chains, we randomly generate some sample paths accordingly. We can figure out a typical temperature and relatively humidity curve (shown in Fig. 3) that correspond well to the realizations in Fig. 2. This implies the desirability of the Markov chains in capturing the weather patterns of the country, and we may use them to generate sample paths while performing **Algorithm** 1, 2 to learn the control policies for the HVAC system.



(a) (b)

Fig. 3. (a) A typical outdoor temperature scenario generated by the Markov chain. (b) A typical outdoor relative humidity scenario generated by the Markov Chain.



(a) (b)

Fig. 4. (a) The outdoor temperature distribution over the day. (b) The outdoor relative humidity distribution over the day.

B. Computation Reduction

The section discusses the following two strategies that can be used to reduce computation in practice.

Strategy I: Generally, though the weather (temperature and humidity) during a whole day spread over wide ranges (see Fig. 2, temperaure: $[22, 34]$°C, relative humidity $[40, 100]\%$), the weather are mostly concentrated in a relative narrow area at each time period of the day, which can be observed in Fig. 4. For example, at 6:00 a.m., the outdoor temperature and relative humidity mainly assemble with $[25, 27]$°C and $[80, 90]\%$, respectively. This phenomenon reveals that we may focus our computation on the scenarios that appear with high probability to exclude the rare cases while estimating the performance gradients.

Strategy II: the main computation of the GBPI is to estimate the performance gradients for the state-action pairs that appear. As the target of HVAC control is to guarantee indoor thermal comfort (to avoid the occurrence of uncomfortable states), we may initialize $\sigma_t(s_t, a_t) = 0$ for all the state-action pairs that result in indoor temperature or relatively humidity out of the range $[23, 28]$°C or $[40, 70]\%$ (uncomfortable ranges). This can quickly reduce the occurrence of discomfort states and suppose to accelerate the converge rate of the method.

We note that the two strategies discussed above are some heuristic rules, which are easy to be implemented without any contributions to the computation burden of GBPI. Moreover, the convergence of the proposed method doesn't depen on such two strategies.

C. Case Studies

This part investigates the HVAC control for a general room with size: $6\text{m}\times5\text{m}\times4\text{m}$ (height×length×width) as shown in Fig. 1. The maximum number of occupants is 5. Intuitively, we discretize the occupany into $L_A = 5$ levels. The other settings for the room and the occupants are gathered in TABLE I, which refer to [10]. We set the comfortable PMV range as $[-0.5, 0.5]$ and the static inputs for the PMV model as typical numbers (TABLE II) according to the ANSI/ASHRAE Standard [46]. The main HVAC parameters are presented in TABLE II. In the case studies, the mean radiant temperature t_t^r is estimated by 2°C higher than the instantaneous indoor air temperature according to the standard [46]. The time-of-use (TOU) price refers to [11].

TABLE I
ROOM & OCCUPANT SETTINGS

Param.	Value & Units	Param.	Value & Units
C_p	$1012\text{J}/(\text{kg}\cdot\text{K})$	m^a	144.6kg
h_{gs}	$2.5\text{W}/\text{m}^2$	m^{wl}	7.2×10^3kg
A_{gs}	10m^2	m^{wr}	8.64×10^3kg
h_w	$0.8\text{W}/\text{m}^2$	C_w	$1.05\times10^3\text{J}/(\text{kg}\cdot\text{K})$
a_w	0.4	Q_o	$40\ \text{J}\,\text{s}^{-1}$
A_{wl}	$20\ \text{m}^2$	H_g	$0.03\ \text{g}\,\text{s}^{-1}$
A_{wr}	24m^2		

TABLE II
HVAC & PMV PARAMETERS

HVAC		PMV	
Param.	Value & Units	Param.	Value & Units
$P^{\text{FAU,fan,Rated}}$	$0.1\ \text{K W}$	v^a	$0.2\text{m}/\text{s}$
$G^{\text{FAU,Rated}}$	$0.01\ \text{kg}\,\text{s}^{-1}$	M	1.0 met
$P^{\text{FAU,fan,Rated}}$	$0.1\ \text{K W}$	W	0
$G^{\text{FCU,Rated}}$	$0.05\ \text{kg}\,\text{s}^{-1}$	I_{cl}	0.155 clo
COP	2.7	P_a	1.01×10^5 Pa

We investigate the GBPI through comparison with the optimal solutions, which can be obtained by sequentially solving problem (24) with assumed accurate information for the weather and the occupancy over each planning horizon H.

$$\min_{G_t^{\text{FAU}},T_t^{\text{FAU}},G_t^{\text{FCU}},T_t^{\text{FCU}}} J(t_k) = \sum_{t=t_k+1}^{t_k+H-1} \{c_t(\eta(C_t^{\text{FCU}}+C_t^{\text{FAU}}) + P_t^{\text{FCU, fan}}+P_t^{\text{FAU, fan}})\Delta t\}$$
$$\text{subjec to} \quad \text{System dynamics}: \ (2)-(5), \quad \text{Operation limits}: \ (11), \quad \text{Thermal comfort}: (12), \quad \forall t \in \{t_k, t_k+1, \cdots, t_k+H-1\}. \tag{24}$$

Except for the time index ($t \in \{t_k, t_k+1, \cdots, t_k+H-1\}$), the other notations of problem (24) keep in accordance with the previous section. We note that problem (24) is non-linear, non-convex, and non-analytical due to the system dynamics and the introduction of PMV model, which makes it generally intractable. Following [11, 14], the linearization techniques are utilized to approximate the problem as a mixed-integer linear programming (MILP), which can be tackled efficiently by many existing toolboxes. The main ideas of the linearizations are to divide the range of state variables into several segments and introduce some redundant indicator (binary) variables. We refer the readers to [14] and [11] for the details of the linearizations for the system dynamics and PMV model.

We investigate the two methods through the following three case studies, in which the state and decision variables are discretized with different step-sizes.

Case I: The discretization intervals for temperature and relative humidity are $2°\text{C}$ and 10%, respectively. The set-point temperature range of FAU and FCU are $\{12, 14, 16\}°\text{C}$, and their supply air flow rates are equally divided into 3 levels.

Case II: The discretization intervals for temperature and relative humidity keep in accordance with *Case I*. However, the set-point temperature of FAU and FCU are fixed as $15°\text{C}$, and their supply air flow rates are equally divided into 5 levels.

Case III: The discretization intervals for temperature and the relative humidity are $1°\text{C}$ and 5%, respectively. The settings for FAU and FCU keep consistent with *Case II*.

While performing the GBPI method, 1000 (*Case I*), 2000 (*Case II*) and 5000 (*Case III*) sample paths are generated to estimate the performance gradients at each iteration. As there exist randomness, the distribution of the HVAC cost yield by the two methods are compared under 100 randomly generated scenarios as shown in Fig. 5 (*Case I*), Fig. 6 (*Case II*) and Fig. 7 (*Case III*). The results imply that the average performance (HVAC cost) gaps are about 11.7% (*Case I*), $12.9\ \%$ (*Case II*) and $6.5\ \%$ (*Case III*), respectively. The performance gaps are attributed to: *i)* the discretization of state variables while learning the policies by the proposed method (no discretization in the MPC method); *ii)* the relatively small number of sample paths used in our proposed method, which are related to the estimation accuracy of performance gradients and the state-action pairs that can be sampled. Generally, the performance gap can be reduced by decreasing the discretization step-size and increasing the number of sample paths. This can be illustrated by inspecting the results of *Case II* and *Case III*. Specifically, a reduction of the average HVAC cost under the finer discretization of the state variables can be observed in *Case III*. However, this case generally requires higher computation cost due to the larger space space and the increasing number of sample paths used. Besides, through *Case I* and *Case II*, we imply that the proposed method can be applied to the cases with fixed or controllable set-point temperature for FAU and FCU.

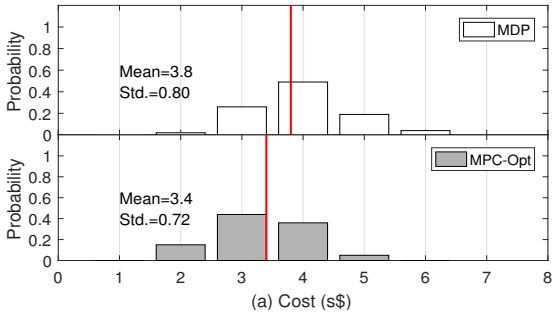


Fig. 5. The histograms of HVAC cost for *Case I*.

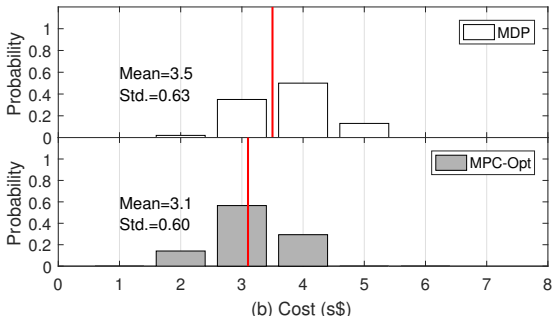


Fig. 6. The histograms of HVAC cost for *Case II*.

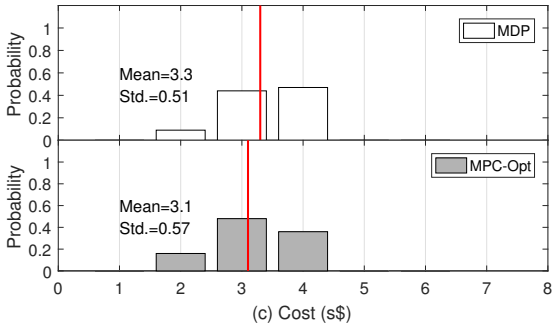


Fig. 7. The histograms of HVAC cost for *Case III*.

We further investigate the two methods by inspecting the results of *Case II*. First, we evaluate the indoor thermal condition under the two methods through a randomly selected scenario. As shown in Fig. 8, the indoor temperature and relative humidity are both maintained in the typical comfortable ranges $[24, 27]°\text{C}$ and $[40, 70]\%$. The indoor thermal comfort under the two methods are also confirmed by inspecting the PMV metrics in Fig. 9. For this scenario, the control of FAU and FCU under the two methods are contrasted in Fig. 10 and Fig 11. We observe that the operation curves of FAU and FCU correspond well under the two methods. That further demonstrates the preferable performance of the proposed method in the paper. Besides, one may observe some interesting phenomenon from the results. *First*, compared with the FCUs, the FAUs are mostly operated in quite lower level

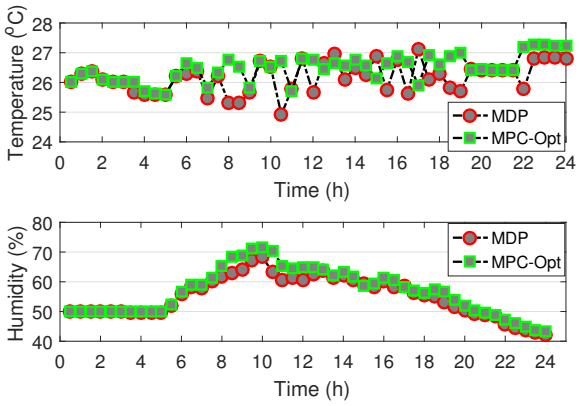


Fig. 8. The indoor temperature and relative humidity curves for a randomly selected scenario under the two methods.

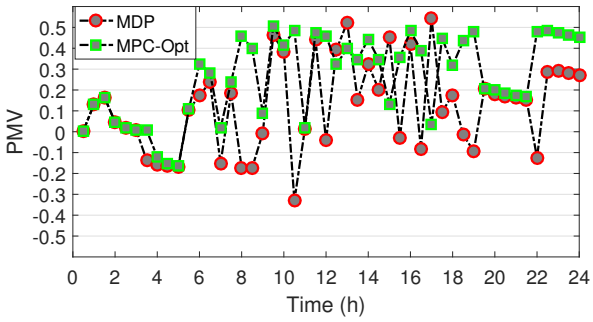


Fig. 9. The PMV curves for a randomly selected scenario under the two methods.

(low fresh air flow rate). The phenomenon is rational as the temperature for the outdoor fresh air is generally higher than that of the recirculated air. Therefore, the HVAC system tends to regulate the FCUs to satisfy the thermal demand while saving energy. This can be further illustrated by comparing the operation patterns of FCU and FAU under each of the two methods. We observe that at the early morning (0:00-10:00), the FAUs are usually operated at a relatively higher level, however, the peak operation time period for the FCUs are 12:00-18::00. Analogously, these phenomena result from the typical weather patterns (i.e., the outdoor temperature is

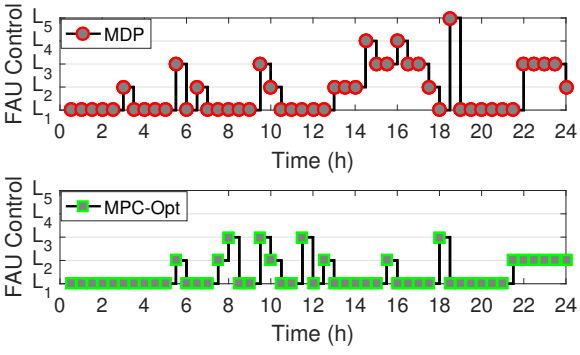


Fig. 10. The control of FAU for a randomly selected scenario under the two methods.

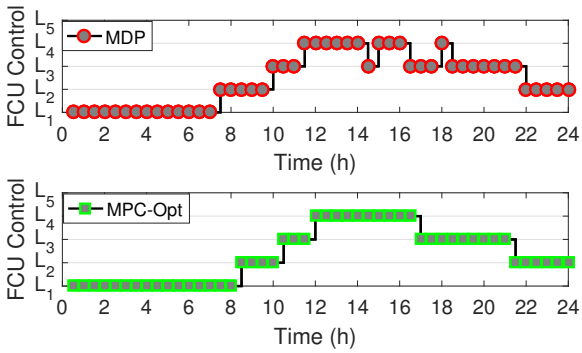


Fig. 11. The control of FAU for a randomly selected scenario under the two methods.

usually relative low at the early morning and late night but tends to arrive at the peak during noon to the afternoon) and the occupancy patterns (i.e., there tends to appear high occupancy during the working hours).

Besides, we investigate the convergence rate and the structures of policies yield by GBPI in *Case II*. First, we inspect the convergence rate of the GBPI in learning the policies. Take the average HVAC cost under the 100 scenarios as an indicator, the convergence rate of the GBPI is exhibited in Fig. 12. We can imply the quite favorable convergence rate of the proposed method as the sub-optimal control policy is approached within about 10 iterations. Besides, as an example, we investigate the random policy (mapping from the state space to the action space) at 9:00 a.m. in Fig. 13 (here only shows the policies for the states that appear in the sample paths). We find that for most of those states, the corresponding probability distributions for their actions are concentrated around some small groups of actions. Exceptionally, we only observe a small group of states, i.e., $\{171, 296, 297, 321, 322, 696, 697, \cdots, \}$ whose probability distributions for their actions are relatively scattered. This phenomenon is mainly attributed to the low occurrence for those states appearing in the sample paths, which results in fewer updates for those state-action pairs while performing the GBPI to learn the policy (at the beginning, we initialize the policy as uniform distribution). This can be confirmed by investigating the occurrence of those states in the 2000 sample paths as shown in Fig. 14. However, as those states are supposed to occur rarely in practice (otherwise they will appear with high probability in the sample paths), the policies of the GBPI can still guarantee a favorable performance.

V. Conclusion

This paper studied the agile, adaptive and energy-efficient control the HVAC system to manage the uncertain thermal load caused by the weather and dynamic occupancy via Markov decision process (MDP). To cope with the computational challenges, a gradient-based policy iteration (GBPI) method based on performance gradients is proposed, and the convergence of the method was investigated. The main advantages of the proposed method against the literature are that *i)* it can be implemented off-line while circumventing the computational

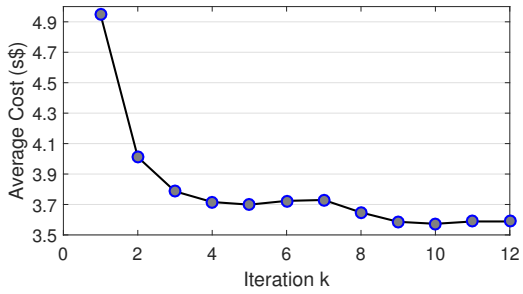


Fig. 12. The convergence rate of GBPI (*Case II*)

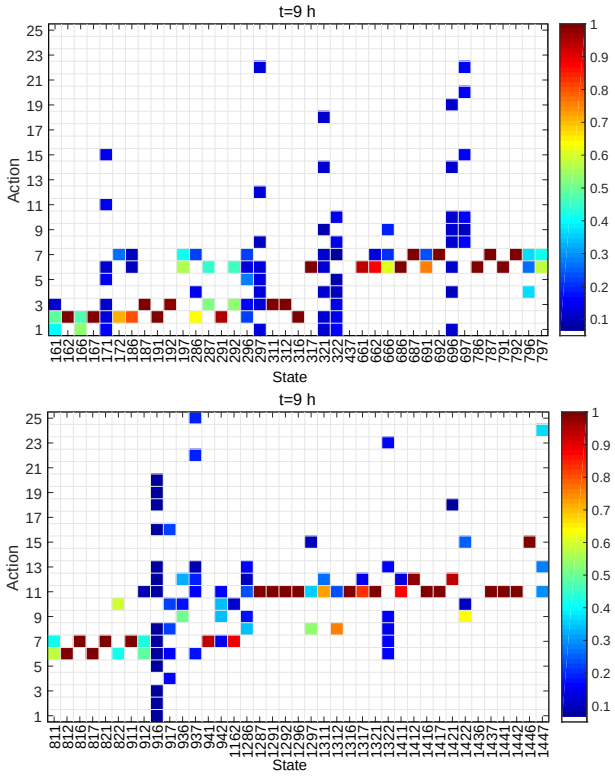


Fig. 13. The random policy at 9:00 a.m. for Case II).

intensity of most existing on-line methods; *ii)* it's convenient and efficient in tacking the intrinsic non-linear and non-convex system dynamics and the non-analytical elaborate thermal comfort model, such as PMV. The performance of the proposed method was illustrated though comparison with the optimal solutions obtained with assumed accurate information. The results implied the favorable performance of the proposed method in optimizing the HVAC cost while maintaining the thermal comfort indicated by the PMV metric.

This paper mainly investigated the properformance of the proposed method on HVAC control for single-room case. The extensions to multi-rooms which share a central HVAC system is trivial and the proposed methd can be implemented in distributed manner. However, for multi-zone cases where there exist thermal couplings due to heat transfer, there still exist challenging issues to be investigated.

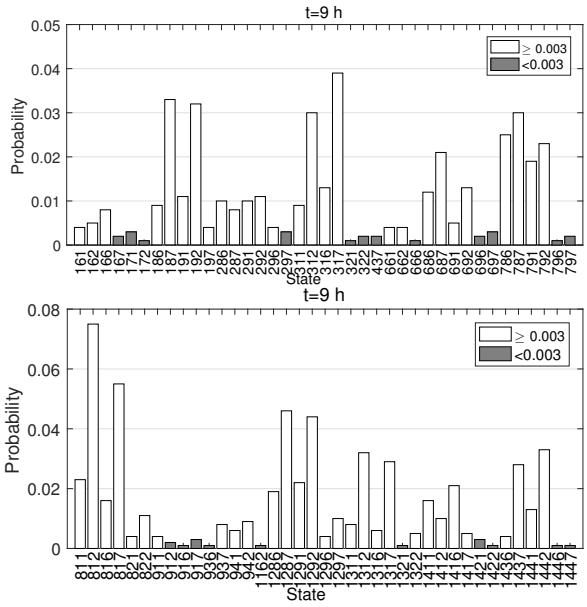


Fig. 14. The state distribution at 9:00 a.m. for *Case II*.

Appendix A
Proof of the Convergence of GBPI

Proof. According to (14), we have the following performance difference equation for the policies over two successive iterations:

$$\begin{aligned}
&J(\boldsymbol{\sigma}^{k+1};S_0)-J(\boldsymbol{\sigma}^{k};S_0)\\
&=\sum_{t=0}^{T-1}\pi_t^{\boldsymbol{\sigma}^{k+1}}\Big[(\boldsymbol{r}_t^{\boldsymbol{\sigma}^{k+1}}-\boldsymbol{r}_t^{\boldsymbol{\sigma}^{k}})+(\boldsymbol{P}_t^{\boldsymbol{\sigma}^{k+1}}-\boldsymbol{P}_t^{\boldsymbol{\sigma}^{k}})\boldsymbol{V}_{t+1}^{\boldsymbol{\sigma}^{k}}\Big]\\
&=\sum_{t=0}^{T-1}\sum_{s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}}\pi_t^{\boldsymbol{\sigma}^{k+1}}(s_t)\Big[(\boldsymbol{r}_t^{\boldsymbol{\sigma}^{k+1}}(s_t)-\boldsymbol{r}_t^{\boldsymbol{\sigma}^{k}}(s_t))\\
&\quad+\sum_{s_{t+1}\in\{1,2,\cdots,|\mathcal{S}_{t+1}|\}}(\boldsymbol{P}_t^{\boldsymbol{\sigma}^{k+1}}(s_{t+1}|s_t)-\boldsymbol{P}_t^{\boldsymbol{\sigma}^{k}}(s_{t+1}|s_t))V_{t+1}^{\boldsymbol{\sigma}^{k}}(s_{t+1})\Big]
\end{aligned}\tag{25}$$

For notation, we define the operator $\Sigma^k(s_t)=\sum_{a_t=1}^{|\mathcal{A}_t|}\sigma_t(s_t,a_t)$, thus we have

$$\begin{aligned}
r_t^{\boldsymbol{\sigma}^{k+1}}(s_t)-r_t^{\boldsymbol{\sigma}^{k}}(s_t)&=\sum_{b_t=1}^{|\mathcal{A}_t|}\big[p_t^{\boldsymbol{\sigma}^{k+1}}(b_t|s_t)-p_t^{\boldsymbol{\sigma}^{k}}(b_t|s_t)\big]r_t(s_t,b_t)\\
&=\sum_{b_t=1}^{|\mathcal{A}_t|}\Big[\frac{\sigma_t^{k+1}(s_t,b_t)}{\Sigma^{k+1}(s_t)}-\frac{\sigma_t^{k}(s_t,b_t)}{\Sigma^{k}(s_t)}\Big]r_t(s_t,b_t)
\end{aligned}\tag{26}$$

$$\begin{aligned}
&\boldsymbol{P}_t^{\boldsymbol{\sigma}^{k+1}}(s_{t+1}|s_t)-\boldsymbol{P}_t^{\boldsymbol{\sigma}^{k}}(s_{t+1}|s_t)\\
&=\sum_{b_t=1}^{|\mathcal{A}_t|}p_t(s_{t+1}|s_t,b_t)(p_t^{\boldsymbol{\sigma}^{k+1}}(b_t|s_t)-p_t^{\boldsymbol{\sigma}^{k}}(b_t|s_t))\\
&=\sum_{b_t=1}^{|\mathcal{A}_t|}p_t(s_{t+1}|s_t,b_t)\Big[\frac{\sigma_t^{k+1}(s_t,b_t)}{\Sigma^{k+1}(s_t)}-\frac{\sigma_t^{k}(s_t,b_t)}{\Sigma^{k}(s_t)}\Big]
\end{aligned}\tag{27}$$

By substituting (26) and (27) into (25), we have

$$\begin{aligned}J(\boldsymbol{\sigma}^{k+1};S_0)-J(\boldsymbol{\sigma}^{k};S_0)=&\sum_{t=0}^{T-1}\sum_{s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}}\Big[\sum_{b_t=1}^{|\mathcal{A}_t|}[\frac{\sigma_t^{k+1}(s_t,b_t)}{\Sigma^{k+1}(s_t)}\\&-\frac{\sigma_t^{k}(s_t,b_t)}{\Sigma^{k}(s_t)}]\big(r_t(s_t,b_t)+V_t^{\boldsymbol{\sigma}^k}(s_t,b_t)\big)\Big]\\=&\sum_{t=0}^{T-1}\sum_{s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}}\sum_{b_t=1}^{|\mathcal{A}_t|}[\frac{\sigma_t^{k+1}(s_t,b_t)}{\Sigma^{k+1}(s_t)}-\frac{\sigma_t^{k}(s_t,b_t)}{\Sigma^{k}(s_t)}]A^{\boldsymbol{\sigma}^k}(s_t,b_t)\end{aligned}\tag{28}$$

where we define $A_t^{\boldsymbol{\sigma}^k}(s_t,b_t)=r_t(s_t,b_t)+V_t^{\boldsymbol{\sigma}^k}(s_t,b_t)$.

From (21), we have

$$\sigma^{k+1}(s_t,a_t)=\sigma_t^k(s_t,a_t)-\gamma^k(s_t,a_t)\Delta\sigma_t^k(s_t,a_t),\quad\forall a_t\in\{1,2,\cdots,|\mathcal{A}_t|\}.$$

with $\Delta\sigma_t^k(s_t,a_t)=\dfrac{\partial J(\boldsymbol{\sigma}^k;S_0)}{\partial\sigma_t^k(s_t,a_t)}$

$$\begin{aligned}&=\pi_t^{\boldsymbol{\sigma}^k}(s_t)\Big[\sum_{b_t=1}^{|\mathcal{A}_t|}\frac{\partial p^{\boldsymbol{\sigma}^k}(b_t|s_t)}{\partial\sigma_t^k(s_t,a_t)}\big(r_t(s_t,b_t)+V_t^{\sigma^k}(s_t,b_t)\big)\Big]\\&=\pi_t^{\sigma^k}(s_t)\Big[\sum_{b_t=1,b_t\neq a_t}^{|\mathcal{A}_t|}\frac{-\sigma_t^k(s_t,b_t)}{[\Sigma^k(s_t)]^2}A_t^{\boldsymbol{\sigma}^k}(s_t,b_t)\\&\qquad+\frac{\Sigma^k(s_t)-\sigma_t^k(s_t,a_t)}{[\Sigma^k(s_t)]^2}A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)\Big]\\&=\frac{\pi_t^{\sigma^k}(s_t)}{[\Sigma^k(s_t)]^2}[\Sigma^k(s_t)A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)-\sum_{b_t=1}^{|\mathcal{A}_t|}\sigma^k(s_t,b_t)A_t^{\boldsymbol{\sigma}^k}(s_t,b_t)]\end{aligned}$$

Besides, we have

$$\begin{aligned}&\frac{\sigma_t^{k+1}(s_t,a_t)}{\Sigma^{k+1}(s_t)}-\frac{\sigma_t^k(s_t,a_t)}{\Sigma^k}\\&=\frac{\Delta\sigma_t^k(s_t,a_t)\cdot\Sigma^k(s_t)+\Delta\Sigma^k(s_t)\cdot\sigma_t^k(s_t,a_t)}{\Sigma^k(s_t)\cdot\Sigma^{k+1}(s_t)}\end{aligned}\tag{29}$$

where $\Delta\Sigma^k(s_t)=\Sigma^{k+1}(s_t)-\Sigma^k(s_t)$.

If the step-sizes are set as $\gamma^k(s_t,b_t)=\frac{\sigma_t^k(s_t,b_t)}{\Sigma^k(s_t)}$ ($\forall b_t\in\{1,2,\cdots,|\mathcal{A}_t|\}$), we have

$$\begin{aligned}\Delta\Sigma^k(s_t)=&\sum_{b_t=1}^{|\mathcal{A}_t|}\boldsymbol{\gamma}_t^k(s_t,b_t)\Delta\sigma_t^k(s_t,b_t)\\=&\frac{\pi_t^{\sigma^k}(s_t)}{[\Sigma^k(s_t)]^2}\Big[\sum_{b_t=1}^{|\mathcal{A}_t|}\gamma^k(s_t,b_t)\Sigma^k(s_t)A_t^{\boldsymbol{\sigma}^k}(s_t,b_t)\\&-\sum_{b_t=1}^{|\mathcal{A}_t|}\gamma^k(s_t,b_t)\sum_{b_t=1}^{|\mathcal{A}_t|}\sigma^k(s_t,b_t)A^{\boldsymbol{\sigma}^k}(s_t,b_t)\Big]\\=&0\end{aligned}\tag{30}$$

Thus, we can imply from (30) that we have $\Sigma^k(s_t)=\Sigma^{k+1}(s_t)$ ($\forall s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}$) with the step-size $\gamma^k(s_t,b_t)=\frac{\sigma_t^k(s_t,b_t)}{\Sigma^k(s_t)}$ ($\forall s_t\in\{1,2,\cdots,|\mathcal{S}_t|\},b_t\in\{1,2,\cdots,|\mathcal{A}_t|\}$).

Then by combining (30) with (29), we have

$$\begin{aligned}&\frac{\sigma_t^{k+1}(s_t,a_t)}{\Sigma^{k+1}(s_t)}-\frac{\sigma_t^k(s_t,a_t)}{\Sigma^k(s_t)}=\frac{-\gamma^k(s_t,a_t)\Delta\sigma_t^k(s_t,a_t)}{\Sigma^k(s_t)}\\&=\frac{-\pi_t^{\sigma^k}(s_t)}{[\Sigma^k(s_t)]^3}\Big[\Sigma^k(s_t)\gamma_t^k(s_t,a_t)A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)\\&\qquad-\gamma^k(s_t,a_t)\sum_{b_t=1}^{|\mathcal{A}_t|}\sigma^k(s_t,b_t)A_t^{\boldsymbol{\sigma}^k}(s_t,b_t)\Big]\end{aligned}\tag{31}$$

By substituting (31) into (28), we have

$$\begin{aligned}&J(\boldsymbol{\sigma}^{k+1};S_0)-J(\boldsymbol{\sigma}^{k};S_0)\\&=\sum_{t=0}^{T-1}\sum_{s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}}\sum_{a_t=1}^{|\mathcal{A}_t|}\Big\{\frac{-\pi_t^{\sigma^k}(s_t)}{[\Sigma^k(s_t)]^3}\Big[\Sigma^k(s_t)\gamma^k(s_t,a_t)A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)\\&\qquad-\gamma^k(s_t,a_t)\sum_{a_t=1}^{|\mathcal{A}_t|}\sigma_t^k(s_t,a_t)A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)\Big]A_t^{\sigma^k}(s_t,a_t)\Big\}\\&=\sum_{t=0}^{T-1}\sum_{s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}}\frac{-\pi_t^{\sigma^k}(s_t)}{[\Sigma^k(s_t)]^2}\Big[\sum_{a_t=1}^{|\mathcal{A}_t|}\sigma_t^k(s_t,a_t)[A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)]^2\\&-\sum_{a_t=1}^{|\mathcal{A}_t|}\frac{\sigma^k(s_t,a_t)}{\Sigma^k(s_t)}A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)\sum_{b_t=1}^{|\mathcal{A}_t|}\sigma_t^k(s_t,b_t)A_t^{\boldsymbol{\sigma}^k}(s_t,b_t)\Big]\end{aligned}\tag{32}$$

We set $\Sigma^0=1$ while initialize the policy σ^0 before the iteration (trivial in practice), thus we have

$$\begin{aligned}&J(\boldsymbol{\sigma}^{k+1};S_0)-J(\boldsymbol{\sigma}^{k};S_0)\\&=\sum_{t=0}^{T-1}\sum_{s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}}\sum_{a_t=1}^{|\mathcal{A}_t|}\Big\{\frac{-\pi_t^{\sigma^k}(s_t)}{[\Sigma^k(s_t)]^3}\Big[\Sigma^k(s_t)\gamma_t^k(s_t,a_t)A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)\\&\qquad-\gamma^k(s_t,a_t)\sum_{a_t=1}^{|\mathcal{A}_t|}\sigma_t^k(s_t,a_t)A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)\Big]A_t^{\sigma^k}(s_t,a_t)\Big\}\\&=\sum_{t=0}^{T-1}\sum_{s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}}\frac{-\pi_t^{\sigma^k}(s_t)}{[\Sigma^k(s_t)]^2}M(s_t)\end{aligned}\tag{33}$$

where we have $M(s_t)=\sum_{a_t=1}^{|\mathcal{A}_t|}\sigma_t^k(s_t,a_t)[A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)]^2-\sum_{a_t=1}^{|\mathcal{A}_t|}\sigma_t^k(s_t,a_t)A_t^{\boldsymbol{\sigma}^k}(s_t,a_t)\sum_{b_t=1}^{|\mathcal{A}_t|}\sigma_t^k(s_t,b_t)A_t^{\boldsymbol{\sigma}^k}(s_t,b_t)$.

As function $f(x)=x^2$ is convex and we have $\sum_{a_t=1}^{|\mathcal{A}_t|}\sigma_t^k(s_t,a_t)=\Sigma^k$, it's easy to figure out that $M(s_t)\geq 0$ ($\forall s_t\in\{1,2,\cdots,|\mathcal{S}_t|\}$) based on the definition of convex functions.

Thus, we conclude that

$$J(\boldsymbol{\sigma}^{k+1};S_0)-J(\boldsymbol{\sigma}^{k};S_0)\leq 0$$

The above illustrates the non-increasing characteristics of the performance function $J(\boldsymbol{\sigma}^k;S_0)$ w.r.t. iteration k. As problem (13) is bounded (e.g., the feasible action space is bounded), we imply that the GBPI (**Algorithm** 2) will converge to a local optima $\bar{\boldsymbol{\sigma}}$ with $k\to\infty$. The remainder illustrates that the method will converge to the global optima if existed.

We assume at least one feasible policy exist for problem (13) and denoted by $\boldsymbol{\sigma}^*$. It is straightforward that we have

$$J(\boldsymbol{\sigma}^*;S_0)\leq J(\overline{\boldsymbol{\sigma}};S_0).$$

We assume the random parameterized policy $\boldsymbol{\sigma}^{\delta} = (1-\delta)\bar{\boldsymbol{\sigma}} + \delta\boldsymbol{\sigma}^{*}$ ($\delta \in [0,1]$) is adopted. Based on (28) ($\Sigma_k = 1,\ \forall k \in \mathbb{N}$), we have

$$\begin{aligned} &J(\boldsymbol{\sigma}^{r}; S_0) - J(\overline{\boldsymbol{\sigma}}; S_0) \\ &= \sum_{t=0}^{T-1} \sum_{s_t \in \{1,2,\cdots,|\mathcal{S}_t|\}} \Big[\sum_{a_t=1}^{|\mathcal{A}_t|} (\sigma_t^{r}(s_t,a_t) - \bar{\sigma}_t(s_t,a_t)) A_t^{\bar{\sigma}}(s_t,a_t) \Big] \\ &= \sum_{t=0}^{T-1} \sum_{s_t \in \{1,2,\cdots,|\mathcal{S}_t|\}} \Big[\sum_{a_t=1}^{|\mathcal{A}_t|} \delta \cdot (\boldsymbol{\sigma}^{*}(s_t,a_t) - \bar{\sigma}_t(s_t,a_t)) A_t^{\bar{\sigma}}(s_t,a_t) \Big] \end{aligned} \tag{34}$$

As $\overline{\boldsymbol{\sigma}}$ is a local optima, we have

$$J(\boldsymbol{\sigma}^{r}; S_0) - J(\overline{\boldsymbol{\sigma}}; S_0) \geq 0 \tag{35}$$

On the other hand, as $\boldsymbol{\sigma}^{*}$ is the global optima, we have

$$\begin{aligned} &J(\boldsymbol{\sigma}^{*}; S_0) - J(\overline{\boldsymbol{\sigma}}; S_0) \\ &= \sum_{t=0}^{T-1} \sum_{s_t \in \{1,2,\cdots,|\mathcal{S}_t|\}} \Big[\sum_{a_t=1}^{|\mathcal{A}_t|} (\sigma_t^{*}(s_t,a_t) - \bar{\sigma}_t(s_t,a_t)) A_t^{\bar{\sigma}}(s_t,a_t) \Big] \\ &\leq 0 \end{aligned} \tag{36}$$

Thus, we have $\bar{\boldsymbol{\sigma}} = \boldsymbol{\sigma}^{*}$, otherwise (34) and (36) contradict with each other. This implies that the GBPI method can converge to the global optima of problem (13). □

REFERENCES

[1] K. Ku, J. Liaw, M. Tsai, T. Liu, *et al.*, "Automatic control system for thermal comfort based on predicted mean vote and energy saving.," *IEEE Trans. Automation Science and Engineering*, vol. 12, no. 1, pp. 378–383, 2015.

[2] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and buildings*, vol. 40, no. 3, pp. 394–398, 2008.

[3] Z. Afroz, G. Shafiullah, T. Urmee, and G. Higgins, "Modeling techniques used in building hvac control systems: A review," *Renewable and Sustainable Energy Reviews*, vol. 83, pp. 64–84, 2018.

[4] A. Afram and F. Janabi-Sharifi, "Theory and applications of hvac control systems–a review of model predictive control (mpc)," *Building and Environment*, vol. 72, pp. 343–355, 2014.

[5] D. Yan, J. Xia, W. Tang, F. Song, X. Zhang, and Y. Jiang, "Dest—an integrated building simulation toolkit part i: Fundamentals," in *Building Simulation*, vol. 1, pp. 95–110, Springer, 2008.

[6] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher, M. J. Witte, *et al.*, "Energyplus: creating a new-generation building energy simulation program," *Energy and buildings*, vol. 33, no. 4, pp. 319–331, 2001.

[7] A. Afram, F. Janabi-Sharifi, A. S. Fung, and K. Raahemifar, "Artificial neural network (ann) based model predictive control (mpc) and optimization of hvac systems: A state of the art review and case study of a residential hvac system," *Energy and Buildings*, vol. 141, pp. 96–113, 2017.

[8] H. Huang, L. Chen, and E. Hu, "A neural network-based multi-zone modelling approach for predictive control system design in commercial buildings," *Energy and buildings*, vol. 97, pp. 86–97, 2015.

[9] Z. Wu, Q.-S. Jia, and X. Guan, "Optimal control of multiroom hvac system: An event-based approach," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 2, pp. 662–669, 2016.

[10] B. Sun, P. B. Luh, Q.-S. Jia, Z. Jiang, F. Wang, and C. Song, "Building energy management: Integrated control of active and passive heating, cooling, lighting, shading, and ventilation systems," *IEEE Transactions on automation science and engineering*, vol. 10, no. 3, pp. 588–602, 2013.

[11] Z. Xu, G. Hu, C. J. Spanos, and S. Schiavon, "Pmv-based event-triggered mechanism for building energy management under uncertainties," *Energy and Buildings*, vol. 152, pp. 73–85, 2017.

[12] M. Maasoumy, A. Pinto, and A. Sangiovanni-Vincentelli, "Model-based hierarchical optimal control design for hvac systems," 2014.

[13] A. Kelman and F. Borrelli, "Bilinear model predictive control of a hvac system using sequential quadratic programming," in *Ifac world congress*, vol. 18, pp. 9869–9874, 2011.

[14] Z. Xu, Q.-S. Jia, and X. Guan, "Supply demand coordination for building energy saving: Explore the soft comfort," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 2, pp. 656–665, 2015.

[15] M. Killian, B. Mayer, and M. Kozek, "Cooperative fuzzy model predictive control for heating and cooling of buildings," *Energy and Buildings*, vol. 112, pp. 130–140, 2016.

[16] F. Jazizadeh, A. Ghahramani, B. Becerik-Gerber, T. Kichkaylo, and M. Orosz, "User-led decentralized thermal comfort driven hvac operations for improved efficiency in office buildings," *Energy and Buildings*, vol. 70, pp. 398–410, 2014.

[17] N. Nassif, S. Kajl, and R. Sabourin, "Optimization of hvac control system strategy using two-objective genetic algorithm," *HVAC&R Research*, vol. 11, no. 3, pp. 459–486, 2005.

[18] S. Wang and X. Jin, "Model-based optimal control of vav air-conditioning system using genetic algorithm," *Building and Environment*, vol. 35, no. 6, pp. 471–487, 2000.

[19] J. Brooks, S. Kumar, S. Goyal, R. Subramany, and P. Barooah, "Energy-efficient control of under-actuated hvac zones in commercial buildings," *Energy and Buildings*, vol. 93, pp. 160–168, 2015.

[20] J. Brooks, S. Goyal, R. Subramany, Y. Lin, T. Middelkoop, L. Arpan, L. Carloni, and P. Barooah, "An experimental investigation of occupancy-based energy-efficient control of commercial building indoor climate," in *53rd IEEE Conference on Decision and Control*, pp. 5680–5685, IEEE, 2014.

[21] J. H. Yoon, R. Baldick, and A. Novoselac, "Dynamic demand response controller based on real-time retail

price for residential buildings," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 121–129, 2014.

[22] P. O. Fanger *et al.*, "Thermal comfort. analysis and applications in environmental engineering.," *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.

[23] M. Klauco and M. Kvasnica, "Explicit mpc approach to pmv-based thermal comfort control.," in *CDC*, pp. 4856–4861, 2014.

[24] J. Cigler, S. Prívara, Z. Váňa, E. Žáčeková, and L. Ferkl, "Optimization of predicted mean vote index within model predictive control framework: Computationally tractable solution," *Energy and Buildings*, vol. 52, pp. 39–49, 2012.

[25] M. Pčolka, E. Žáčeková, R. Robinett, S. Čelikovskỳ, and M. Šebek, "Bridging the gap between the linear and nonlinear predictive control: Adaptations for efficient building climate control," *Control Engineering Practice*, vol. 53, pp. 124–138, 2016.

[26] S. Goyal, H. A. Ingley, and P. Barooah, "Effect of various uncertainties on the performance of occupancy-based optimal control of hvac zones.," in *CDC*, pp. 7565–7570, 2012.

[27] Y. Ma, A. Kelman, A. Daly, and F. Borrelli, "Predictive control for energy efficient buildings with thermal storage: Modeling, stimulation, and experiments," *IEEE Control Systems*, vol. 32, no. 1, pp. 44–64, 2012.

[28] Y. Ma, J. Matusko, and F. Borrelli, "Stochastic model predictive control for building hvac systems: Complexity and conservatism.," *IEEE Trans. Contr. Sys. Techn.*, vol. 23, no. 1, pp. 101–116, 2015.

[29] B. Kouvaritakis and M. Cannon, "Stochastic model predictive control," *Encyclopedia of Systems and Control*, pp. 1350–1357, 2015.

[30] F. Oldewurtel, C. N. Jones, A. Parisio, and M. Morari, "Stochastic model predictive control for building climate control," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 3, pp. 1198–1205, 2013.

[31] A. Parisio, L. Fabietti, M. Molinari, D. Varagnolo, and K. H. Johansson, "Control of hvac systems via scenario-based explicit mpc," in *53rd IEEE conference on decision and control*, pp. 5201–5207, IEEE, 2014.

[32] M. Klaučo and M. Kvasnica, "Explicit mpc approach to pmv-based thermal comfort control," in *53rd IEEE conference on decision and control*, pp. 4856–4861, IEEE, 2014.

[33] Y. Ma and F. Borrelli, "Fast stochastic predictive control for building temperature regulation," in *American Control Conference (ACC), 2012*, pp. 3075–3080, IEEE, 2012.

[34] Y. Ma, S. Vichik, and F. Borrelli, "Fast stochastic mpc with optimal risk allocation applied to building control systems," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 7559–7564, IEEE, 2012.

[35] X. Zhang, G. Schildbach, D. Sturzenegger, and M. Morari, "Scenario-based mpc for energy-efficient building climate control under weather and occupancy uncertainty," in *2013 European Control Conference (ECC)*, pp. 1029–1034, IEEE, 2013.

[36] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[37] B. Sun, P. B. Luh, Q.-S. Jia, Z. Jiang, F. Wang, and C. Song, "An integrated control of shading blinds, natural ventilation, and hvac systems for energy saving and human comfort," in *Automation Science and Engineering (CASE), 2010 IEEE Conference on*, pp. 7–14, IEEE, 2010.

[38] Z. Wu, Q.-S. Jia, and X. Guan, "Optimal control of multiroom hvac system: An event-based approach," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 2, pp. 662–669, 2015.

[39] Q.-S. Jia, J. Wu, Z. Wu, and X. Guan, "Event-based hvac control—a complexity-based approach," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 4, pp. 1909–1919, 2018.

[40] N. Radhakrishnan, S. Srinivasan, R. Su, and K. Poolla, "Learning-based hierarchical distributed hvac scheduling with operational constraints," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 5, pp. 1892–1900, 2017.

[41] R. Jia, R. Dong, S. S. Sastry, and C. J. Sapnos, "Privacy-enhanced architecture for occupancy-based hvac control," in *Cyber-Physical Systems (ICCPS), 2017 ACM/IEEE 8th International Conference on*, pp. 177–186, IEEE, 2017.

[42] W. Shen, G. Newsham, and B. Gunay, "Leveraging existing occupancy-related data for optimal control of commercial office buildings: A review," *Advanced Engineering Informatics*, vol. 33, pp. 230–242, 2017.

[43] M. L. Puterman, "Markov decision processes: Discrete stochastic dynamic programming," 1994.

[44] Y. Zhao, *Optimization theories and methods for Markov decision processes in resource scheduling of networked systems*. PhD thesis, Tsinghua University, 2010.

[45] Q.-S. Jia, J.-X. Shen, Z. Xu, and X. Guan, "Simulation-based policy improvement for energy management in commercial office buildings.," *IEEE Trans. Smart Grid*, vol. 3, no. 4, pp. 2211–2223, 2012.

[46] "Thermal Environmental Conditions for Human Occupancy," standard, Standing Standard Project Committee (SSPC), Mar. 2017.