# Approximate Dynamic Programming for Building Control Problems with Occupant Interactions

Donghwan Lee, Seungjae Lee, Panagiota Karava, and Jianghai Hu

***Abstract*— The goal of this paper is to study potential applicability and performance of approximate dynamic programming (ADP) for building control problems. It is known that occupants' stochastic behavior affects the thermal dynamics of building spaces. We consider an occupant stochastic behavior model which depends on environmental variables and apply ADP with illustrative scenarios of occupant-building interactions. Through simulations, we demonstrate the validity of ADP-based control designs for building control problems with occupants.**

## I. Introduction

The main goal of building control problems is to balance between the energy consumption and occupants' comfort in building spaces. There has been a great amount of research interests in energy consumption and comfort management in buildings [1], [2]. Model predictive control (MPC) is a popular optimal control scheme in the presence of various constraints and objectives. For this reason, it has been widely used for building control problems [3], [4]. However, MPC uses predictions of system's future output trajectories, which are computed by using a mathematical model of the system. Therefore, its performance is sensitive to unpredictable disturbances and uncertainties. In this respect, one of the main difficulties in building control problems is the presence of stochastic uncertainties and disturbances, such as weather and occupant interactions, that cannot be exactly predicted in general.

To resolve this problem, robust MPC [5] and stochastic MPC [6], [7] can be used. The stochastic MPC has been widely studied for building controls problems [2], [8]–[10]. Many stochastic MPC approaches assume that the system disturbances are Gaussian. To meet more practical needs, scenario-based (or sample-based) MPC [11]–[15] can be applied to cope with generic non-Gaussian stochastic disturbances. There are still more challenging practical situations for which the scenario-based MPC is not applicable. An example is a stochastic system with stochastic disturbances/uncertainties whose probability distributions depend on environmental factors such as the current state variables of the control system. As a result, the probability depends on the control policy and the corresponding design parameters as well. In this case, realizations of state trajectories generated by using fixed design parameters cannot reflect changes of the parameters themselves. Such cases arise in many applications. For example, stochastic systems with disturbances modelled by Markov chains were considered in [16], [17] for hybrid electric vehicle powertrain management problems, where transition probabilities of the Markov chain depend on the state of the dynamic system. One possible approach to solve optimal control problems for the complicated stochastic systems is to use approximate dynamic programming (ADP) [18]–[20] (or reinforcement learning [20] from the machine learning context). For instance, [16] uses infinite-horizon ADP, while [17] applies finite-horizon ADP. Another example arises in building control problems that consider occupant interactions with the building systems. The role of occupants is significant in the thermal dynamics of building spaces [21]–[23]. Occupant models based on Markov chains have been studied in [23], [24] for building control problems. Noting that the thermal preferences induce occupant actions that perturb the thermal dynamics of building spaces, a building space with occupants is a stochastic system whose probabilistic behavior depends on the state variables.

The goal of this paper is to study an application of ADP to building control problems with occupant interactions. Versions of the ADP are sometimes called reinforcement learning (RL) from the machine learning context when it is used with a model-free or simulation-based methods. RL is a family of unsupervised learning schemes for agents interacting with unknown environment, and has been widely studied in [19], [25]–[28]. Model-free RL for building control problems has been studied in several researches, for instance [29]–[34], to find a balance among energy savings, high comfort, and indoor air quality. However, the previous studies do not consider occupant interactions in the building thermal dynamics. In this paper, we assume that a stochastic model of occupant behavior is given. Based on the model, an approximate optimal control policy is designed by using simulation-based Q-learning [35], which is a class of ADPs. A contribution of this paper is the presentation of illustrative scenarios where ADP can be applied to building control systems with occupant interactions, assessing potential of ADP in those cases. In addition, we present the convergence of dynamic programming with exit probabilities of the state space.

## II. Preliminaries

Throughout the paper, the following notations will be used: $\mathbb{N}$ and $\mathbb{N}_+$: sets of nonnegative and positive integers, respectively; $\mathbb{R}$: set of real numbers; $\mathbb{R}^n$: $n$-dimensional

This material is based upon work supported by the National Science Foundation under Grant No. 1539527

D. Lee and J. Hu are with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906, USA lee1923@purdue.edu, jianghai@purdue.edu.

S. Lee and P. Karava are with the Department of Civil Engineering, Purdue University, West Lafayette, IN 47906, USA lee1904@purdue.edu, pkarava@purdue.edu.

Euclidean space; $\mathbb{R}^{n\times m}$: set of all $n\times m$ real matrices; $A^T$: transpose of matrix $A$; $\mathbb{S}^n$ (resp. $\mathbb{S}^n_+$, $\mathbb{S}^n_{++}$): set of symmetric (resp. positive semi-definite, positive definite) $n\times n$ matrices; $|S|$: cardinality of finite set $S$; $\mathbb{E}[\cdot]$: expectation operator; $\mathbb{P}[\cdot]$: probability of event.

In this paper, we consider the discrete-time stochastic system

$$\begin{aligned}\mathbf{x}(k+1) &= f(\mathbf{x}(k),\, \mathbf{u}(k),\, \mathbf{w}(k),\, \mathbf{z}(k)), \quad \mathbf{x}(0) = x_0 \in X,\\ \mathbf{z}(k+1) &\sim p(\mathbf{x}(k)), \quad \mathbf{z}(0) \sim \mu,\end{aligned} \tag{1}$$

where $k \in \mathbb{N}$ is the time step, $\mathbf{x}(k) \in X$ is the state, $X \subset \mathbb{R}^n$ is a compact state space, $\mathbf{u}(k) \in U$ is the control input, $U \subset \mathbb{R}^{m_u}$ is a compact control space, $\mathbf{w}(k) \in \mathbb{R}^{m_w}$ is a random variable representing disturbances and uncertainties, and each $\mathbf{w}(k)$ is independent of other random variables, and $(\mathbf{z}(k))_{k=0}^{\infty}$, is a stochastic process with finite states $S := \{1, 2, \ldots, |S|\}$. $\mathbf{z}(0) \sim \mu$ implies $\mathbb{P}[\mathbf{z}(0) = i] = \mu_i$, and $\mathbf{z}(k+1) \sim p(\mathbf{x}(k))$ implies that the stochastic process $\mathbf{z}(k), k \in \mathbb{N}_+$, evolves according to $\mathbb{P}[\mathbf{z}(k+1) = i|\mathbf{x}(k) = x(k)] = p_i(x(k))$ with the transition probability $p(x(k)) := \begin{bmatrix} p_1(x(k)) & \ldots & p_{|S|}(x(k))\end{bmatrix}^T$, and $x(k)$ is a realization of $\mathbf{x}(k)$. In other words, the transition probability depends on the current state $x(k)$ of (1).

We note that $(\mathbf{z}(k))_{k=0}^{\infty}$ is a special case of Markov chain, and the first convergence result of dynamic programming (DP) in this paper holds for the general Markov chain case. Note that (1) is a Markov decision process (MDP) [18], where the continuous and discrete state-spaces coexist and interact with each other.

## III. Dynamic Programming

For a given nonnegative Lebesgue measurable stage cost function, $g : \mathbb{R}^n \times \mathbb{R}^m \times S \to \mathbb{R}_+$, and control input space $U \subset \mathbb{R}^m$, the cost associated with a given admissible state-feedback control policy $\pi : X \times S \to U$ and initial states $x_0 \in X$, $z_0 \in S$, is

$$J^{\pi}(x_0,\, z_0) := \mathbb{I}_X(x_0)\mathbb{E}_{x_0,\, z_0}[\psi^{\pi}((\mathbf{x}(k),\, \mathbf{z}(k))_{k=0}^{\infty})] \tag{2}$$

where

$$\psi^{\pi}((\mathbf{x}(i),\, \mathbf{z}(i))_{i=0}^{\infty}) := \sum_{i=0}^{\tau(x_0,\, z_0;\pi)-1} \alpha^i g(\mathbf{x}(i), \mathbf{u}(i),\, \mathbf{z}(i)),$$

$\mathbf{u}(i) = \pi(\mathbf{x}(i), \mathbf{z}(i))$, $\alpha \in [0,\, 1)$ is called the discount factor, $\mathbb{I}_X$ is the indicator function, $\tau(x_0,\, z_0; \pi)$ is the first time instant the trajectory $\mathbf{x}(k)$ exits $X$ given $\mathbf{x}(0) = x_0$, $\mathbf{z}(0) = z_0$, and $\mathbb{E}_{x_0,\, z_0}[\cdot]$ is a shorthand notation for $\mathbb{E}[\cdot|\mathbf{x}(0) = x_0,\, \mathbf{z}(0) = z_0]$. The set of all admissible state-feedback control policies is denoted by $\Pi$. In addition, we make the following assumption.

*Assumption 1:* The cost per stage $g$ satisfies $|g(x,\, u,\, i)| \leq M$ for all $(x,\, u,\, i) \in X \times U \times S$, where $M$ is some scalar.

Under Assumption 1, the quantity (2) is always finite, and hence well defined. The optimal cost is

$$J^*(x_0,\, z_0) := \inf_{\pi\in\Pi} J^{\pi}(x_0,\, z_0).$$

For any bounded function $J$ such that $J(x_0, z_0) = 0$, $\forall (x_0, z_0) \in X^c \times S$, define the operator

$$\begin{aligned}&(TJ)(x_0,\, z_0)\\ &:= \mathbb{I}_X(x_0) \inf_{u\in U} \mathbb{E}_{x_0,z_0}\left[g(x_0, u, z_0) + \alpha J(\mathbf{x}(1), \mathbf{z}(1))\right]\\ &= \mathbb{I}_X(x) \inf_{u\in U} \left[g(x_0, u, z_0) + \alpha \sum_{j=1}^{|S|} p_j(x_0)\mathbb{E}[J(f(x_0,\, u,\, w,\, j), j)]\right].\end{aligned} \tag{3}$$

The optimal cost $J^*$ satisfies $TJ^* = J^*$, called Bellman's equation, and the sequence $(J_k)_{k=0}^{\infty}$ generated by the dynamic programming (DP) algorithm (value iteration), $J_{k+1} = TJ_k$, $J_0 \equiv 0$, converges uniformly to $J^*$ under Assumption 1.

*Theorem 1:* The sequence $(J_k)_{k=0}^{\infty}$ generated by the DP algorithm

$$J_{k+1}(x_0,\, z_0) = (TJ_k)(x_0,\, z_0), \quad (x_0,\, z_0) \in X \times S$$

with $J_0 \equiv 0$ converges to $J^*$.

*Proof:* See Appendix I ∎

*Remark 1:* Note that $(\mathbf{z}(k))_{k=0}^{\infty}$ is a special case of Markov chains, and Theorem 1 can be directly applied to the more general case where $(\mathbf{z}(k))_{k=0}^{\infty}$ is a Markov chain. A convergence result of DP for MDP, where continuous and discrete state-spaces coexist and interact with each other, was addressed in [16, Theorem 2,Theorem 3]. However, the proof in [16] cannot be directly applied to our case because for the system in (1), the MDP has stochastic disturbances in continuous spaces.

If $J^*$ is known, then the optimal state-feedback control policy can be computed as

$$u^*(x_0,\, z_0) := \arg\inf_{u\in U} \mathbb{E}_{x_0,z_0}\left[g(x_0, u, z_0) + \alpha J^*(\mathbf{x}(1), \mathbf{z}(1))\right] \tag{4}$$

provided that the infimum is attained. Moreover, Q-factor [35] is defined as

$$\begin{aligned}&Q^*(x_0,\, z_0,\, u)\\ &:= \mathbb{I}_X(x_0)\mathbb{E}_{x_0,z_0}\left[g(x_0, u, z_0) + \alpha J^*(\mathbf{x}(1), \mathbf{z}(1))\right].\end{aligned} \tag{5}$$

By comparing this definition with (4), the optimal policy can be expressed as $u^*(x_0,\, z_0) := \arg\inf_{u\in U} Q^*(x_0,\, z_0,\, u)$. In addition, one has $J^*(x_0, z_0) = \inf_{u\in U} Q^*(x_0,\, z_0,\, u)$. Similarly to $T$, if we define the operator $F$

$$\begin{aligned}&(FQ)(x_0,\, z_0,\, u)\\ &:= \mathbb{I}_X(x_0)\mathbb{E}_{x_0,z_0}\left[g(x_0, u, z_0) + \alpha \inf_{\bar{u}\in U} Q(\mathbf{x}(1), \mathbf{z}(1),\, \bar{u})\right],\end{aligned}$$

then, (5) can be written as $Q^* = FQ^*$, which is equivalent to the Bellman equation. The Q-value iteration, $Q_{k+1} = FQ_k$, $Q_0 \equiv 0$, generates sequence $(Q_k)_{k=0}^{\infty}$ that converges to $Q^*$ under the same condition as in the DP.

In the building control problem of our interest, $\mathbf{z}(k)$ describes occupant thermal preferences. Therefore, it is practical to assume that $\mathbf{z}(k)$ is not available in real time.

*Assumption 2:* $\mathbf{x}(k)$ is measured in real time, but $\mathbf{z}(k)$ cannot be measured.

To design an optimal control policy under Assumption 2, (2) is modified as

$$J^{\pi}(x_0) := \mathbb{I}_X(x_0)\mathbb{E}_{x_0}\left[\sum_{k=0}^{\tau(x_0;\pi)-1} \alpha^k g(\mathbf{x}(k), \mathbf{u}(k), \mathbf{z}(k))\right],$$

where $\tau(x_0;\pi)$ is the first time instant the trajectory $\mathbf{x}(k)$ exits $X$ given $\mathbf{x}(0) = x_0$. Consider the optimal cost

$$J^*(x) := \inf_{\pi\in\Pi} J^{\pi}(x). \tag{6}$$

In this case, the operator (3) and the Bellman equation cannot be well formed because the next state evolution cannot be entirely determined based on the current state information, i.e., the Markov property does not hold. However, we can construct an augmented system that satisfies the Markov property. In particular, Figure 1 shows a graph which describes the dependencies of random variables. From the figure, it is clear that the augmented state vector $\tilde{\mathbf{x}}(k) = \begin{bmatrix}\mathbf{x}(k) & \mathbf{x}(k+1)\end{bmatrix}^T$ has enough information to determine the distributions of $\mathbf{x}(k+2)$. Define $\tilde{\mathbf{w}}(k) := \mathbf{w}(k+1)$, $\tilde{\mathbf{z}}(k) :=$

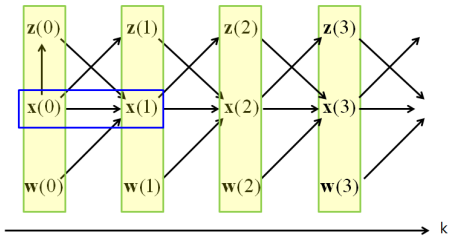


Fig. 1. Graph describing dependencies of random variables.

$\mathbf{z}(k+1)$, $\tilde{\mathbf{u}}(k) := \mathbf{u}(k+1)$, $k \in \{0, 1, 2, \ldots\}$, and define the corresponding stage cost such that $\tilde{g}(\tilde{\mathbf{x}}(k), \tilde{\mathbf{u}}(k), \tilde{\mathbf{z}}(k)) = g(\mathbf{x}(k+1), \mathbf{u}(k+1), \mathbf{z}(k+1))$, $k \in \{0, 1, 2, \ldots\}$. Then, the augmented state satisfies

$$\begin{aligned}\tilde{\mathbf{x}}(k+1) &= \tilde{f}(\tilde{\mathbf{x}}(k), \tilde{\mathbf{u}}(k), \tilde{\mathbf{w}}(k), \tilde{\mathbf{z}}(k)) \\ &:= \begin{bmatrix}\mathbf{x}(k+1) \\ f(\mathbf{x}(k+1), \mathbf{u}(k+1), \mathbf{w}(k+1), \mathbf{z}(k+1))\end{bmatrix}.\end{aligned}$$

Define the corresponding cost function

$$\tilde{J}^{\pi}(\tilde{x}_0) := \mathbb{I}_{X\times X}(\tilde{x}_0)\mathbb{E}_{\tilde{x}_0}\left[\sum_{k=0}^{\tau(\tilde{x}_0;\pi)-1} \alpha^k \tilde{g}(\tilde{\mathbf{x}}(k), \tilde{\mathbf{u}}(k), \tilde{\mathbf{z}}(k))\right],$$

where $\tilde{x}_0 \in X \times X$, and $\tilde{J}^*(\tilde{x}_0) := \inf_{\pi\in\Pi} \tilde{J}^{\pi}(\tilde{x}_0)$. If the distribution of $\mathbf{z}(k+1)$ depends only on partial coordinates of $\mathbf{x}(k)$, i.e., $P\mathbf{x}(k)$ where $P$ is a projection matrix that projects onto the partial coordinates, then the augmented state can be replaced with $\tilde{\mathbf{x}}(k) = \begin{bmatrix}P\mathbf{x}(k) & \mathbf{x}(k+1)\end{bmatrix}^T$, and the augmented system can be defined as

$$\begin{aligned}\tilde{\mathbf{x}}(k+1) &= \tilde{f}(\tilde{\mathbf{x}}(k), \tilde{\mathbf{u}}(k), \tilde{\mathbf{w}}(k), \tilde{z}(k)) \\ &:= \begin{bmatrix}P\mathbf{x}(k+1) \\ f(\mathbf{x}(k+1), \mathbf{u}(k+1), \mathbf{w}(k+1), \mathbf{z}(k+1))\end{bmatrix}.\end{aligned}$$

Now, we obtain a system of the form (1), and the result in Theorem 1 can be directly applied.

## IV. Building Control with Occupant Interactions

### A. Building Model

In this paper, we consider a $3\text{m} \times 3\text{m}$ private office space with a $2.5\text{m}^2$ south facing window, and its RC (resistor-capacitor) circuit analogy is given in Figure 2. To reduce the order of the model, we use one node for air in the room and another node collecting all the thermal mass in the room, where $T_\text{a}$ is the air temperature ($^\circ C$), $T_\text{o}$ is the outdoor air

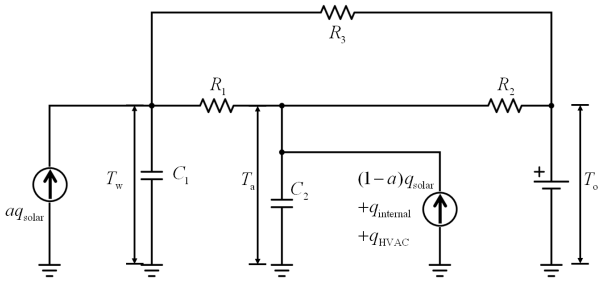


Fig. 2. RC circuit analogy

temperature ($^\circ C$), $T_\text{w}$ is the temperature of the aggregated mass node ($^\circ C$), $q_\text{solar}$ is the solar radiation ($W$), $q_\text{internal}$ is the internal heat ($W$), $q_\text{HVAC}$ is the heating/cooling rate of the HVAC system ($W$). We assume that the room is conditioned by a VAV system so that $q_\text{HVAC}$ directly affects $T_\text{a}$. Since we use low order model, we assume that the air node includes some portion of surfaces in the room which absorb radiative heat and release the heat quickly to the air. To determine appropriate values of the parameters of the circuit, we conducted a building energy simulation with EnergyPlus 8.7.0 in [36], and estimated the parameters minimizing the root-mean-square error between the air temperatures calculated by the EnergyPlus simulation and the low order model. The values of parameters are summarized in Table I. The dynamic system model is given as

TABLE I
VALUES OF THE PARAMETERS OF THE CIRCUIT IN FIGURE 2

| Parameter | Value | Unit |
|---|---|---|
| $R_1$ | 0.0084197 | $^\circ C/W$ |
| $R_2$ | 0.044014 | $^\circ C/W$ |
| $R_3$ | 4.38 | $^\circ C/W$ |
| $C_1$ | 9861100 | $J/^\circ C$ |
| $C_2$ | 128560 | $J/^\circ C$ |
| $a$ | 0.55 | – |

$$\begin{aligned}C_2\dot{T}_\text{a}(t) &= \frac{T_\text{o}(t) - T_\text{a}(t)}{R_2} + \frac{T_\text{w}(t) - T_\text{a}(t)}{R_1} \\ &+ (1-a)q_\text{solar}(t) + q_\text{HVAC}(t) + q_\text{internal}(t), \\ C_1\dot{T}_\text{w}(t) &= \frac{T_a(t) - T_w(t)}{R_1} + \frac{T_\text{o}(t) - T_\text{w}(t)}{R_3} + aq_\text{solar}(t).\end{aligned}$$

A discrete time representation can be obtained by using the Euler discretization with a sampling time of $\Delta t$

$$T_\text{a}(k+1) - T_\text{a}(k) = \frac{\Delta t}{C_2R_2}(T_\text{o}(k) - T_\text{a}(k))$$

$$+\frac{\Delta t}{C_2R_1}(T_{\rm w}(k)-T_{\rm a}(k))+\frac{\Delta t(1-a)}{C_2}q_{\rm solar}(k)$$
$$+\frac{\Delta t}{C_2}q_{\rm HVAC}(k)+\frac{\Delta t}{C_2}q_{\rm internal}(k),$$
$$T_{\rm w}(k+1)-T_{\rm w}(k)=\frac{\Delta t}{C_1R_1}(T_{\rm a}(k)-T_{\rm w}(k))$$
$$+\frac{\Delta t}{C_1R_3}(T_{\rm o}(k)-T_{\rm w}(k))+\frac{\Delta ta}{C_1}q_{\rm solar}(k),$$

where $k \in \mathbb{N}$ is the discrete time step. In this paper, we consider $\Delta t = 10$min sampling time. In the building control literature, the time step is usually chosen to be $\Delta t = 30$min. The reason we consider finer time steps is for quicker responses to occupant's actions.

Now, we assume that there is an occupant in the room, and the occupant's stochastic behavior affects the system dynamics. In particular, define the stochastic process $(\mathbf{z}(k))_{k=0}^{\infty}$ with the state space $S = \{1, 2, 3\}$, which represents the occupant's feeling of cold, comfort, and hot, respectively. Its probability depends on the current indoor temperature $T_{\rm a}(k)$, and its probability density function $p_{\mathbf{z}}(z; T_{\rm a})$ is obtained by the Bayesian modelling approach in [37]. The values of the probability for different values of $T_{\rm a}$ are depicted in Figure 3. Consider some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{A}$ be a space

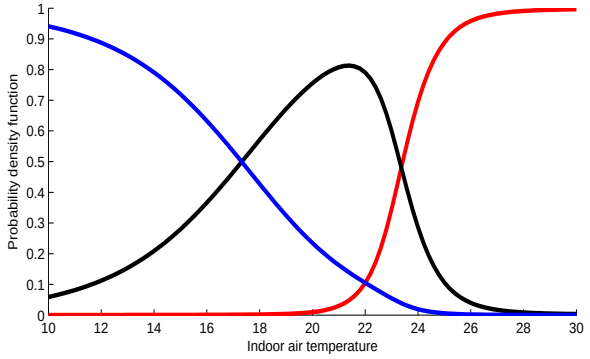


Fig. 3. The probability density function $p_{\mathbf{z}}(1; T_{\rm a})$ (blue), $p_{\mathbf{z}}(2; T_{\rm a})$ (black), $p_{\mathbf{z}}(3; T_{\rm a})$ (red) for different $T_{\rm a}$

of occupant's actions, and let $\mathcal{I}$ be some information space. The information space $\mathcal{I}$ is a set of variables that affect occupant actions. For example, the values of $\mathbf{z}(k)$ can be an element of $\mathcal{I}$ because it is used to induce occupant actions. The occupant's actions are modelled as a map $M : \mathcal{I} \times \Omega \to \mathcal{A}$. We consider one possible scenario of occupant's actions described below.

Occupant's overriding on current temperature set point: The occupant can use a control panel to increase, decrease, or maintain the current temperature. The reference signal has the dynamic equation $T_{\rm ref}(k+1) = T_{\rm ref}(k) + M(\mathbf{z}(k))$, where $T_{\rm ref}(k)$ is the current reference signal and $M(\mathbf{z}(k))$ is occupant's control input. For example, if $\mathbf{z}(k) = 1$ or $\mathbf{z}(k) = 3$, the occupant changes the set point with probability 0.5. In particular, the map $M(\mathbf{z}(k)) \in \{-1, 0, 1\}$ is given by

$$M(\mathbf{z}(k)) = \begin{cases} \omega(k), & \text{if } \mathbf{z}(k) = 1 \\ 0, & \text{if } \mathbf{z}(k) = \text{comfort} \\ -\omega(k), & \text{if } \mathbf{z}(k) = 3 \end{cases},$$

where

$$\omega(k) = \begin{cases} 0 & \text{with probability } 0.5 \\ 1 & \text{with probability } 0.5 \end{cases}.$$

We assume that the temperature set point varies within the range $15 \le T_{\rm ref}(k) \le 30$. The internal heat due to electronic appliances and occupant's body is given by $q_{\rm internal}(k) = 145$ $(W)$. In summary, one obtains a state-space model $x(k+1) = Ax(k) + Bu(k) + Dw(k)$ with $u(k) = q_{\rm HVAC}(k)$,

$$x(k) = \begin{bmatrix} T_{\rm a}(k) \\ T_{\rm a}(k+1) \\ T_{\rm w}(k+1) \\ T_{\rm ref}(k+1) \end{bmatrix}, \; w(k) = \begin{bmatrix} q_{\rm solar}(k+1) \\ q_{\rm internal}(k+1) \\ T_{\rm o}(k+1) \\ M(\mathbf{z}(k+1)) \end{bmatrix},$$

and

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1-\frac{\Delta t}{C_2R_2}-\frac{\Delta t}{C_2R_1} & \frac{\Delta t}{C_2R_1} & 0 \\ 0 & \frac{\Delta t}{C_1R_1} & 1-\frac{\Delta t}{C_1R_3}-\frac{\Delta t}{C_1R_1} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 \\ \frac{\Delta t}{C_2} \\ 0 \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \frac{\Delta t(1-a)}{C_2} & \frac{\Delta t}{C_2} & \frac{\Delta t}{C_2R_2} & 0 \\ \frac{\Delta ta}{C_1} & 0 & \frac{\Delta t}{C_1R_3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The stage cost function is set to be $g(\mathbf{x}(k), \mathbf{u}(k)) = (\mathbf{T}_{\rm in}(k+1) - \mathbf{T}_{\rm ref}(k+1))^2 + 0.0001\mathbf{u}(k)^2$. In addition, we consider the space space $\mathbf{T}_{\rm in}(k) \in [10, 30]$, $\mathbf{T}_{\rm in}(k+1) \in [10, 30]$, $\mathbf{T}_{\rm w}(k+1) \in [10, 30]$, $\mathbf{T}_{\rm ref}(k+1) \in [10, 30]$, and the control space is $\mathbf{u}(k) \in \{-1000, 0, 1000\}$.

### *B. Approximate dynamic programming*

Consider the Q-factor $Q(x, u) = Q_{\rm LTI}(x, u) + \hat{Q}(x, u; \theta)$, where $Q_{\rm LTI}$ is the Q-factor obtained based on the LTI system $(A, B)$, and $\hat{Q}(x, u; \theta)$ is an additive term to be determined in order to compensate the first term. Note that $Q_{\rm LTI}(x, u)$ is fixed and can be exactly computed by using classical LQR results. We apply ADP with the linear function approximation $\hat{Q}(x, u; \theta) = \phi(x, u)^T\theta$ (see [19] for details) with vectors $\phi(x, u) \in \mathbb{R}^q$ and $\theta \in \mathbb{R}^q$, $q \in \mathbb{N}_+$, defined as follows: Each element of the vector $\phi(x, u)$ is called a feature; $\phi_i(x, u)$ denotes the value of feature $i$ for state-control input pair $(x, u)$. The feature function $\phi : X \times U \to \mathbb{R}^q$ maps each state-control input pair to a vector of feature values; $\theta \in \mathbb{R}^q$ is the weight vector specifying the contribution of each feature across all state-control input pairs. We will use Gaussian radial basis functions as feature functions of the Q-factor, i.e.,

$$\phi_j(x, u) := \exp\left(-\left\| \begin{bmatrix} x \\ u \end{bmatrix}^2 - c_j \right\|^2 / 2\mu_j^2\right),$$

with centers $c_j$ manually placed in the state-control input space. The number of feature functions used is 2700. To

solve the optimal control problem, we apply the trajectory-based value iteration (TBVI) [19], which is a simulation-based fast ADP algorithm for large-scale problems. Simulation results are given in Figure 4, where the first figure depicts $\mathbf{T}_{\text{in}}(k)$ (black solid line) and $\mathbf{T}_{\text{ref}}(k)$ (blue dashed line), the second figure is a realization of $\mathbf{z}(k)$, and the third one is the control input history $\mathbf{u}(k)$. Histograms of costs of the two methods are compared in Figure 5 with total $1000$ simulations. The average cost of the proposed control is $168$, while $245$ for the LQR control. The cost saving is due to the occupant behaviors, which are considered by the proposed ADP method but not by the LQR control.

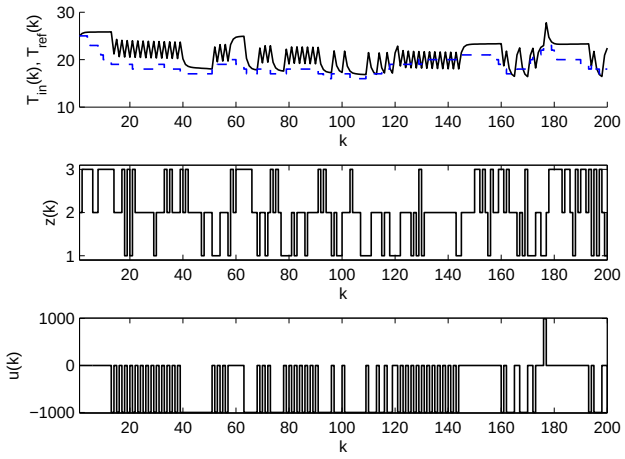


Fig. 4. Simulation results: (1) $\mathbf{T}_{\text{in}}(k)$ (black solid line) and $\mathbf{T}_{\text{ref}}(k)$ (blue dashed line); (2) window open/close state; (3) realization of $\mathbf{z}(k)$; (4) control input history $\mathbf{u}_{\text{ac}}(k)$

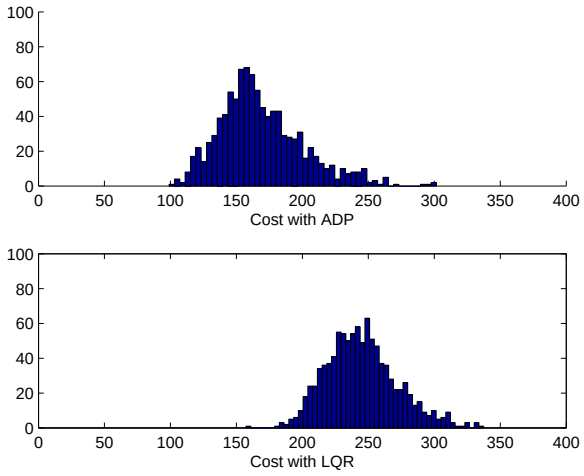


Fig. 5. Cost histograms of LQR and proposed control.

## Conclusion

In this paper, we have studied infinite-horizon ADP for building control problems with occupant interactions. Since occupant thermal preferences and actions are dependent on state variables of the building model, so are their probability density functions as well. In this case, existing stochastic MPC approaches including scenario-based MPC are not applicable. ADP is one possible approach to solve optimal control problems for these systems. Through simulation studies, we have demonstrated that the ADP is suitable for building control problems with occupant interactions.

## References


[1] A. I. Dounis and C. Caraiscos, "Advanced control systems engineering for energy and comfort management in a building environment–A review," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 6, pp. 1246–1261, 2009.

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

[3] Y. Ma, A. Kelman, A. Daly, and F. Borrelli, "Predictive control for energy efficient buildings with thermal storage," *IEEE Control system magazine*, vol. 32, no. 1, pp. 44–64, 2012.

[4] Y. Ma, F. Borrelli, B. Hencey, B. Coffey, S. Bengea, and P. Haves, "Model predictive control for the operation of building cooling systems," *IEEE Transactions on Control Systems Technology*, vol. 20, no. 3, pp. 796–803, 2012.

[5] A. Bemporad and M. Morari, "Robust model predictive control: A survey," in *Robustness in identification and control*. Springer, 1999, pp. 207–226.

[6] D. Chatterjee, P. Hokayem, and J. Lygeros, "Stochastic receding horizon control with bounded control inputs: a vector space approach," *IEEE Transactions on Automatic Control*, vol. 56, no. 11, pp. 2704–2710, 2011.

[7] J. A. Primbs and C. H. Sung, "Stochastic receding horizon control of constrained linear systems with state and control multiplicative noise," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 221–230, 2009.

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

[9] F. Oldewurtel, A. Parisio, C. N. Jones, M. Morari, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and K. Wirth, "Energy efficient building climate control using stochastic model predictive control and weather predictions," in *American control conference (ACC), 2010*, 2010, pp. 5100–5105.

[10] F. Oldewurtel, A. Parisio, C. N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari, "Use of model predictive control and weather forecasts for energy efficient building climate control," *Energy and Buildings*, vol. 45, pp. 15–27, 2012.

[11] L. Blackmore, A. Bektassov, M. Ono, and B. C. Williams, "Robust, optimal predictive control of jump markov linear systems using particles," in *International Workshop on Hybrid Systems: Computation and Control*, 2007, pp. 104–117.

[12] L. Blackmore, M. Ono, A. Bektassov, and B. C. Williams, "A probabilistic particle-control approximation of chance-constrained stochastic predictive control," *IEEE transactions on Robotics*, vol. 26, no. 3, pp. 502–517, 2010.

[13] G. C. Calafiore and L. Fagiano, "Stochastic model predictive control of LPV systems via scenario optimization," *Automatica*, vol. 49, no. 6, pp. 1861–1866, 2013.

[14] G. Schildbach, L. Fagiano, C. Frei, and M. Morari, "The scenario approach for stochastic model predictive control with bounds on closed-loop constraint violations," *Automatica*, vol. 50, no. 12, pp. 3009–3018, 2014.

[15] G. C. Calafiore and L. Fagiano, "Robust model predictive control via scenario optimization," *IEEE Transactions on Automatic Control*, vol. 58, no. 1, pp. 219–224, 2013.

[16] I. V. Kolmanovsky, L. Lezhnev, and T. L. Maizenberg, "Discrete-time drift counteraction stochastic optimal control: Theory and application-motivated examples," *Automatica*, vol. 44, no. 1, pp. 177–184, 2008.
[17] L. Johannesson, M. Asbogard, and B. Egardt, "Assessing the potential of predictive control for hybrid vehicle powertrains using stochastic dynamic programming," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 71–83, 2007.
[18] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996.
[19] A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, J. P. How *et al.*, "A tutorial on linear function approximators for dynamic programming and reinforcement learning," *Foundations and Trends® in Machine Learning*, vol. 6, no. 4, pp. 375–451, 2013.
[20] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 1998.
[21] A. Aswani, N. Master, J. Taneja, D. Culler, and C. Tomlin, "Reducing transient and steady state electricity consumption in hvac using learning-based model-predictive control," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 240–253, 2012.
[22] F. Oldewurtel, D. Sturzenegger, and M. Morari, "Importance of occupancy information for building climate control," *Applied energy*, vol. 101, pp. 521–532, 2013.
[23] J. R. Dobbs and B. M. Hencey, "Model predictive hvac control with online occupancy model," *Energy and Buildings*, vol. 82, pp. 675–684, 2014.
[24] J. Page, D. Robinson, N. Morel, and J.-L. Scartezzini, "A generalised stochastic model for the simulation of occupant presence," *Energy and buildings*, vol. 40, no. 2, pp. 83–98, 2008.
[25] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPSs," *arXiv preprint arXiv:1507.06527*, 2015.
[26] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[27] M. Riedmiller, "Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method," in *European Conference on Machine Learning*, 2005, pp. 317–328.
[28] Y. Engel, S. Mannor, and R. Meir, "Bayes meets Bellman: The Gaussian process approach to temporal difference learning," in *ICML*, vol. 20, no. 1, p. 154.
[29] G. P. Henze and J. Schoenmann, "Evaluation of reinforcement learning control for thermal energy storage systems," *HVAC&R Research*, vol. 9, no. 3, pp. 259–275, 2003.
[30] L. Yang, Z. Nagy, P. Goffin, and A. Schlueter, "Reinforcement learning for optimal control of low exergy buildings," *Applied Energy*, vol. 156, pp. 577–586, 2015.
[31] R. Hafner and M. Riedmiller, "Reinforcement learning in feedback control," *Machine learning*, vol. 84, no. 1-2, pp. 137–169, 2011.
[32] S. Liu and G. P. Henze, "Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 2: Results and analysis," *Energy and buildings*, vol. 38, no. 2, pp. 148–161, 2006.
[33] Z. Yu and A. Dexter, "Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning," *Control Engineering Practice*, vol. 18, no. 5, pp. 532–539, 2010.
[34] Z. Cheng, Q. Zhao, F. Wang, Y. Jiang, L. Xia, and J. Ding, "Satisfaction based q-learning for integrated lighting and blind control," *Energy and Buildings*, vol. 127, pp. 43–55, 2016.
[35] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995, vol. 1, no. 2.
[36] U. D. of Energy, "Energyplustm 8.7.0 documentation," https://energyplus.net/documentation, [Online Available].
[37] S. Lee, I. Bilionis, P. Karava, and A. Tzempelikos, "A bayesian approach for probabilistic classification and inference of occupant thermal preferences in office buildings," *Building and Environment*, vol. 118, pp. 323–343, 2017.

## APPENDIX I
## PROOF OF THEOREM 1

We first find another characterization of $J_k(x_0,\, z_0)$ in terms of a sum of stage cost functions.

*Lemma 1:* For any fixed $k \geq 1$, $J_k$ is described as

$$J_k(x_0,\, z_0) = \mathbb{I}_X(x_0) \inf_{\pi\in\Pi} \mathbb{E}_{x_0,\, z_0}[\psi_k^\pi((\mathbf{x}(i),\, \mathbf{z}(i))_{i=0}^\infty)],$$

where

$$\psi_k^\pi((\mathbf{x}(i),\, \mathbf{z}(i))_{i=0}^\infty) := \sum_{i=0}^{\min\{\tau(x_0,\, z_0;\pi)-1,\, k-1\}} \alpha^i g(\mathbf{x}(i), u(i),\, \mathbf{z}(i)).$$

*Proof:* The claim will be proved by an induction argument. Since $J_0 \equiv 0$, $J_1(x_0,\, z_0)$ is given by

$$\begin{aligned} &J_1(x_0,\, z_0) = (TJ_0)(x_0,\, z_0) \\ &= \mathbb{I}_X(x_0) \inf_{u\in U} \mathbb{E}_{x_0,z_0}\left[g(x_0, u, z_0)\right] \\ &= \mathbb{I}_X(x_0) \\ &\times \inf_{\pi\in\Pi} \mathbb{E}_{x_0,z_0}\left[\sum_{i=0}^{\min\{\tau(x_0,\, z_0;\pi)-1,\, 0\}} \alpha^i g(\mathbf{x}(i), u(i),\, \mathbf{z}(i))\right]. \end{aligned}$$

Now, suppose for $k \geq 2$

$$\begin{aligned} &J_{k-1}(x_0,\, z_0) \\ &= \mathbb{I}_X(x_0) \\ &\times \inf_{\pi\in\Pi} \mathbb{E}_{x_0,z_0}\left[\sum_{i=0}^{\min\{\tau(x_0,\, z_0;\pi)-1,\, k-2\}} \alpha^i g(\mathbf{x}(i), u(i),\, \mathbf{z}(i))\right] \end{aligned} \tag{7}$$

holds. Then,

$$\begin{aligned} &J_k(x_0,\, z_0) = (TJ_{k-1})(x_0,\, z_0) \\ &= \mathbb{I}_X(x_0) \inf_{u\in U} \mathbb{E}_{x_0,z_0}\left[g(x_0, u, z_0) + \alpha J_{k-1}(\mathbf{x}(1), \mathbf{z}(1))\right]. \end{aligned}$$

By conditioning on the exit time $\tau(x_0,\, z_0;\pi)$, the expectation in the last equation is expressed as

$$\begin{aligned} &\mathbb{E}_{x_0,z_0}\left[g(x_0, u(0), z_0)|\, \tau(x_0,\, z_0;\pi) = 1\right] \\ &\times \mathbb{P}[\tau(x_0,\, z_0;\pi) = 1] \\ &+ \mathbb{E}_{x_0,z_0}\left[\ \chi\ \middle|\, \tau(x_0,\, z_0;\pi) \geq 2\right] \mathbb{P}[\tau(x_0,\, z_0;\pi) \geq 2], \end{aligned} \tag{8}$$

where

$$\begin{aligned} \chi &:= g(x_0, u(0), z_0) \\ &+ \sum_{i=1}^{\min\{\tau(\mathbf{x}(1),\, \mathbf{z}(1);\pi)-1,\, k-2\}+1} \alpha^i g(\mathbf{x}(i), u(i),\, \mathbf{z}(i)), \end{aligned}$$

and the second term is obtained by the induction hypothesis (7). In the second term, the quantity $\min\{\tau(\mathbf{x}(1),\, \mathbf{z}(1);\pi) - 1,\, k-2\} + 1$ is rewritten as

$$\begin{aligned} &\min\{\tau(\mathbf{x}(1),\, \mathbf{z}(1);\pi) - 1,\, k-2\} + 1 \\ &= \min\{\tau(\mathbf{x}(1),\, \mathbf{z}(1);\pi),\, k-1\} \\ &= \min\{\tau(x_0,\, z_0;\pi) - 1,\, k-1\}. \end{aligned}$$

Therefore, (8) is identical to

$$\mathbb{E}_{x_0,z_0}\left[\sum_{i=0}^{\min\{\tau(x_0,\, z_0;\pi)-1,\, k-1\}} \alpha^i g(\mathbf{x}(i), u(i),\, \mathbf{z}(i))\right],$$

and the desired result follows. ∎

Proof of Theorem 1:

For any fixed $\pi \in \Pi$, define

$$J_k^\pi(x_0,\, z_0) := \mathbb{I}_X(x_0)\mathbb{E}_{x_0,\, z_0}[\psi_k^\pi((\mathbf{x}(i),\, \mathbf{z}(i))_{i=0}^\infty)].$$

For any $k \geq 1$ and sample path $(x(i),\, z(i))_{i=0}^\infty$, we have

$$|\psi_k^\pi((x(i),\, z(i))_{i=0}^\infty)| \leq \left| \sum_{i=0}^{\min\{\tau(x_0,\, z_0;\pi)-1,\, k-1\}} \alpha^i M \right|$$

$$\leq M \left| \sum_{i=0}^{\infty} \alpha^i \right| \leq M \frac{1}{1-\alpha}. \tag{9}$$

Therefore, $J_k^\pi(x_0,\, z_0)$ is bounded. Since $J_k^\pi(x_0,\, z_0)$ is non-decreasing in $k$, the point-wise limit $\lim_{k\to\infty} J_k^\pi(x_0,\, z_0)$ exists. Choose a $\varepsilon$-suboptimal control policy $\pi^\varepsilon \in \Pi$ such that $J^{\pi^\varepsilon}(x_0,\, z_0) \leq J^*(x_0,\, z_0) + \varepsilon$. Since $\lim_{k\to\infty} \psi_k^\pi = \psi^\pi$ pointwise and (9) holds, by the dominated convergence theorem, we have

$$\begin{aligned}
&\lim_{k\to\infty} J_k^{\pi^\varepsilon}(x_0,\, z_0) \\
&= \lim_{k\to\infty} \mathbb{I}_X(x_0)\mathbb{E}_{x_0,\, z_0}[\psi_k^{\pi^\varepsilon}((\mathbf{x}(i),\, \mathbf{z}(i))_{i=0}^\infty)] \\
&= \mathbb{I}_X(x_0)\mathbb{E}_{x_0,\, z_0}[\lim_{k\to\infty} \psi_k^{\pi^\varepsilon}((\mathbf{x}(i),\, \mathbf{z}(i))_{i=0}^\infty)] \\
&= \mathbb{I}_X(x_0)\mathbb{E}_{x_0,\, z_0}[\psi^{\pi^\varepsilon}((\mathbf{x}(i),\, \mathbf{z}(i))_{i=0}^\infty)] \\
&= J^{\pi^\varepsilon}(x_0,\, z_0).
\end{aligned} \tag{10}$$

Since $J_k(x_0,\, z_0) = \mathbb{I}_X(x_0) \inf_{\pi\in\Pi} J_k^\pi(x_0,\, z_0)$ by Lemma 1, we have $J_k(x_0,\, z_0) \leq J_k^{\pi^\varepsilon}(x_0,\, z_0)$. Combining it with (10) leads to

$$\begin{aligned}
\lim_{k\to\infty} J_k(x_0,\, z_0) &\leq \lim_{k\to\infty} J_k^{\pi^\varepsilon}(x_0,\, z_0) \\
&= J^{\pi^\varepsilon}(x_0,\, z_0) \leq J^*(x_0,\, z_0) + \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we have $\lim_{k\to\infty} J_k(x_0,\, z_0) \leq J^*(x_0,\, z_0)$. To prove the reversed direction, note that

$$J^*(x_0,\, z_0) \leq J_k(x_0,\, z_0)$$
$$+ \mathbb{I}_X(x_0)\mathbb{E}_{x_0,\, z_0}\left[ \sum_{i=\min\{\tau(x_0,\, z_0;\pi)-1,\, k-1\}+1}^{\tau(x_0,\, z_0;\pi)-1} \alpha^i g(\mathbf{x}(i), u(i),\, \mathbf{z}(i)) \right],$$

where $\pi \in \Pi$ is arbitrary. Taking the limit $k \to \infty$ on the right-hand side yields $J^*(x_0,\, z_0) \leq \lim_{k\to\infty} J_k(x_0,\, z_0)$. This completes the proof.