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—The heating, ventilation, and air-conditioning (HVAC) system account for substantial energy use in buildings, whereas a large group of occupants is still not actually feeling comfortable staying inside. This poses the issue of developing energy-efficient HVAC control, i.e., reduce energy use (cost) while simultaneously enhancing human comfort. This brief pursues the objective and studies the stochastic optimal HVAC control subject to uncertain thermal demand (i.e., the weather and occupancy). Particularly, we involve the elaborate predicted mean vote (PMV) thermal comfort model in the optimization. The problem is computationally challenging due to the nonlinear and nonanalytical constraints imposed by the system dynamics and PMV model. We make the following contributions to address it. First, we formulate the problem as a Markov decision process (MDP) which is a desirable modeling technique capable of handling the complexities. Second, we propose a gradient-based learning (GB-L) method for progressively learning a stochastic control policy off-line and store it for on-line execution. Third, we prove the learning method's converge to the optimal policies theoretically, and its performance (i.e., energy cost, thermal comfort, and on-line computation) for HVAC control via simulations. The comparisons with the existing model predictive control based relaxation (MPC-R) method which is assumed with accurate future information and supposed to provide the near-optimal bounds show that though there exists some discount in energy cost reduction (i.e., 6.5%), the proposed method can enable efficient on-line implementation (less than 1 s) and provide a high probability of thermal comfort under uncertainties.

Index Terms—Energy-efficient, heating, ventilation and air-conditioning (HVAC) control, Markov decision process (MDP), predicted mean vote (PMV), stochastic policy.

NOMENCLATURE

t	Time index.
α_w	The absorption coefficient of walls.
A_{gs}	The area of glass window [m^2].
$A_{\text{wl}}/A_{\text{wr}}$	The area of left/right wall [m^2].
C_p	The air specific heat [J/(kg · K)].
C_w	The wall capacity [J/(kg · K)].
c_t	The electricity price [s\$/kW].
η	The reciprocal of coefficient of performance (COP) of chiller.
$G_t^{\text{fau}}/G_t^{\text{fcu}}$	The supply air flow rate of FAU/FCU [kgs^{-1}].
$G^{\text{fcu},r}$	The nominal air flow rate of FCU [kgs^{-1}].
$G^{\text{fau},r}$	The nominal air flow rate of FAU [kgs^{-1}].
$\underline{G}^{\text{fau}}/\underline{G}^{\text{fcu}}$	The lower bound of damper position for FAU/FCU [kgs^{-1}].
$\overline{G}^{\text{fau}}/\overline{G}^{\text{fcu}}$	The upper bound of damper position for FAU/FCU [kgs^{-1}].
H_t^o/H_t^a	Outdoor/indoor relative humidity [%].
$H_t^{\text{fau}}/H_t^{\text{fcu}}$	Relative humidity of supply air by FAU/FCU [%].
h_{gs}/h_w	Heat transfer coefficient of window/walls [J/(m^2 · °C)].
m^a	The mass of indoor air [kg].
$m^{\text{wl}}/m^{\text{wr}}$	The mass of left/right wall [kg].
$F^{\text{fcu},r}$	The nominal fan power of FCU [kW].
$F^{\text{fau},r}$	The nominal fan power of FAU [kW].
Q^o/H^g	The average internal heat/humidity generation rate per occupant [Js^{-1}].
Q^{d}	The average heat generation rate of electrical devices per occupant [Js^{-1}].
Q_t^w	Solar radiation density [J/m^2 · s].
T_t^o/T_t^a	Outdoor/indoor temperature [°C].
$T_t^{\text{wl}}/T_t^{\text{wr}}$	The left/right wall temperature [°C].
$T_t^{\text{fau}}/T_t^{\text{fcu}}$	The set-point temperature of FAU/FCU [°C].
$\underline{T}^{\text{fau}}/\underline{T}^{\text{fcu}}$	The lower bound of set-point temperature of FAU/FCU [°C].
$\overline{T}^{\text{fau}}/\overline{T}^{\text{fcu}}$	The upper bound of set-point temperature of FAU/FCU [°C].

I. INTRODUCTION

HEATING, ventilation, and air-conditioning (HVAC) system accounts for 40%–50% of energy use in buildings for providing occupant comfort [1], [2]. Whereas a large group is still not actually feeling comfortable staying inside [3]. This negative situation is substantially attributed to the insensible operation of HVAC systems, such as settled ON–OFF switch and fixed thermostat settings, and so on. Fortunately, the advances of information and communication technology (ICT) including smart sensing, data storing, and processing technologies are likely to turn it around [4]. Advanced HVAC control is expected for energy-efficient buildings, i.e., reduce energy use (cost) while simultaneously enhancing human comfort [5], [6].

A. Literature

The recent decades have seen massive progress towards improving building energy efficiency. Particularly, various

modeling techniques have been established for optimizing HVAC control [7]. However, the *system complexity* and *uncertainties* are still two major challenges we are facing for realization. The HVAC system comprises various heat and mass transfer equipment such as the chiller, the heating/cooling coils and the air-handling equipment, making the overall system dynamics highly nonlinear and nonconvex [7]. Besides, the HVAC control requires to respond to the uncertain thermal disturbances caused by the weather and the occupants, etc.

To address the non-linearity, the typical solution methods include sequential quadratic programming (SQP) [8], mixed-integer programming (MIP) [9], [10] and metaheuristic algorithms (e.g., genetic algorithm [11] and particle swarm optimization [12], and so on). These works have mostly focused on developing approximation or relaxation techniques to deal with the nonlinear system dynamics. Besides, complete information or accurate predictions for dynamic thermal demand are usually assumed for computation (see [8] and [10] for examples).

However, the uncertainties caused by the weather and occupancy cannot be underestimated for practice. Both the weather and indoor occupancy fluctuate over time, affecting the thermal demand to be responded by the HVAC operation. In the literature, the model predictive control (MPC) technique has been extensively discussed for HVAC control, especially deterministic MPC based on short-term predictions [13]. The basic idea of MPC is to rely on a model to predict the future system dynamic process and use the predictions to make a local optimal decision for the current stage. However, MPC is still being rarely used in buildings primarily due to the two obstacles: 1) the lack of an appropriate model to be deployed in the MPC controller due to the system complexity [9] and 2) the difficulty to obtain accurate predictions for future thermal demand [14]. Moreover, deterministic MPC without considering the uncertainties of predictions has been found not work well for HVAC control [15]. Aware of that, stochastic MPC [15], [16] and explicit MPC [17], [18] have been suggested to account for the (prediction) uncertainties. They minimize the average HVAC energy cost and provide thermal comfort with probability guarantees via chance constraints. These MPCs can reduce performance variance over deterministic MPCs, but face the challenges of handling the chance constraints [19], especially accounting for the on-line computation burden. To compensate the deficiency, Parisio *et al.* [17] and Klaučo and Kvasnica [18] discussed explicit MPCs for on-line HVAC control with assumed linear models.

While improving energy efficiency of HVAC systems has raised extensive attention, we are still on the way to achieve the target. On the one hand, a control method that can handle the intrinsic nonlinearity and uncertainties is required. On the other hand, the current practices have mostly used static temperature ranges to indicate human thermal comfort (see [16], [19]), which may not actually provide the thermal comfort warranted [3]; for the occupants' thermal sensation is determined by multiple parameters: indoor air temperature, mean radiant temperature, humidity, air velocity, metabolic activity, and clothing insulation, and so on [20]. Therefore, it necessitates integrating an elaborate thermal comfort model in the optimization. Moreover, we can expect a more reasonable thermal comfort model can help save energy by avoiding overcooling or overheating [1], [9].

B. Our Contributions

To advance energy-efficient buildings, this brief studies the optimal HVAC control subject to uncertain thermal demand (i.e., weather and the occupants). To enhance thermal comfort, we involve the elaborate predicted mean vote (PMV) thermal comfort model [20] in the optimization. Particularly, we deploy a Markov decision process (MDP) formulation [21], [22] which is capable of accommodating the nonlinear system dynamics and the nonanalytical PMV model for HVAC control but with the computational challenges to be addressed. To handle it, we propose a gradient-based learning (GB-L) algorithm to progressively learn the optimal stochastic control policies off-line and store it for on-line execution. We prove the method's convergence to the optimal policies theoretically, and moreover, we demonstrate the performance (i.e., energy cost, thermal comfort, and on-line computation) for HVAC control via simulations. The comparisons with the existing MPC-based relaxation (MPC-R) method which is assumed with accurate future information and supposed to provide the near-optimal bounds, show though there exists some discount in energy cost reduction, the proposed method can enable efficient on-line implementation and high probability of thermal comfort under uncertainties.

The remainder is structured as follows. Section II gives the MDP formulation, Section III introduces the GB-L method and its convergence, Section IV evaluates the method 's performance for HVAC control via simulations, and Section V concludes this brief.

II. PROBLEM

A. HVAC System Configuration

An HVAC system mainly comprises chiller and air handling units, with the latter integrated with cooling/heating coils and driven fan. The HVAC system generally works in the way that the air handling units first cool/heat and dehumidify the air to the set-points (temperature and humidity) and then drive the supply air to the duct network connected to the rooms using the fans. Amid this process, the cooling/heating coils rely on the chiller providing circulated chilled water. Therefore, the chiller, cooling/heating coils, and fan account for the major parts of HVAC's energy consumption.

In this brief, we study the type of HVAC systems encompassing a fresh air unit (FAU) and a fan coil unit (FCU) for handling the recirculated air and fresh air separately [2]. We focus on the HVAC control for a specific office room for cooling with the thermal demand mainly caused by the heat gain from the outside, indoor occupants and electrical devices in use. To provide indoor comfort with the least energy use, we study the optimal control inputs for the HVAC system including supply airflow rates and the set-points of air handling units (i.e., FAU and FCU). We note there exist other types of HVAC systems that have integrated the FAU and FCU into one unit [23]. This brief just uses the type of HVAC system as an example to build a general framework, and the formulation and method are supposed to be extended accordingly.

Since the building thermal dynamics is a slow process and the arrival/ departure of occupants are spontaneous, we discuss the problem in a discrete-time setting with a sampling and control interval of $\Delta t = 30$ min over a daily circle (i.e., $T = 48$ stages).

B. MDP Formulation

A standard MDP is composed of system states, control, performance function, and system dynamics. In the following, we give their specific definitions for HVAC control.

1) System State: The system state is the input for computing HVAC control at each stage. Obviously, it should characterize the cooling demand determined by the present indoor condition (T_t^a, H_t^a), the weather (T_t^o, H_t^o) and the occupancy (N_t^a): $S_t = [T_t^o, H_t^o, T_t^a, H_t^a, N_t^a]^\top$.

2) Control Variables: Since the air flow rates and the set-points of FAU and FCU are to be decided, we have the control

$$A_t = \left[G_t^{\text{fau}}, T_t^{\text{fau}}, G_t^{\text{fcu}}, T_t^{\text{fcu}} \right]^\top.$$

3) System Dynamics: We use the gray-box model deduced from the energy and mass conservation equations to capture the indoor thermal and humidity dynamics [2], [24] and Markov chains to model the weather and indoor occupancy patterns.

Temperature: The indoor temperature variation is the interplay of HVAC control, thermal disturbance (i.e., the weather and occupancy), and the building thermal inertia, which can be described by the equations [2], [24]

$$
\begin{aligned}
C_p m^a & \left(T_{t+1}^a - T_t^a \right) \\
&= N_t^a \left(Q^o + Q^d \right) \Delta_t + h_{\text{gs}} A_{\text{gs}} \left(T_t^o - T_t^a \right) \Delta_t \\
&\quad + h_{\text{w}} A_{\text{wl}} \left(T_t^{\text{wl}} - T_t^a \right) \Delta_t + h_{\text{w}} A_{\text{wr}} \left(T_t^{\text{wr}} - T_t^a \right) \Delta_t \\
&\quad + G_t^{\text{fau}} \left(T_t^{\text{fau}} - T_t^a \right) \Delta_t + G_t^{\text{fcu}} \left(T_t^{\text{fcu}} - T_t^a \right) \Delta_t
\end{aligned}
\tag{1}
$$

where the first term captures the heat generated by the occupants and electrical devices (e.g., laptops, monitors, and desktops). The next three terms calculate the heat gains from the glass window and walls, and the last two model the cooling power provided by the FAU and FCU.

As (1) shows, the inside and outside interact through the walls. We have the wall thermal dynamics

$$
\begin{aligned}
C_w m^{\text{wl}} \left(T_{t+1}^{\text{wl}} - T_t^{\text{wl}} \right) &= h_w A_{\text{wl}} \left(T_t^a - T_t^{\text{wl}} \right) \Delta_t \\
C_w m^{\text{wr}} \left(T_{t+1}^{\text{wr}} - T_t^{\text{wr}} \right) &= h_w A_{\text{wr}} \left(T_t^a - T_t^{\text{wr}} \right) \Delta_t + \alpha_w A_{\text{wr}} Q_t^w \Delta_t.
\end{aligned}
\tag{2}
$$

Humidity: According to [2], we have the indoor humidity dynamics

$$
\begin{aligned}
m^a \left(H_{t+1}^a - H_t^a \right) &= N_t^a H^g \Delta_t + G_t^{\text{fau}} \left(H_t^{\text{fau}} - H_t^a \right) \Delta_t \\
&\quad + G_t^{\text{fcu}} \left(H_t^{\text{fcu}} - H_t^a \right) \Delta_t
\end{aligned}
\tag{3}
$$

where the first term calculates the humidity exhaled by the occupants, and the other two model the dehumidifying of FAU and FCU for the circulated air. The dehumidifying intensity of FAU and HCU are determined by their saturation set-points, and we have $H_t^{\text{fau}} = \min(H_t^o, H^{\text{fau,sat}})$ and $H_t^{\text{fcu}} = \min(H_t^a, H^{\text{fcu,sat}})$.

Occupancy: We use Markov chain to capture the occupancy patterns [25], [26]

$$\Pr\left(N_{t+1}^a = j | N_t^a = i \right) = \boldsymbol{P}_t^N[i, j], \ \forall i, j, \in \{1, \cdots, L_N\} \tag{4}$$

where $\boldsymbol{p}_t^N \in \mathbb{R}^{L_N}$ denotes the transition probability matrix of occupancy at time t. The integers i, j represents the occupancy levels within the L_N segments.

Weather Condition: Similarly, we use Markov chains to capture the outdoor temperature and humidity dynamics

$$
\begin{aligned}
\Pr\left(T_{t+1}^o = j | T_{t+1}^o = i \right) &= \boldsymbol{P}_t^T[i, j], \ \forall i, j \in \{1, \cdots, L_T\}. \\
\Pr\left(H_{t+1}^o = j | H_{t+1}^o = i \right) &= \boldsymbol{P}_t^H[i, j], \ \forall i, j \in \{1, \cdots, L_H\}
\end{aligned}
\tag{5}
$$

where $\boldsymbol{P}_t^T$ and $\boldsymbol{P}_t^H$ represent the transition probability matrices. In particular, we equally divide the outdoor temperature and humidity ranges into L_T and L_H segments. In this work, $\boldsymbol{P}_t^T$ and $\boldsymbol{P}_t^H$ are estimated from historical weather data in Singapore. We refer the readers to [27] for interest.

4) Objective Function: Considering the energy use is hard to inspect in practice, we select the electricity bill of HVAC system as the objective

$$J = \mathbb{E} \left[\sum_{t=0}^{T-1} c_t \left\{ \eta \left(C_t^{\text{fcu}} + C_t^{\text{fau}} \right) + F_t^{\text{fcu}} + F_t^{\text{fau}} \right\} \Delta_t \right] \tag{6}$$

where the cooling power $C_t^{\text{fcu}}, C_t^{\text{fau}}$ and the fan power $F_t^{\text{fcu}}, F_t^{\text{fau}}$ for the FAU and FCU are calculated as [2]

$$
\begin{aligned}
C_t^{\text{fcu}} &= C_p G_t^{\text{fcu}} \left(T_t^a - T_t^{\text{fcu}} \right) + C_p G_t^{\text{fcu}} \left[H_t^a \left(2500 + 1.84 T_t^a \right) \right. \\
&\quad \left. - H_t^{\text{fcu}} \left(2500 + 1.84 T_t^{\text{fcu}} \right) \right]
\end{aligned}
\tag{7a}
$$

$$
\begin{aligned}
C_t^{\text{fau}} &= C_p G_t^{\text{fau}} \left(T_t^o - T_t^{\text{fau}} \right) + C_p G_t^{\text{fau}} \left[H_t^o \left(2500 + 1.84 T_t^o \right) \right. \\
&\quad \left. - \left(2500 + 1.84 T_t^{\text{fau}} \right) \right]
\end{aligned}
\tag{7b}
$$

$$F_t^{\text{fcu}} = F^{\text{fcu,r}} \left(G_t^{\text{fcu}} / G^{\text{fcu},r} \right)^3 \tag{7c}$$

$$F_t^{\text{fau}} = F^{\text{fau,r}} \left(G_t^{\text{fau}} / G^{\text{fau},r} \right)^3. \tag{7d}$$

5) PMV Model: To enhance thermal comfort, this brief capitalizes on the elaborate PMV model to capture the occupants' satisfaction [20]. The PMV model is nonanalytical and implicitly characterized by a number of equations. For brevity, we denote the PMV model as

$$\text{pmv}_t = \text{PMV}\left(M, W, T_t^a, H_t^a, t_t^r, v_t^a, I_{\text{cl}} \right) \tag{8}$$

where the inputs include: metabolic rate M (W/m^2), mechanic work intensity W (W/m^2), indoor air temperature T_t^a (°C), relative humidity H_t^a (%), mean radiation temperature t_t^r (°C), indoor air velocity v_t^a (m^2), and clothing insulation I_{cl} (m^{2}K/W). The PMV model establishes the mapping from indoor condition to the occupants' average satisfaction within the range $[-3, 3]$. Particularly, the numbers -3, 0, 3 indicate too cold, ideal, and too hot, respectively.

6) Constraints: The operation of the HVAC system should comply with the physical limits of dampers within the FAU and FCU (9a), as well as the chiller capacity that determines the temperature set-points (9b)

$$\underline{G}^{\text{fau}} \leq G_t^{\text{fau}} \leq \overline{G}^{\text{fau}}, \quad \underline{G}^{\text{fcu}} \leq G_t^{\text{fcu}} \leq \overline{G}^{\text{fcu}}. \tag{9a}$$

$$\underline{T}^{\text{fau}} \leq T_t^{\text{fau}} \leq \overline{T}^{\text{fau}}, \quad \underline{T}^{\text{fcu}} \leq T_t^{\text{fcu}} \leq \overline{T}^{\text{fcu}}. \tag{9b}$$

We use $\underline{\text{pmv}}$ and $\overline{\text{pmv}}$ to capture the human comfort requirement characterized by the PMV value, that is

$$\underline{\text{pmv}} \leq \text{pmv}_t \leq \overline{\text{pmv}} \tag{10}$$

where we usually have $\underline{\text{pmv}} = -0.5$ and $\overline{\text{pmv}} = 0.5$.

7) Optimization Problem: Overall, the optimal control of the HVAC system to optimize energy cost while respecting occupant comfort can be described as

$$\min_{\pi} \; J(S_0, \pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{T-1} c_t \{\eta(C_t^{\text{fcu}} + C_t^{\text{fau}}) + F_t^{\text{fcu}} + F_t^{\text{fau}}\}\Delta_t\right]$$

s.t. System dynamics:$(1)-(5)$, Operation limits: (9),

$\quad$ Thermal comfort: (10), $\forall t \in \{0, 1, \ldots, T-1\}$ $\tag{11}$

where $\pi = (\pi_0, \pi_1, \ldots, \pi_{T-1})$ denotes the control policy over the optimization horizon. S_0 is the initial system state. At each time t, the control $\pi_t\colon \mathcal{S}_t \to \mathcal{A}_t$ establishes a mapping from the state space $\mathcal{S}_t$ to the action space $\mathcal{A}_t$. $\mathbb{E}^{\pi}$ denotes the expectation under policy π.

We have formulated the optimal control of the HVAC system as a standard MDP (11). However, searing for an optimal policy is nontrivial and remains to be addressed. Dynamic programming (DP) for problem (11) requires to transverse Q-factors for the Cartesian product of state and action space backward or forward to identify the optimal deterministic policy [22]. This is computationally impractical due to: 1) the multiple uncertainties resulting in complicated state transitions and 2) the large state and action space posing intensive computation.

III. GB-L

To handle (11), we propose a GB-L algorithm to progressively learn a stochastic policy off-line and store it for on-line implementation. To account for the multiple uncertainties, the Monte Carlo (MC) technique [28] is employed to simulate system dynamics and estimate the performance gradients of policies. Particularly, for the HVAC control, we establish two specific strategies for reducing computation by exploring the problem features.

A. Notations

We use the lower cases s_t and a_t, b_t, c_t to represent state and action instances at time t, respectively. We use the integer sets $\mathcal{S}_t \triangleq \{1, 2, \ldots, |\mathcal{S}_t|\}$ and $\mathcal{A}_t \triangleq \{1, 2, \cdots, |\mathcal{A}_t|\}$ to denote the state and action space, where the operator $|\cdot|$ defines the cardinality. $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{T-1})^T$ denotes a stochastic policy, where $\theta_t \in \mathbb{R}^{|\mathcal{S}_t| \times |\mathcal{A}_t|}$ establishes the mapping from the state space $\mathcal{S}_t$ to the action space $\mathcal{A}_t$ with $\theta_t(s_t, a_t) \in [0, 1]$ denoting the probability to take action a_t at state s_t. The lower cases $p_t(s_{t+1}|s_t, a_t)$ and $p_t^{\theta}(s_{t+1}|s_t)$ indicate the state transition probability when taking action a_t or policy $\boldsymbol{\theta}$, and $\boldsymbol{P}_t^{\theta} = [p_t^{\theta}(s_{t+1}|s_t)] \in \mathbb{R}^{|\mathcal{S}_t| \times |\mathcal{S}_{t+1}|}$ is the transition probability matrix. The superscript k denotes the iteration.

B. GB-L

To handle the various constraints of problem (11), we redefine an augmented stage-cost function

$$r_t(s_t, a_t) = c_t[\eta(C_t^{\text{fcu}} + C_t^{\text{fau}}) + F_t^{\text{fcu}} + F_t^{\text{fau}}]\Delta_t + I_t(\mathcal{X}_t) \tag{12}$$

where the first part is the energy cost and the second part quantifies the penalty of constraint violations. Particularly, we use $\mathcal{X}_t$ to indicate the set of constraints at time t, comprising (1)–(5), (9) and (10). We have the indicator function $I_t(\cdot)$ such that $I(A) = 1$ with the condition A true, otherwise $\mathbb{I}(A) = 0$.

Our method is inspired by the performance difference equation [29] that for any two stochastic policies $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$, the induced performance difference can be quantified by

$$J(\boldsymbol{\mu}; S_0) - J(\boldsymbol{\sigma}; S_0) = \sum_{t=0}^{T-1} \pi_t^{\mu}[(\boldsymbol{r}_t^{\mu} - \boldsymbol{r}_t^{\sigma}) + (\boldsymbol{P}_t^{\mu} - \boldsymbol{P}_t^{\sigma})V_{t+1}^{\sigma}] \tag{13}$$

where we have

$$\boldsymbol{\pi}_t^{\theta} = (\pi_t^{\theta}(1), \pi_t^{\theta}(2), \cdots, \pi_t^{\theta}(|\mathcal{S}_t|))^{\top},$$
$$\boldsymbol{r}_t^{\theta} = (r_t^{\theta}(1), r_t^{\theta}(2), \cdots, r_t^{\theta}(|\mathcal{S}_t|))^{\top},$$
$$\boldsymbol{V}_{t+1}^{\theta} = (V_{t+1}^{\theta}(1), \cdots, V_{t+1}^{\theta}(|\mathcal{S}_{t+1}|))^{\top}$$

denote the state distribution, stage-cost and performance potential for policy $\theta \in \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$. In particular, the performance potential is defined as

$$V_{t+1}^{\theta}(s_{t+1}) = \mathbb{E}^{\theta}\left[\sum_{\tau=t+1}^{T-1} r_{\tau}(s_{\tau}, a_{\tau})\right], \quad \text{with } a_{\tau} = \boldsymbol{\theta}(s_{\tau}). \tag{14}$$

From the perspective of perturbation analysis (PA), the stochastic policy $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ can be viewed as the base and perturbed policy. The intuitive interpretation of (13) is that the performance of the perturbed policy $\boldsymbol{\mu}$ can be constructed by the performance potentials $V_{t+1}^{\sigma}(s_{t+1})$ of base policy $\boldsymbol{\sigma}$. However, it requires the explicit formula of state distribution π^{μ} under the perturbed policy $\boldsymbol{\mu}$, which is generally unknown *a priori*, making it impractical to use (13) for policy update.

To handle it, we derive a differential formula by defining a structured random policy $\boldsymbol{\delta}$ which adopts policy $\boldsymbol{\sigma}$ with probability δ and policy $\boldsymbol{\mu}$ with probability $1-\delta$. For policy $\boldsymbol{\delta}$, we can easily identify the state transition probability matrix $\boldsymbol{P}_t^{\delta} = \boldsymbol{P}_t^{\sigma} + \delta\Delta\boldsymbol{P}_t$ with $\Delta\boldsymbol{P}_t = \boldsymbol{P}_t^{\mu} - \boldsymbol{P}_t^{\sigma}$, and the stage-cost vectors $\boldsymbol{r}_t^{\delta} = \boldsymbol{r}_t^{\sigma} + \delta\Delta\boldsymbol{r}_t$ with $\Delta\boldsymbol{r}_t = \boldsymbol{r}_t^{\mu} - \boldsymbol{r}_t^{\sigma}$. In this regard, (13) can be translated into

$$J(\boldsymbol{\delta}; S_0) - J(\boldsymbol{\sigma}; S_0) = \sum_{t=0}^{T-1} \pi_t^{\delta}[\Delta\boldsymbol{r}_t + \Delta\boldsymbol{P}_t V_{t+1}^{\sigma}]. \tag{15}$$

By assuming $\delta \to 0$, we have the performance differential equation

$$\frac{dJ(\boldsymbol{\delta}; S_0)}{d\delta} = \lim_{\delta \to 0} \sum_{t=0}^{T-1} \pi_t^{\delta}[\Delta\boldsymbol{r}_t + \Delta\boldsymbol{P}_t V_{t+1}^{\sigma}]. \tag{16}$$

Equivalently, we have

$$\frac{\partial J(\boldsymbol{\sigma}; S_0)}{\partial \sigma_t(s_t, a_t)} = \pi_t^{\sigma}(s_t)\left[\frac{\partial r_t^{\sigma}(s_t)}{\partial \sigma_t(s_t, a_t)} + \sum_{s_{t+1} \in \mathcal{S}_{t+1}} \frac{\partial p_t^{\sigma}(s_{t+1}|s_t)}{\partial \sigma_t(s_t, a_t)} V_{t+1}^{\sigma}(s_{t+1})\right]. \tag{17}$$

By substituting $r_t^{\sigma}(s_t) = \sum_{a_t=1}^{|\mathcal{A}_t|} p_t^{\sigma}(a_t|s_t)r_t(s_t, a_t)$, $p_t^{\sigma}(s_{t+1}|s_t) = \sum_{a_t=1}^{|\mathcal{A}_t|} p_t^{\sigma}(a_t|s_t)p(s_{t+1}|s_t, a_t)$ into (17),

we have

$$\frac{\partial J(\boldsymbol{\sigma}; S_0)}{\partial \sigma_t(\boldsymbol{s}_t, \boldsymbol{a}_t)}$$

$$= \pi_t^\sigma(\boldsymbol{s}_t) \left[\sum_{b_t=1}^{|\mathcal{A}_t|} \frac{\partial p_t^\sigma(\boldsymbol{b}_t|\boldsymbol{s}_t)}{\partial \sigma_t(\boldsymbol{s}_t, \boldsymbol{a}_t)} r_t(\boldsymbol{s}_t, \boldsymbol{b}_t) \right.$$

$$\left. + \sum_{s_{t+1}\in\mathcal{S}_{t+1}} \sum_{b_t=1}^{|\mathcal{A}_t|} \frac{\partial p_t^\sigma(\boldsymbol{b}_t|\boldsymbol{s}_t)}{\partial \sigma_t(\boldsymbol{s}_t, \boldsymbol{a}_t)} p_t(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{b}_t) V_{t+1}^\sigma(\boldsymbol{s}_{t+1}) \right]$$

$$= \pi_t^\sigma(\boldsymbol{s}_t) \left[\sum_{b_t=1}^{|\mathcal{A}_t|} \frac{\partial p_t^\sigma(\boldsymbol{b}_t|\boldsymbol{s}_t)}{\partial \sigma_t(\boldsymbol{s}_t, \boldsymbol{a}_t)} \left(r_t(\boldsymbol{s}_t, \boldsymbol{b}_t) + V_t^\sigma(\boldsymbol{s}_t, \boldsymbol{b}_t) \right) \right] \quad (18)$$

where $V_t^\sigma(\boldsymbol{s}_t, \boldsymbol{b}_t) = \sum_{s_{t+1}\in\mathcal{S}_{t+1}} p_t(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{b}_t) V_{t+1}^\sigma(\boldsymbol{s}_{t+1})$.

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

$$\frac{\partial p_t^\sigma(\boldsymbol{b}_t|\boldsymbol{s}_t)}{\partial \sigma_t(\boldsymbol{s}_t, \boldsymbol{a}_t)} = \begin{cases} \frac{\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t(\boldsymbol{s}_t, \boldsymbol{c}_t) - \sigma_t(\boldsymbol{s}_t, \boldsymbol{a}_t)}{\left[\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t(\boldsymbol{s}_t, \boldsymbol{c}_t) \right]^2}, & \boldsymbol{b}_t = \boldsymbol{a}_t \\ \frac{-\sigma(\boldsymbol{s}_t, \boldsymbol{b}_t)}{\left[\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t(\boldsymbol{s}_t, \boldsymbol{c}_t) \right]^2}, & \boldsymbol{b}_t \neq \boldsymbol{a}_t \end{cases} . \quad (19)$$

Equation (18) can be interpreted as the performance gradients of policy $\boldsymbol{\sigma}$, and thus we can establish the standard policy update procedure

$$\boldsymbol{\sigma}^{k+1} = \boldsymbol{\sigma}^k - \boldsymbol{\gamma}^k \cdot \nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0) \quad (20)$$

where we have $\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0) = [(\partial J(\boldsymbol{\sigma}^k; S_0)/\partial \sigma_t^k(\boldsymbol{s}_t, \boldsymbol{a}_t))]$. $\boldsymbol{\gamma}^k = [\gamma_t^k(\boldsymbol{s}_t, \boldsymbol{a}_t)]$ denotes the step-size at iteration k.

For the policy update formula (20), we note two problems to be addressed: 1) computing the performance gradients $\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0)$ and 2) determining the step-size $\boldsymbol{\gamma}^k$. As there exists randomness both for the weather and occupancy, it is impractical to analytically estimate the performance gradients under expectation. To overcome this difficulty, the MC method [28] is adopted to simulate the system dynamics and estimate the performance gradients $\nabla_\sigma J(\boldsymbol{\sigma}^k; S_0)$ for any given policy $\boldsymbol{\sigma}$. We refer the readers to our extended version [27] for details. For the step-size $\boldsymbol{\gamma}^k$, which is closely related to the method's converge, we have the main results in Theorem 1.

Theorem 1: For any given initial policy $\boldsymbol{\sigma}^0$, the GB-L method can converge to an optimal policy of problem (11) with the selected step size $\gamma_t^k(\boldsymbol{s}_t, \boldsymbol{a}_t) = (\sigma_t^k(\boldsymbol{s}_t, \boldsymbol{a}_t)/\sum_{c_t=1}^{|\mathcal{A}_t|} \sigma_t^k(\boldsymbol{s}_t, \boldsymbol{c}_t))$ $(\forall \boldsymbol{s}_t \in \mathcal{S}_t, \boldsymbol{a}_t \in \mathcal{A}_t)$ and the performance gradients $\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0)$ estimated accurately enough.

The proof refers to Appendix A [27].

Remark 1: As Theorem 1 illustrates, the method's convergence depends on the estimation accuracy of the performance gradients. Generally, any accurate estimation can be approached by increasing the number of sample paths. However, we want to explain that the practical implementation only requires an estimation that preserves the (performance) order of the (action) candidates.

An example: for any two action a and b, we only require an estimation that $\tilde{J}(a) \leq \tilde{J}(b)$ if $J(a) \leq J(b)$ where $J(\cdot)$ and $\tilde{J}(\cdot)$ denote the real and estimated value.

Remark 2: The method's convergence does not depend on the selection of initial policy $\boldsymbol{\sigma}^0$, which can be identified from the proof of Theorem 1. However, a better initial policy is expected to yield faster convergence and less iterations.

The implementation of GB-L is shown in Algorithm 1 which comprises the two main steps: 1) estimate the performance gradients (step 3) and 2) update the policy (step 4). We select $\|\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0)\|_2 \leq \epsilon$ (ϵ is positive threshold) as the stopping criterion. Note that the learning process can be carried out off-line and then store the obtained policy for on-line implementation. In such a setting, we can expect efficient on-line execution as it only requires to identify the action at each stage from the storage based on the system states.

Algorithm 1 GB-L

1: **Initialize** $k \to 0$, $\boldsymbol{\sigma}^0$.
2: **Iteration:**
3: Estimate the performance gradients $\nabla_{\boldsymbol{\sigma}^k} J(\boldsymbol{\sigma}^k; S_0)$ based on MC method.
4: Update policy:

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

5: Stop if the **stopping criterion** is reached, otherwise $k = k + 1$ and go to **Step** 3.

C. Computation Reduction

We note the main computation burden of GL-B lies in estimating the performance gradients of policies (step 3 of Algorithm 1). According to (18), the performance potentials of the state-action pairs are required to build the performance gradients. For HVAC control, both the state and control encompass concatenated variables, causing large state and action space. Therefore, the computation is an underlying issue to be concerned. To handle it, we establish two strategies to reduce the computation of GB-L for HVAC control by studying the problem features.

Strategy I—Concentrate on the High Probability States: This strategy is promoted by the data analysis [27]. We observe though the outdoor temperature and humidity spread wide ranges throughout the day (i.e., temperature [22, 34] °C and relative humidity [40, 100]%), their variations over each specific time period are quite small. For instance, the outdoor temperature and humidity mostly lie in the range of [26, 28] °C and [75, 95]%) at 6:00 A.M. (see Section IV-A [27]). This suggests us to concentrate our computation on the states within those ranges while learning as the states outside are rare and their actions have less impact on the overall performance.

Strategy II—Pick a Good Initial Policy $\boldsymbol{\sigma}^0$: The operation of the HVAC system should satisfy indoor comfort (avoid the occurrence of discomfort states). For HVAC control, we have some insight about the comfortable temperature and humidity ranges. For example, the temperature out of [23, 28] °C or humidity out of [40, 70]% will cause discomfort (we can evaluate the PMV metrics). Therefore, instead of a random pick, we can have a more reasonable initial policy $\boldsymbol{\sigma}^0$ by assigning the entries corresponding to the indoor temperature and humidity out of the ranges as *zero*. This is expected to reduce the iterations.

IV. CASE STUDIES

In this section, we evaluate the performance of the proposed method for HVAC control via simulations. We compare the method with existing MPC-based relaxation (MPC-R) method on the energy saving performance, thermal comfort, and on-line computing cost.

Simulation Settings: We consider an office of size 6 m × 5 m × 4 m and occupied by five staffs. We equally divide the occupancy into $L_A = 6$ levels, corresponding to $0, 1, \ldots, 5$ occupants. The static inputs of the PMV model refer to the ANSI/ASHRAE Standard [30]. The other parameter settings can refer to our extended version [27].

Considering most existing on-line HVAC control is deployed in MPC, we compare the GB-L with an MPC method. Particularly, to identify the performance bounds, we assume perfect information (i.e., weather and occupancy) for the MPC in our simulations, which is usually not available in realization but to create a benchmark here. As for the MPC method with a receding horizon H, we have the formulation

$$\min_{\substack{G_t^{\text{fau}}, G_t^{\text{fcu}}, \\ T_t^{\text{fau}}, T_t^{\text{fcu}}}} J(t_k) = \sum_{t=t_k}^{t_k+H-1} c_t \{\eta(C_t^{\text{fcu}} + C_t^{\text{fau}}) + F_t^{\text{fcu}} + F_t^{\text{fau}}\} \Delta_t$$

s.t. System dynamics : $(1)-(3)$, Operation limits: (9),
Thermal comfort: (10), $\forall t \in \{t_k, t_k+1, \cdots, t_k+H-1\}$. (22)

As discussed, solving problem (22) exactly is a nontrivial task due to the nonlinear and nonanalytical constraints imposed by the system dynamics and the PMV model. To build the comparisons, we settle for an existing relaxed solution method (referred to MPC-R) [9], [10] as it is the scarcely available and capable one for problem (22) to our best knowledge. The main point of MPC-R is to convert the nonlinear constraints into a sequence of combinatorial linear constraints through piecewise linearization and approximation so as to be tackled by some commercial solvers, like CPLEX.

We compare the two methods under three different settings regarding the state and control discretization. Particularly, the state discretization is only required by GB-L not the MPC-R.

S-1: The discretization pace of temperature and relative humidity are 2 °C and 10%, respectively. The set-point temperature levels for FAU and FCU are {12, 14, 16} °C, and their supply airflow rates are equally divided into three levels.

S-2: The discretization of temperature and relative humidity are as *S-1*. The set-point temperature of FAU and FCU are fixed as 15 °C, and their supply airflow rates are equally divided into five levels.

S-3: The discretization pace of temperature and relative humidity are 1 °C and 5%. The settings for FAU and FCU are as S-2.

For the GB-L, we use 1000 (S-1), 2000 (S-2) and 5000 (S-3) sample paths to estimate the performance gradients while learning the policy. As there exist uncertainties, we compare the performance (i.e., energy cost, thermal comfort and on-line computation cost) of two methods under 100 randomly generated scenarios. First of all, we study the distributions of induced energy cost by the two methods as shown in Fig. 1.
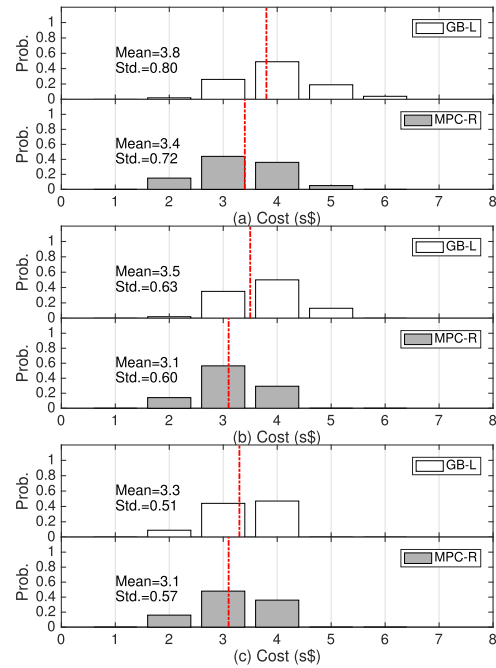

Fig. 1. Histograms of HVAC costs. (a) S-1. (b) S-2. (c) S-3.

TABLE I
ON-LINE COMPUTING TIME OF GB-L AND MPC-R

#Settings	GB-L		MPC-R	
	Mean (sec.)	Std (sec.)	Mean (sec.)	Std(sec.)
S-1	0.48	0.09	259.83	40.82
S-2	0.58	0.13	247.41	41.91
S-3	0.43	0.06	335.39	43.14

We see some performance discount of GB-L over MPC-R: 11.7% (S-1), 12.9% (S-2) and 6.5% (S-3), respectively. This is reasonable as we have assumed accurate information for the MPC-R to provide the near-optimal bounds. Besides, the variation of performance of the GB-L with the different settings can be attributed to: 1) the state discretization pace and 2) the estimation accuracy of performance gradients affected by the number of sample paths used. Reasonably, we can expect to shrink the performance gap further with a finer-grained state discretization and increased sample paths if more computation is acceptable. This is inferred from the comparison of S-2 and S-3.

Further, we study the on-line computation cost for the two methods. Particularly, GB-L learns the optimal control policy off-line and stores it for on-line execution. Therefore, the main on-line computation lies in identifying the control from the state-action mapping table. Whereas the implementation of MPC-R requires to solve problem (22) based on the updated information at each stage. For the different settings, the mean and standard deviation (std.) of on-line computation[1] throughout the optimization horizon are contrasted in Table I. We see the GB-L can respond instantly (less than 1 s), whereas the MPC-R takes 4–5 min. This demonstrates GB-L can enable efficient on-line implementation.

Subsequently, we evaluate the thermal comfort under the two methods by inspecting the indoor temperature, humidity,

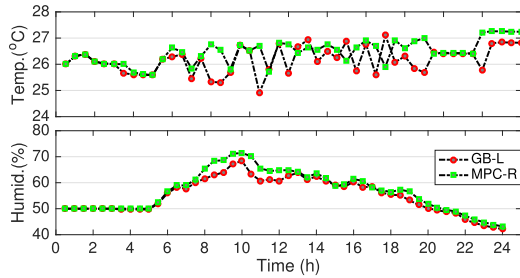[1]Simulation platform: MATLAB R2016a, Window i7-5500 CPU@2.40GHZ.

Fig. 2. Indoor temperature (Temp.) and relative humidity (Humid.) for a specific scenario with the GB-L and MPC-R.
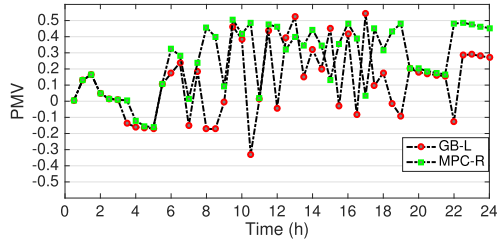

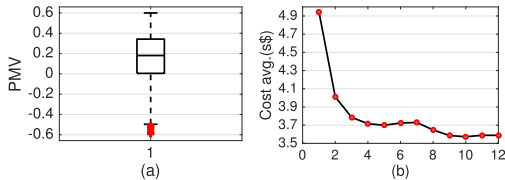Fig. 3. PMV curves for a scenario with the GB-L and MPC-R method.

Fig. 4. (a) Distribution of PMV value with the GB-L (S-2). (b) Convergence rate of GB-L (S-2).

and the PMV value for a randomly picked scenario with S-2. As shown in Fig. 2, we observe both the indoor temperature and relative humidity are maintained in the typical comfortable range $[24, 27]$ °C and $[40, 70]$% with the two methods. More notably, as shown in Fig. 3, the PMV curve lies in the specified range of $[-0.5, 0.5]$. However, to be noted is that we may observe comfort violations with other realizations due the uncertainties. To evaluate the thermal comfort with GB-L under uncertainties, we study the distribution of PMV value under the 100 realizations. As indicated in Fig. 4(a), the PMV value is almost maintained within the range of $[-0.5, 0.6]$. Particularly, It is reasonable to see a minor upper violation (i.e., 0.6) as the energy cost saving target is set. Furthermore, we find the exact thermal comfort, i.e., PMV $[-0.5, 0.5]$ is achieved with a probability of 93%. This demonstrates the desirable performance of the GB-L for providing thermal comfort under uncertainties.

Use this scenario as an example, we also study the control inputs for the FAU and FCU. As shown in Fig. 5, we see quite similar operation patterns of FAU/FCU with the two methods. This implicitly illustrates the energy-saving performance of GB-L that we observed in Fig. 1. Besides, we can observe some other interesting phenomenon that the FAU are mostly operated in quite lower level versus the FCU. This is as rational as the FAU handles outdoor fresh air with a general higher temperature than that of the recirculated air managed by the FCU. To fulfill the energy-saving objective, the HVAC system tends to activate the FCU instead of FAU in priority to reduce cooling load. Besides, we observe the operation
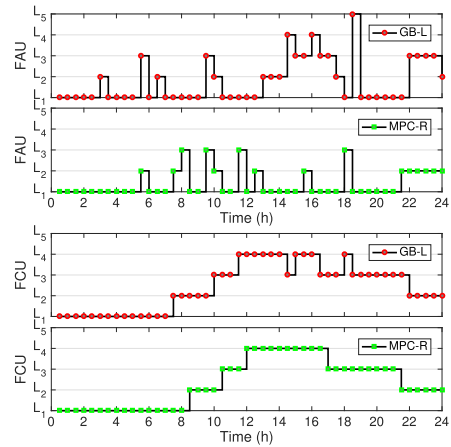

Fig. 5. Control of FAU and FCU for a specific scenario with the GB-L and MPC-R method (L_1–L_5 denote the air flow rate levels).
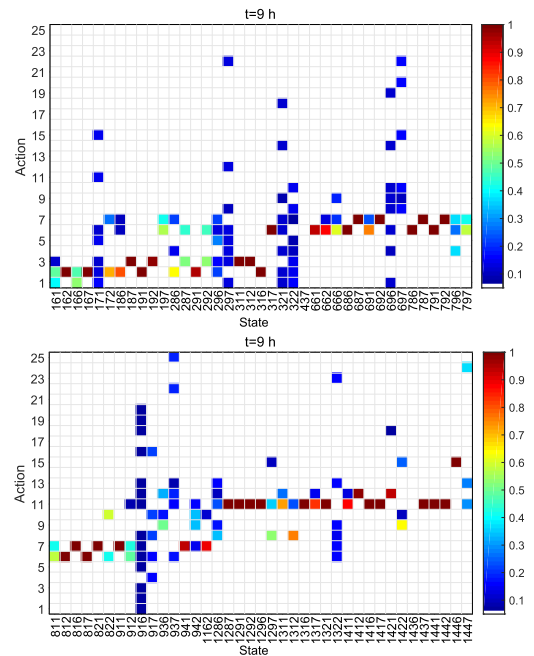

Fig. 6. Stochastic policy obtained by GB-L for 9:00 A.M. (S-2).

patterns, especially the FCU, correspond well to the typical weather patterns (i.e., lower outdoor temperature in the early morning and late night, and higher temperature in the noon) and the occupancy patterns (i.e., high occupancy during the working hours and low occupancy during nonworking hours). The demonstrates the capability of GB-L responding to the uncertain thermal disturbance caused by the weather and occupancy.

Though the GB-L is implemented off-line, the convergence speed is still an important and concerned issue. Therefore, we inspect the convergence rate of the GB-L with S-2. We inspect the average energy cost of 100 realizations while executing Algorithm 1. Take the average energy cost as an indicator, we obtain the convergence rate in Fig. 4(b). For this case, it takes about ten iterations to approach the "optima."

Last, we visualize the obtained stochastic policy (a mapping from the state space to the action space) at 9:00 A.M. in Fig. 6 (only for the visited states in the sample paths) as an example. From the figure, we can have some insights regarding the

characteristics of GB-L for HVAC control. We note that for many states, the probability (to take action) is concentrated within a small group of actions. However, for some states like {171, 296, 297, 321, 322, 696, 697, ...}, the probability distributions scatter more diversely on the action space. The latter is attributed to the low occurrence of those states in the sample paths, resulting in fewer updates. Conversely, this implies those states occur with lower probability and deserve less computation.

V. CONCLUSION

This brief studied the energy-efficient control of HVAC systems subject to uncertain thermal disturbances caused by the weather and the occupants. To enhance thermal comfort, we incorporate the elaborate PMV thermal comfort model in the optimization. This problem suffers computational challenges from the nonlinear and nonanalytical constraints imposed by the thermal dynamics and PMV model. To handle it, we formulated the problem as an MDP and proposed a GB-L method for learning the optimal stochastic control policy off-line and stored it for efficient on-line implementation. We prove the method's converge to the optimal policies theoretically. In parallel, we demonstrated the method's performance (i.e., energy cost, thermal comfort, and on-line computing) for HVAC control via simulations. The comparisons with the existing MPC-based relaxation (MPC-R) method which was assumed with accurate future information and expected to provide the near-optimal bounds, showed that though there exists some performance loss in energy cost savings, the proposed method can enable efficient on-line implementation and provide a high probability of thermal comfort under uncertainties.

This brief has used the singe office as an example to establish the GB-L method. An interesting future direction is its extension to multizone commercial buildings. However, that is not straightforward due to the concatenated state and action space to handle. From our insight, one possible solution is to adopt the *"one-agent-at-a-time"* idea (i.e., sequentially learning the control policies for each individual zones) proposed in [31] to address the computational issue.

REFERENCES

[1] K. L. Ku, J. S. Liaw, M. Y. Tsai, and T. S. Liu, "Automatic control system for thermal comfort based on predicted mean vote and energy saving," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 1, pp. 378–383, Jan. 2015.

[2] 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 Trans. Autom. Sci. Eng.*, vol. 10, no. 3, pp. 588–602, Jul. 2013.

[3] M. Frontczak, S. Schiavon, J. Goins, E. Arens, H. Zhang, and P. Wargocki, "Quantitative relationships between occupant satisfaction and satisfaction aspects of indoor environmental quality and building design," *Indoor Air*, vol. 22, no. 2, pp. 119–131, Apr. 2012.

[4] S. Kejriwal and S. Mahajan. (2016). *Smart Buildings: How IoT Technology AIMS to Add Value for Real Estate Companies*. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/real-estate/deloitte-nl-fsi-real-estate-smart-buildings-how-iot-technology-aims-to-add-value-for-real-estate-companies.pdf

[5] M. Mossolly, K. Ghali, and N. Ghaddar, "Optimal control strategy for a multi-zone air conditioning system using a genetic algorithm," *Energy*, vol. 34, no. 1, pp. 58–66, Jan. 2009.

[6] M. Fasiuddin and I. Budaiwi, "HVAC system strategies for energy conservation in commercial buildings in saudi arabia," *Energy Buildings*, vol. 43, no. 12, pp. 3457–3466, Dec. 2011.

[7] Z. Afroz, G. Shafiullah, T. Urmee, and G. Higgins, "Modeling techniques used in building HVAC control systems: A review," *Renew. Sustain. Energy Rev.*, vol. 83, pp. 64–84, Mar. 2018.

[8] A. Kelman and F. Borrelli, "Bilinear model predictive control of a HVAC system using sequential quadratic programming," in *Proc. IFAC World Congr.*, vol. 18, 2011, pp. 9869–9874.

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

[10] Z. Xu, Q.-S. Jia, and X. Guan, "Supply demand coordination for building energy saving: Explore the soft comfort," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 2, pp. 656–665, Apr. 2015.

[11] A. Garnier, J. Eynard, M. Caussanel, and S. Grieu, "Predictive control of multizone heating, ventilation and air-conditioning systems in non-residential buildings," *Appl. Soft Comput.*, vol. 37, pp. 847–862, Dec. 2015.

[12] Q. Zou, J. Ji, S. Zhang, M. Shi, and Y. Luo, "Model predictive control based on particle swarm optimization of greenhouse climate for saving energy consumption," in *Proc. World Automat. Congr.*, 2010, pp. 123–128.

[13] A. Afram and F. Janabi-Sharifi, "Theory and applications of HVAC control systems—A review of model predictive control (MPC)," *Building Environ.*, vol. 72, pp. 343–355, Feb. 2014.

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

[15] F. Oldewurtel *et al.*, "Use of model predictive control and weather forecasts for energy efficient building climate control," *Energy Buildings*, vol. 45, pp. 15–27, Feb. 2012.

[16] F. Oldewurtel, C. N. Jones, A. Parisio, and M. Morari, "Stochastic model predictive control for building climate control," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 3, pp. 1198–1205, May 2014.

[17] A. Parisio, L. Fabietti, M. Molinari, D. Varagnolo, and K. H. Johansson, "Control of HVAC systems via scenario-based explicit MPC," in *Proc. 53rd IEEE Conf. Decis. Control*, Dec. 2014, pp. 5201–5207.

[18] M. Klauco and M. Kvasnica, "Explicit MPC approach to PMV-based thermal comfort control," in *Proc. 53rd IEEE Conf. Decis. Control*, Dec. 2014, pp. 4856–4861.

[19] B. Kouvaritakis and M. Cannon, "Stochastic model predictive control," in *Encyclopedia of Systems and Control*. London, U.K.: Springer, 2015, pp. 1350–1357.

[20] P. O. Fanger, *Thermal Comfort: Analysis and Applications in Environmental Engineering*. Copenhagen, Denmark: Danish Technical Press, 1970.

[21] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2014.

[22] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 1994.

[23] N. Radhakrishnan, S. Srinivasan, R. Su, and K. Poolla, "Learning-based hierarchical distributed HVAC scheduling with operational constraints," *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 5, pp. 1892–1900, Sep. 2018.

[24] Z. Wu, Q.-S. Jia, and X. Guan, "Optimal control of multiroom HVAC system: An event-based approach," *IEEE Trans. Control Syst. Technol.*, vol. 24, no. 2, pp. 662–669, Mar. 2016.

[25] R. Jia, R. Dong, S. S. Sastry, and C. J. Spanos, "Privacy-enhanced architecture for occupancy-based HVAC control," in *Proc. 8th Int. Conf. Cyber-Phys. Syst. (ICCPS)*, Apr. 2017, pp. 177–186.

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

[27] Y. Yang, G. Hu, and C. J. Spanos, "Stochastic optimal control of HVAC system for energy-efficient buildings," 2019, *arXiv:1911.00840*. [Online]. Available: http://arxiv.org/abs/1911.00840

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

[29] Y. Zhao, "Optimization theories and methods for Markov decision processes in resource scheduling of networked systems," Ph.D. dissertation, Tsinghua Univ., Beijing, China, 2010.

[30] *Thermal Environmental Conditions for Human Occupancy*, Standard 55, ANSI/ASHRAE, Standing Standard Project Committee, Mar. 2017.

[31] D. Bertsekas, "Multiagent rollout algorithms and reinforcement learning," 2019, *arXiv:1910.00120*. [Online]. Available: http://arxiv.org/abs/1910.00120