2-5-2018

Simulation-Based Approximate Policy Gradient and Its Building Control Application

Donghwan Lee
Illinois, donghwan@illinois.edu

Seungjae Lee
Lyles School of Civil Engineering, Purdue University, West Lafayette, Indiana USA / Center for High Performance Buildings, Ray W. Herrick Laboratories, Purdue University, West Lafayette, Indiana USA, lee1904@purdue.edu

Panagiota Karava
School of Civil Engineering and Division of Construction Engineering and Management, Purdue University, United States of America, pkarava@purdue.edu

Jianghai Hu
Purdue University, jianghai@purdue.edu

Simulation-Based Approximate Policy Gradient and Its Building Control Application*

Donghwan Lee †, Seungjae Lee, Panagiota Karava ‡, and Jianghai Hu§

February 5, 2018

Abstract

The goal of this paper is to study the potential applicability of a stochastic approximation-based policy gradient method for optimal office building HVAC (Heating, Ventilation, and Air Conditioning) control systems. A real-world building thermal dynamics with occupant interactions is the main focus of this paper. It is a complex stochastic system in the sense that its statistical properties depend on its state variables. In this case, existing approaches, for instance, stochastic model predictive control methods, cannot be applied to optimal control designs. As a remedy, we approximate the gradient of the cost function using simulations and use a gradient descent type algorithm to design a suboptimal control policy. We assess its performance through a simulation study of building HVAC systems.

1 Introduction

The goal of this paper is to study a stochastic approximate algorithm for stochastic optimal control designs and assess its applicability to building climate control scenarios, which are important stochastic control applications. Recently, there has been a great amount of research interest in energy consumption and comfort management in buildings [1]. One of their main goals is to balance between the energy consumption and occupants' comfort in work environments. The presence of stochastic uncertainties and disturbances, such as weather and occupant interactions, is a major concern in building environment research as they degrade the performance of the control systems.

This paper considers the building control problem with a particular focus on occupant interactions. The role of occupants is significant in the thermal dynamics of building spaces [2–6]. In particular, the thermal preferences of occupants induce their actions, which potentially perturb the thermal dynamics of building spaces. It is a special class of complex stochastic systems in the sense that the statistical behavior of the occupant's actions interact with the system evolution: occupant thermal preference models [7–9] depend on environmental factors, for example, the indoor air temperature. For this reason, developments of effective stochastic control methods become of prime importance.

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

†D. Lee is with the Department of Mechanical Science and Engineering, University of Illinois, Urbana-Champaign, IL 61801, USA lee1923@purdue.edu,donghwan@illinois.edu.

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

§J. Hu is with the Department of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906, USA jianghai@purdue.edu.

Literature Review: Optimal control designs for stochastic systems have been an active area of research during the last decades. A classical approach is the linear quadratic Gaussian (LQG) control, which computes optimal control policy for linear time-invariant (LTI) systems with Gaussian random disturbances. Another main research direction is stochastic model predictive control (SMPC) [10]. The SMPC was extensively studied in [11–14] for building control systems with Gaussian random disturbances and without occupant interactions. However, Gaussian disturbances cannot describe more complicated behavior of real-world systems, and it is of great importance to develop optimal control designs for systems with more generic stochastic disturbances to meet practical needs. For example, systems with disturbances/uncertainties that depend on Markov chains arise in many applications, for instance, the vehicle path-planning [15, 16], macroeconomic model [17], economic models of government expenditure [18], vehicle controls with driver's behavior models [19, 20], and hybrid electric vehicle powertrain management [21]. For building control systems with Markov chain occupancy models, SMPCs were developed in [5, 22], and a nonlinear MPC was used in [23] with occupancy pattern models expressed as hidden Markov models. Another tractable approach is the scenario-based (or sample-based) approximation approaches [15,16,24–27]. Its advantage is the ability to cope with generic probability distributions as long as a sufficient number of random samples can be obtained. The approach was successfully applied to many engineering applications, for instance, robot path-planning problems in [15, 16] and the aircraft conflict detection in [24]. For the building problems, the scenario approaches were investigated in [28], where samples of the external temperature, the solar radiation, and the room occupancy are generated by using an empirical statistic model. Approximate dynamic programming (ADP) [29–31] (or reinforcement learning [31] from the machine learning context) is another possibility. For building applications, ADP was studied in several researches, for instance [32–35]. In [32], a closed-loop satisfaction based system is developed for optimal control strategy of blinds and lights. In this approach, users' feelings are sent to the system via a human-machine interface to construct comfort models, which are time-dependent and deterministic.

Challenges: However, the existing results did not consider occupant behavior models that depend on the dynamics of the building spaces, for example, the indoor air temperature. In building environment research, advanced occupant thermal preference models have been developed, e.g., [7–9], where occupant's thermal preferences are expressed as probability mass functions that depend on environmental factors, for example, the indoor air temperature. The optimal control of such stochastic systems cannot be formulated as easily solvable optimization problems. For instance, the scenario-based control design schemes are problematic in this case. Consider the case that probabilities of random disturbances/uncertainties depend on the current state variables. Then, the probabilities depend on the control policy and the corresponding design parameters as well. Therefore, state samples generated by using fixed controller parameters cannot reflect changes of the controller parameters themselves. Beside the building problems, such cases arise in many applications. For example, stochastic systems with disturbances modelled by Markov chains were considered in [21, 36] for hybrid electric vehicle powertrain management problems, where transition probabilities of the Markov chain depend on the state of the dynamic system. These problems were addressed in [21, 36] by using ADP.

Statement of Contributions: We apply the convolutional function smoothing and stochastic gradient approximation procedure [37–40] to the aforementioned stochastic optimal control problem. The smoothed function is an approximation of the original one, and described by an expectation of the original function with respect to a random variable that perturbs its optimization parameters. Benefits are two-fold: the smoothed approximation is differentiable, and its gradient can be stochastically approximated by sample means. Similarly to the scenario-based control design

approaches, this method uses samples of random variables. A main difference relies on the fact that in most scenario-based approximations, the samples are used in optimization formulations, while in our approach, they are used to estimate the cost function gradient. Then, a gradient descent type algorithm is applied to the cost function minimization of the finite-horizon optimal control problem. More recent advances of the convolutional function smoothing procedure can be found in [41] for convex optimizations and in [40] for more rigorous and general analysis of the approach. The proposed approach can be regarded as a class of policy gradient RL methods [42–45] with pros and cons compared to existing approaches. In addition, the proposed method is applied to building HVAC (Heating, Ventilation, and Air Conditioning) system control problems, and the validity of the proposed approach is evaluated through a simulation study. We apply an occupant thermal preference model based on an advanced Bayesian modelling approach in [9], where the thermal preferences are expressed as probability mass functions that depend on the indoor air temperature. The occupant's feeling or cognition induces actions that perturb thermal dynamics. To meet practical needs, we also consider the output-feedback control where only partial information of the state vector is available. The technical report is an extension of the conference version [46].

2 Preliminaries

The adopted notation is as follows: $\mathbb{N}$ and $\mathbb{N}_+$: sets of nonnegative and positive integers, respectively; $\mathbb{R}$: set of real numbers; $\mathbb{R}_+$: set of nonnegative real numbers; $\mathbb{R}_{++}$: set of positive real numbers; $\mathbb{R}^n$: n-dimensional Euclidean space; $\mathbb{R}^{n\times m}$: set of all $n \times m$ real matrices; A^T: transpose of matrix A; $\|\cdot\|$: any norm of a vector or a matrix; for a set $\mathcal{S}$, $|\mathcal{S}|$: cardinality of the set S; $\mathbb{E}\{\cdot\}$: expectation operator; $\mathcal{N}(x, \Sigma)$: Gaussian distribution with mean x and covariance matrix Σ; $\Pi_U(\cdot)$: projection onto set U; w.p.: "with probability"; w.r.t.: "with respect to."

Consider the discrete-time stochastic system

$$\mathbf{x}(k+1) = f(\mathbf{x}(k), \mathbf{u}(k), \mathbf{w}(k)), \quad \mathbf{x}(0) = z \in \mathbb{R}^n, \tag{1}$$

where $k \in \{1, 2, \dots, N\}$ is the finite time step, $\mathbf{x}(k) \in \mathbb{R}^n$ is the state, $\mathbf{u}(k) \in U$ is the control input, U is a compact set, $\mathbf{w}(k) \in W$ is a random/deterministic variable representing disturbances and uncertainties with a certain distribution, and W is a compact set. Since U and W are bounded, there exists a compact set $X \in \mathbb{R}^n$ such that $\mathbf{x}(k) \in X$ for all $k \in \{1, 2, \dots, N\}$. Therefore, without loss of generality, we can assume that the state-space is X. In this paper, we consider the finite-horizon stochastic optimal control problem.

Problem 1 (Optimal control problem (OCP))**.** *Denote by*

$$\mathcal{F}_k := \{\mathbf{w}(0), \dots, \mathbf{w}(k-1), \mathbf{x}(0), \dots, \mathbf{x}(k), \mathbf{u}(0), \dots, \mathbf{u}(k-1)\}$$

the history of the system until time k, and define the stage cost function $c_k : X \times U \to \mathbb{R}_+$ for all $k \in \{1, \dots, N\}$. For a given initial state $z \in \mathbb{R}^n$, solve for $(u_k(\mathcal{F}_k))_{k=0}^{N-1}$

$$(u_k^*(\mathcal{F}_k))_{k=0}^{N-1} := \underset{(u_k(\mathcal{F}_k))_{k=0}^{N-1}}{\arg\min}\ \mathbb{E}_{z\sim\mu}\left\{\sum_{k=0}^{N} c_k(\mathbf{x}(k), u_k(\mathcal{F}_k))\right\},$$

where $(\mathbf{x}(k))_{k=0}^{N}$ is a stochastic process which obeys the dynamics in (1), $\mathbf{x}(0) = z \sim \mu$ indicates that the initial state $\mathbf{x}(0)$ follows a certain probability distribution μ which is defined in a bounded sample space, $u_k(\mathcal{F}_k)$ are maps from $\mathcal{F}_k$ to U, and the terminal cost c_N only depends on the state. The minimization is taken over a set of all maps from $\mathcal{F}_k$ to U.

The goal is to solve Problem 1 approximately by using a class of approximate policy gradient methods [42–45]. In particular, we consider the parameterized state-feedback policy $u_k : \mathbb{R}^n \times \mathcal{C} \to U$

$$u_k(\mathcal{F}_k) = \Pi_U(\pi_k(x(k);\theta)), \quad k \in \{1, \dots, N-1\}, \tag{2}$$

where $\pi_k : \mathbb{R}^n \times \mathcal{C} \to \mathbb{R}^m$ is a state-feedback control policy, $\theta \in \mathcal{C}$ is a parameter vector to be determined, $\mathcal{C} \subseteq \mathbb{R}^q$ is a convex set, and $\Pi_U(\cdot)$ is the projection onto $\mathcal{C}$. The projection operator is used to ensure $u_k(\mathcal{F}_k) \in U$. A simplified stochastic optimal control problem is given as follows.

Problem 2 (Simplified OCP (SOCP)). *For a given initial state $z \in \mathbb{R}^n$, solve for $\theta \in \mathcal{C}$*

$$\theta^* := \arg\min_{\theta \in \mathcal{C}} J(\theta),$$

where

$$J(\theta) := \mathbb{E}_{z\sim\mu}\left\{\sum_{k=0}^{N} c_k(\mathbf{x}(k), u_k(\mathbf{x}(k);\theta))\right\}. \tag{3}$$

3 Smoothing and stochastic approximation

Since the probability mass function depends on θ, it is difficult to compute, if exists, the gradient of $J(\theta)$ with respect to θ. Moreover, in many practical applications, J is not smooth, and the gradient of $J(\theta)$ does not exist for some $\theta \in \mathcal{C}$. In this paper, we consider the generic scenario where J is non-smooth, non-convex, and even the evaluation of the function value $J(\theta)$ at a single point $\theta \in \mathcal{C}$ is almost impossible numerically as well as analytically. To this end, the convolution function smoothing approach in [37] is applied. In particular, we consider the smoothed function approximate of J given by the convolution

$$\hat{J}_\beta(\theta) := \int_{\mathbb{R}^q} J(\theta - \eta) h_\beta(\eta) d\eta = \mathbb{E}_\eta\{J(\theta - \eta)\},$$

where $\mathbb{E}_\eta\{\cdot\}$ implies that the expectation is taken with respect to η, the kernel function h_β is a probability mass function satisfying certain properties (see [37], [39] for details), and $\beta > 0$ is a parameter that controls the dispersion of h_β. It is known that $\hat{J}_\beta(\theta)$ is differentiable even when $J(\theta)$ is not. Moreover, for a sufficiently large β, $\hat{J}_\beta(\theta)$ tends to become convex, and non-convex optimization method based on this property was investigated in [39]. On the other hand, for a sufficiently small β, $\hat{J}_\beta$ approximates J with arbitrarily small errors. In this respect, we further simplify the problem by approximating J into a smooth function $\hat{J}_\beta$ (but possibly not convex).

Problem 3 (β-Approximated SOCP (β-ASOCP)). *For a given initial state $z \in \mathbb{R}^n$ and parameter $\beta > 0$, solve for $\theta \in \mathcal{C}$*

$$\theta^* := \arg\min_{\theta \in \mathcal{C}} \hat{J}_\beta(\theta).$$

One of the possible choices for the kernel h_β is the Gaussian density function

$$h_\beta(\eta) = \frac{1}{(2\pi)^{n/2}\beta^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{q} (\eta_i/\beta)^2\right),$$

i.e., $\eta \sim \mathcal{N}(0, \beta^2 I_q)$. In this case, $\hat{J}_\beta$ can be described by $\hat{J}_\beta(\theta) = \mathbb{E}_\eta\{J(\theta+\beta\eta)\}$, where $\eta \sim \mathcal{N}(0, I_q)$. Its exact gradient has the simple form

$$\nabla_\theta \hat{J}_\beta(\theta) = \frac{1}{\beta}\mathbb{E}_\eta\{J(\theta+\beta\eta)\eta\} = \frac{1}{\beta}\mathbb{E}_\eta\{[J(\theta+\beta\eta) - J(\theta)]\eta\}. \tag{4}$$

The gradient (4) can be computed by using the property of the convolution

$$\hat{J}_\beta(\theta) := \int_{\mathbb{R}^q} J(\theta-\eta)h_\beta(\eta)d\eta = \int_{\mathbb{R}^q} J(\eta)h_\beta(\theta-\eta)d\eta,$$

taking the gradient of the last term w.r.t θ, and using the derivative of the normal distribution. It also has the equivalent double-sided version

$$\nabla_\theta \hat{J}_\beta(\theta) = \frac{1}{2\beta}\mathbb{E}_\eta\{[J(\theta+\beta\eta) - J(\theta-\beta\eta)]\eta\}.$$

The double-sided form can be obtained by changing η to $-\eta$ in (4), adding two identical integrals, and dividing the results by two as discussed in [39].

An experimental evidence was given in [39] to demonstrate that the double-sided gradient form provides better performance for optimization problems. Computation of the expectation is numerically intractable when the dimension of η is large. For this reason, the stochastic gradient estimate can be used

$$\nabla_\theta \hat{J}_\beta(\theta) \cong \frac{1}{2\beta}\frac{1}{N_\eta}\sum_{i=1}^{N_\eta}[J(\theta+\beta\eta^{(i)}) - J(\theta-\beta\eta^{(i)})]\eta^{(i)},$$

where $\eta^{(i)}$ denotes the i-th sample of $\eta \sim \mathcal{N}(0, I_q)$ and $N_\eta \in \mathbb{N}_+$. Since J cannot be evaluated in our problem setting, we further approximate J by using the sample average

$$J(\theta) \cong \tilde{J}(\theta) := \frac{1}{N_J}\sum_{i=1}^{N_J}\sum_{k=0}^{N} c_k(x^{(i)}(k), u_k(x^{(i)}(k);\theta)),$$

where $N_J \in \mathbb{N}_+$, $(x^{(i)}(k))_{k=0}^N$ is the i-th realization of the process $(\mathbf{x}(k))_{k=0}^N$. Combining the two estimates, our gradient estimator is

$$\nabla_\theta \hat{J}_\beta(\theta) \cong g(\theta) := \frac{1}{2\beta}\frac{1}{N_\eta}\sum_{i=1}^{N_\eta}[\tilde{J}(\theta+\beta\eta^{(i)}) - \tilde{J}(\theta-\beta\eta^{(i)})]\eta^{(i)}. \tag{5}$$

Now, since the approximate gradient is available, a gradient descent type algorithm [47] can be used to solve Problem 3.

Algorithm 1 Stochastic Algorithm for β-ASOCP

1: Initialize $\theta_0 \in \mathbb{R}^q$, set $t = 0$ and a convex set $\mathcal{C} \subseteq \mathbb{R}^q$.
2: **repeat**
3: $\quad \theta_{t+1} = \Pi_{\mathcal{C}}(\theta_t - \gamma_t g(\theta_t))$
4: $\quad t \leftarrow t+1$
5: **until** t is sufficiently large.

Remark 1. *Algorithm 1 can be regarded as a class of policy gradient reinforcement learning [42–45]. Philosophically, it is most similar to the finite-difference policy gradient [43], where the simultaneous perturbation stochastic gradient approximation (SPSA) [48] is used to estimate the gradient. One of differences is that in SPSA, the cost function J needs to be differentiable while in Algorithm 1, it does not. Most likelihood-radio gradient estimators [42, 44] are applicable to systems with discrete state spaces and stochastic policies. For systems with continuous state spaces and deterministic policies, the approach is problematic [45]. Although its counterpart exist as in [45], they should be implemented in the actor-critic framework [29].*

Remark 2. *1) An advantage of the proposed approach is its simplicity and applicability to broad classes of systems. A disadvantage is that there exist many tuning issues, for instance, the choices of the kernels h_β and parameterization structures. 2) Algorithm 1 can be applied for deterministic systems. However, Algorithm 1 is especially useful with stochastic systems than deterministic systems because in the former case, dynamic programming algorithms can be more easily applied, e.g. solving Riccati equations for deterministic linear systems.*

Definition 1 (Lipschitz continuity, [49, 50])**.** *Function $f : I \to \mathbb{R}^s$ is Lipschitz continuous on I with constant $L_0(f) > 0$ if $\|f(x) - f(y)\| \leq L_0(f)\|x - y\|$, for all $x, y \in I$, where $L_0(f) \in \mathbb{R}_{++}$ is called the Lipschitz constant and $I \subseteq \mathbb{R}^q$. $f : I \to \mathbb{R}^s$ is called locally Lipschitz continuous on I if for every x in I, there exists a neighborhood $\mathcal{N}_x$ of x such that f is Lipschitz continuous on $\mathcal{N}_x$.*

Throughout the paper, for any Lipschitz continuous function f, its Lipschitz constant will be denoted by $L_0(f)$. Moreover, we assume that J is Lipschitz continuous on $\mathcal{C}$ with constant $L_0(J) > 0$.

Assumption 1. *J is Lipschitz continuous on $\mathcal{C}$ with constant $L_0(J) > 0$.*

If Assumption 1 holds with $\mathcal{C} = \mathbb{R}^q$, we can prove the convergence of Algorithm 1 to the stationary point θ^* of $\hat{J}_\beta$ which satisfies $\nabla_\theta \hat{J}_\beta(\theta^*) = 0$. Although the convergence is proved by using existing results, e.g., [29, Prop. 4.1], we provide its proof in Appendix 7 for completeness of the presentation.

Proposition 1. *Suppose that Assumption 1 holds with $\mathcal{C} = \mathbb{R}^q$, and*

$$\lim_{t\to\infty} \gamma_t = 0, \quad \sum_{t=0}^{\infty} \gamma_t = \infty, \quad \sum_{t=0}^{\infty} \gamma_t^2 < \infty. \tag{6}$$

Then, the following properties hold with probability one:

1. *The sequence $(\hat{J}_\beta(\theta_t))_{t=0}^{\infty}$ converges.*
2. *$\lim_{t\to\infty} \nabla \hat{J}_\beta(\theta_t) = 0$.*
3. *Every limit point of $(\theta_t)_{t=0}^{\infty}$ is a stationary point of $\hat{J}_\beta$.*

Proof. See Appendix 7. □

Although Algorithm 1 is easier to implement in the case $\mathcal{C} = \mathbb{R}^q$, Assumption 1 is difficult to hold if $\mathcal{C} = \mathbb{R}^q$. It will be explained in more detail in the next section with LTI systems. For this reason, we will also consider the case that $\mathcal{C}$ is a convex and compact subset of $\mathbb{R}^q$. In this case, Assumption 1 holds under mild conditions, while the convergence proof becomes more tricky. To prove the convergence, we adopt the following assumption on $\mathcal{C}$.

Assumption 2 ([51, Assumption 3]). *The set $\mathcal{C}$ is described by*

$$\mathcal{C} := \{\theta \in \mathbb{R}^q : h_j(\theta) \leq 0, j \in \{1, \ldots, p\}\},$$

which is nonempty and compact, where $p \in \mathbb{N}_+$, $h_j : \mathbb{R}^q \to \mathbb{R}$, $j \in \{1, \ldots, p\}$, are continuously differentiable in a neighborhood of $\partial\mathcal{C}$ (boundary of $\mathcal{C}$), for any $\theta \in \partial\mathcal{C}$, $\{\nabla h_j(\theta), j \in \mathcal{A}(\theta)\}$ is a linearly independent collection of vectors, and $\mathcal{A}(\theta)$ is the active set defined as $\mathcal{A}(\theta) = \{j : h_j(\theta) = 0, \theta \in \mathcal{C}\}$.

Define the set of stationary points of $\hat{J}_\beta$ on $\mathcal{C}$ as

$$\mathcal{L} := \{\xi \in \mathcal{C} : -\nabla_\theta \hat{J}(\xi) \in \mathcal{N}_C(\xi)\},$$

where $\mathcal{N}_\mathcal{C}(\xi)$ is the normal cone, i.e., $\mathcal{N}_\mathcal{C}(\xi) := \{v \in \mathbb{R}^q : v^T(\xi - \xi') \geq 0, \forall \xi' \in \mathcal{C}\}$. We can establish the convergence of Algorithm 1.

Proposition 2. *Suppose that Assumption 1 and Assumption 2 hold and $\hat{J}_\beta(\mathcal{L})$ has an empty interior. Let $(\theta_t)_{t=0}^\infty$ be the stochastic process generated by Algorithm 1 with the step-size rule in (6). Then, with probability one, $\lim_{k\to\infty} \operatorname{dist}(\theta_k, \mathcal{L}) = 0$, where $\operatorname{dist}(x, \mathcal{L}) := \inf_{y\in\mathcal{L}} \|x - y\|$.*

Proof. See Appendix 8. □

4 Analysis for LTI systems

From the discussions of the previous section, the Lipschitz continuity of J is essential. Consider the stochastic LTI system

$$x(k+1) = Ax(k) + Bu(k) + Dw(k), \quad x(0) = z \in \mathbb{R}^n, \tag{7}$$

where $k \in \{0, \ldots, N-1\}$, and $w(k)$ is a random vector whose probability mass function is dependent on the state $x(k)$. Under some conditions, we can prove that J corresponding to the LTI system in (7) is Lipschitz continuous on $\mathcal{C}$. In particular, the LTI system (7) can be described by the equation $\bar{x} = \bar{A}z + \bar{B}\bar{u}(\bar{x}, \theta) + \bar{D}\bar{w}$, where

$$\bar{x} := \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N) \end{bmatrix}, \quad \bar{u}(\bar{x}, \theta) := \begin{bmatrix} u_0(x(0); \theta) \\ u_1(x(1); \theta) \\ \vdots \\ u_{N-1}(x(N-1); \theta) \end{bmatrix},$$

and

$$\bar{w} := \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(N-1) \end{bmatrix}, \quad \bar{A} := \begin{bmatrix} I_n \\ A \\ \vdots \\ A^N \end{bmatrix}, \quad \bar{B} := \begin{bmatrix} 0 & \cdots & 0 \\ B & \ddots & \vdots \\ \vdots & \ddots & 0 \\ A^{N-1}B & \cdots & B \end{bmatrix}, \quad \bar{D} := \begin{bmatrix} 0 & \cdots & 0 \\ D & \ddots & \vdots \\ \vdots & \ddots & 0 \\ A^{N-1}D & \cdots & D \end{bmatrix}.$$

Summarizing, both $\bar{x}$ and $\bar{u}$ can be written as functions of θ and $\bar{w}$, i.e., $\bar{x}(\theta, \bar{w}) = \bar{A}z + \bar{B}\bar{u}(\bar{x}(\theta, \bar{w}), \theta) + \bar{D}\bar{w}$. Define

$$J(\theta, \bar{w}) := \sum_{k=0}^{N} c_k(x(k; \theta, \bar{w}), u_k(x(k; \theta, \bar{w}), \theta)),$$

where $x(k;\theta,\bar{w})$ is the state vector at time k for given θ and $\bar{w}$. Assume that $\bar{w} \in W^N$ is a discrete random variable with the probability mass function $p_{\bar{w}}(\bar{w};\bar{x}(\theta,\bar{w}))$. Then, J in (3) can be described by

$$J(\theta) = \sum_{\bar{w}\in D} J(\theta;\bar{w}) p_{\bar{w}}(\bar{w};\bar{x}(\theta,\bar{w})).$$

In this case, the probability mass function depends on θ. For notational simplicity, we write $p_{\bar{w}}(\bar{w};\bar{x}(\theta,\bar{w})) = p_{\bar{w}}(\bar{w};\theta)$. We can prove that if the stage cost function, probability mass function, and the control policy are Lipschitz, then J is also Lipschitz.

Proposition 3. *Consider the LTI system in (7). Suppose that*

1. *the stage cost function $c_k : X \times U \to \mathbb{R}_+$ is Lipschitz on $X \times U$ for all $k \in \{1,\dots,N\}$;*
2. *for any $\bar{w} \in W^N$, $p_{\bar{w}}(\bar{w};\cdot)$ is Lipschitz continuous on $\mathbb{R}^{n(N+1)}$;*
3. *for any $k \in \{1,\dots,N-1\}$, the parameterized control policy $u_k : X \times \mathcal{C} \to \mathbb{R}^m$ is Lipschitz continuous on $X \times \mathcal{C}$;*
4. *there exists $G > 0$ such that $J(\theta;\bar{w}) \le G, \forall \theta \in \mathcal{C}, \bar{w} \in W^N$.*

Then, J is Lipschitz continuous on $\mathcal{C}$.

Proof. See Appendix 9. □

Note that the third statement of Proposition 3 is difficult to be satisfied when $\mathcal{C} = \mathbb{R}^q$. For instance, if $\pi_k : \mathbb{R}^n \times \mathcal{C} \to \mathbb{R}^m$ in (2) is a linear state feedback, i.e., $\pi_k(x,\theta) = Fx$, where F is a state-feedback gain matrix and $\theta = vec(F)$ ($vec(F)$ is a vectorization of F), then π_k is a bilinear function, which is only locally Lipschitz. Although $u_k : \mathbb{R}^n \times \mathcal{C} \to \mathbb{R}^m$ in (2) is bounded due to the projection map, it is still not Lipschitz on $\mathcal{C} = \mathbb{R}^q$. If $\mathcal{C}$ is a compact subset of $\mathbb{R}^q$, then by using the local Lipschitz assumption and the compactness of $\mathcal{C}$, it is easy to prove that $\pi_k(x,\theta) = Fx$ is Lipschitz in $\mathcal{C}$. Therefore, the convergence can be guaranteed by Proposition 2.

5 Application: Building HVAC System Control with Occupant

In this paper, we consider a 3m × 3m private office space with a 2.5m^2 south facing window, and its RC (resistor-capacitor) circuit analogy is given in Figure 1. 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_a is the air temperature ($^\circ C$), T_o is the outdoor air temperature ($^\circ C$), T_w is the temperature of the aggregated mass node ($^\circ C$), q_solar is the solar radiation (W), q_internal is the internal heat (W), q_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_HVAC directly affects T_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 [52], 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 1. The dynamic system model is given as

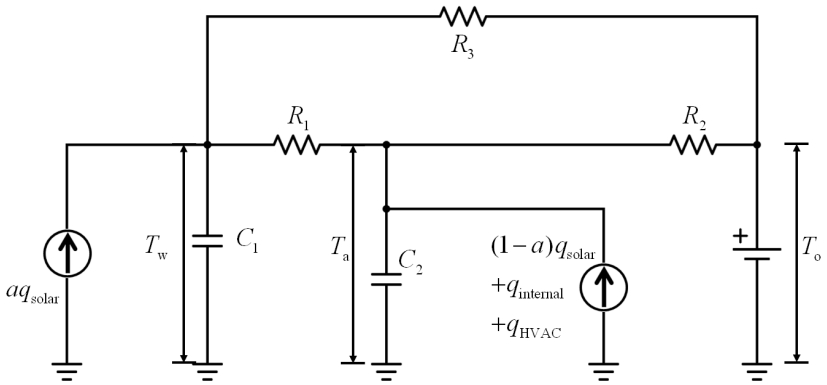

Figure 1: RC circuit analogy

Table 1: Values of the parameters of the circuit in Figure 1

Parameter	Value	Unit
R_1	0.0084197	$°C/W$
R_2	0.044014	$°C/W$
R_3	4.38	$°C/W$
C_1	9861100	$J/°C$
C_2	128560	$J/°C$
a	0.55	–

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

A discrete time representation can be obtained by using the Euler discretization with a sampling time of Δt

$$\begin{aligned} T_{\text{a}}(k+1) - T_{\text{a}}(k) =& \frac{\Delta t}{C_2R_2}(T_{\text{o}}(k) - T_{\text{a}}(k)) + \frac{\Delta t}{C_2R_1}(T_{\text{w}}(k) - T_{\text{a}}(k)) \\ &+ \frac{\Delta t(1-a)}{C_2}q_{\text{solar}}(k) + \frac{\Delta t}{C_2}q_{\text{HVAC}}(k) + \frac{\Delta t}{C_2}q_{\text{internal}}(k), \\ T_{\text{w}}(k+1) - T_{\text{w}}(k) =& \frac{\Delta t}{C_1R_1}(T_{\text{a}}(k) - T_{\text{w}}(k)) + \frac{\Delta t}{C_1R_3}(T_{\text{o}}(k) - T_{\text{w}}(k)) + \frac{\Delta ta}{C_1}q_{\text{solar}}(k), \end{aligned}$$

where $k \in \mathbb{N}$ is the discrete time step. In this paper, we consider $\Delta t = 10$min sampling time with 24 hours time horizon. 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. Therefore, the discrete time horizon is $N = 144$. Moreover, the real weather data $(T_{\text{o}}(k), q_{\text{solar}}(k))$ for $k \in \{1, \dots N\}$ collected during the day 30th, July, 2017, is used (see Figure 2).

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 $(\mathbf{z}(k))_{k=0}^{144}$ with the state space $S = \{1, 2, 3\}$, which represents the occupant's feeling of cold, comfort, and hot, respectively.

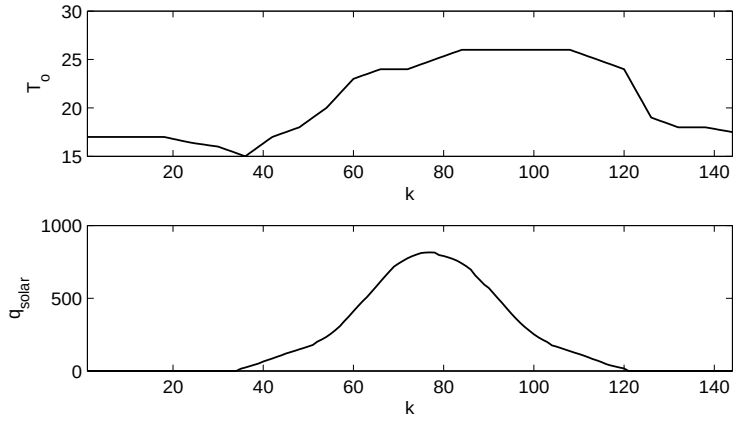

Figure 2: Weather data $(T_{\mathrm{o}}(k), q_{\mathrm{solar}}(k))$ for $k \in \{1, \ldots, N\}$.

Its probability depends on the current indoor temperature $T_{\mathrm{a}}(k)$, and its probability mass function $p_{\mathbf{z}}(z; T_{\mathrm{a}})$ is obtained by the Bayesian modelling approach in [9]. The values of the probability for different values of T_{a} are depicted in Figure 3. Consider some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let

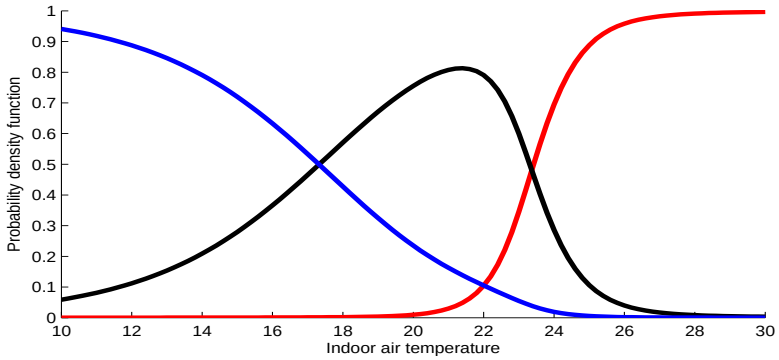

Figure 3: The probability mass function $p_{\mathbf{z}}(1; T_{\mathrm{a}})$ (blue), $p_{\mathbf{z}}(2; T_{\mathrm{a}})$ (black), $p_{\mathbf{z}}(3; T_{\mathrm{a}})$ (red) for different T_{a}

R be a space of occupant's actions, and let D be some information space. The information space D is a set of variables that affect occupant actions. For example, the values of $z(k)$ can be an element of D because it is used to induce occupant actions. The occupant's actions are modelled as a map $M : D \times \Omega \to R$. In this example, we consider two possible scenarios of occupant's actions, i.e., $M = \begin{bmatrix} M_1 & M_2 \end{bmatrix}^T$, described below. Arguments of M will be omitted for notational simplicity.

1. Occupancy (M_1): The occupant arrives at the room at time $\mathbf{w}_1$ uniformly distributed within $\{48, \ldots, 54\}$ (between 8am and 9am), and leaves the room at time $\mathbf{w}_2$ uniformly distributed within $\{96, \ldots, 114\}$ (between 4pm and 7pm). The map $M_1(k, \mathbf{w}_1, \mathbf{w}_2) \in \{0, 1\}$ is

$$M_1 = \begin{cases} 0, & \text{if } k < \mathbf{w}_1 \text{ or } k > \mathbf{w}_2 \\ 1, & \text{otherwise} \end{cases} .$$

2. Occupant's overriding on set point (M_2): The occupant uses a control panel to increase, decrease, or maintain the current temperature set point. The set point has the dynamic

equation $T_{\text{ref}}(k+1) = T_{\text{ref}}(k)+M_2$, where $T_{\text{ref}}(k)$ is the current set point, and M_2 is occupant's control input. If $\mathbf{z}(k) = 1$, then

$$M_2 = \begin{cases} 0, & \text{w.p.} \quad 0.4 \\ M_1, & \text{w.p.} \quad 0.3 \\ 2M_1, & \text{w.p.} \quad 0.2 \\ 3M_1, & \text{w.p.} \quad 0.1 \end{cases},$$

if $\mathbf{z}(k) = 2$, then $M_2 = 0$, and if $\mathbf{z}(k) = 3$, then

$$M_2 = \begin{cases} 0, & \text{w.p.} \quad 0.4 \\ -M_1, & \text{w.p.} \quad 0.3 \\ -2M_1, & \text{w.p.} \quad 0.2 \\ -3M_1, & \text{w.p.} \quad 0.1 \end{cases}.$$

The set point is assumed to vary within the range $15 \leq T_{\text{ref}}(k) \leq 30$.

Accordingly, the internal heat is given by $q_{\text{internal}}(k) = 75 + 70M_1$ (W), where the first term 75 is internal heat due to electronic products, and the second term $70M_1$ indicates the heat produced by the occupant's body. To construct a state-space model, two additional state variables are considered in addition to $T_{\text{a}}(k)$ and $T_{\text{w}}(k)$. The first additional state variable is $T_{\text{ref}}(k)$ described by $T_{\text{ref}}(k+1) = T_{\text{ref}}(k) + M_2$. The second one is the outdoor air temperature $T_{\text{o}}(k)$ described by $T_{\text{o}}(k+1) = T_{\text{o}}(k) + \Delta T_{\text{o}}(k)$, where $\Delta T_{\text{o}}(k) = T_{\text{o}}(k+1) - T_{\text{o}}(k)$. Even though T_{o} is a disturbance, by including it as a state variable, the state-space model better captures the real system dynamics, and improve the performance of the linear quadratic regulator (LQR) control policy, which will be used to initialize the proposed control design algorithm. In summary, one obtains a state-space model $x(k+1) = Ax(k) + Bu(k) + Dw(k)$ with $u(k) = q_{\text{HVAC}}(k)$,

$$x(k) = \begin{bmatrix} T_{\text{a}}(k) \\ T_{\text{w}}(k) \\ T_{\text{ref}}(k) \\ T_{\text{o}}(k) \end{bmatrix}, \quad w(k) = \begin{bmatrix} q_{\text{solar}}(k) \\ 75 + 70M_1 \\ M_2 \\ \Delta T_{\text{o}}(k) \end{bmatrix},$$

and

$$A = \begin{bmatrix} 1 - \frac{\Delta t}{C_2R_2} - \frac{\Delta t}{C_2R_1} & \frac{\Delta t}{C_2R_1} & 0 & \frac{\Delta t}{C_2R_2} \\ \frac{\Delta t}{C_1R_1} & 1 - \frac{\Delta t}{C_1R_1} - \frac{\Delta t}{C_1R_3} & 0 & \frac{\Delta t}{C_1R_3} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} \frac{\Delta t}{C_2} \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} \frac{\Delta t(1-a)}{C_2} & \frac{\Delta t}{C_2} & 0 & 0 \\ \frac{\Delta ta}{C_1} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

The stage cost function at time k is defined as $c_k(x(k), u(k)) = x(k)^T Q x(k) + u(k)^T R u(k)$ if $\mathbf{w}_1 \leq k \leq \mathbf{w}_2$ and $c_k(x(k), u(k)) = u(k)^T R u(k)$ otherwise, where $R = 0.00001$ and

$$Q = \begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix}^T \begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix}.$$

Note that $x(k)^T Q x(k) = (T_{\text{a}}(k) - T_{\text{ref}}(k))^2$ is the square of the set point tracking error, and $u(k)^T R u(k)$ represents the control input energy. R is a weight to balance between the tracking performance and the energy saving. Its value was chosen by experiments.

5.1 Output-feedback control with 24 hours time-horizon

In this subsection, we present an output-feedback control structure. Let $F_{\mathrm{LQR}}(k), k \in \{0, \ldots, N-1\}$, be the finite-horizon LQR state-feedback gain for (A, B). Assume that the control input q_{HVAC} is saturated when $q_{\mathrm{HVAC}} > 1000W$ or $q_{\mathrm{HVAC}} < -1000W$. We use a parameterized control policy (2) of the following form:

$$u_k(x(k); \theta_c) = \Pi_U(F_{\mathrm{LQR}}(k)x(k) + \pi_k(x(k); \theta_c)), \tag{8}$$

where Π_U is the projection onto a convex set U, and $U = [-1000, 1000]$, $\pi_k(x(k); \theta_c)$ is an additive control input to be determined, which compensates the LQR control policy to improve the performance. In addition, consider the parameterization of $\pi_k(x(k); \theta_c)$

$$\begin{aligned}\pi_k(x(k); \theta_c) = {} & F_1 x(k) + \phi(k-40)F_2 x(k) + \phi(k-80)F_3 x(k) \\ & + \phi(k-120)F_4 x(k) + \phi(x_1(k)-20)F_5 x(k) \\ & + \phi(x_1(k)-25)F_6 x(k) + \phi(x_1(k)-30)F_7 x(k) \\ & + \phi(x_3(k)-20)F_8 x(k) + \phi(x_3(k)-25)F_9 x(k) \\ & + \phi(x_3(k)-30)F_{10} x(k),\end{aligned} \tag{9}$$

where ϕ is the Gaussian radial basis function $\phi(t) := \exp(-t^2/(2\sigma^2))$, $\sigma = 10$, and θ_c is any vectorization of $\{F_1, \cdots, F_{10}\}$. To meet practical needs, we will apply a output-feedback control scheme. In particular, the wall temperature T_{w} cannot be exactly measured in building control applications. The measurable output vector $y(k) \in \mathbb{R}^3$ is $y(k) = Cx(k)$, where $C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Since (A, C) is observable, we can consider the Kalman filter to estimate the current state

$$\hat{x}(k+1) = A\hat{x}(k) + L_{\mathrm{Kal}}(k)(C\hat{x}(k) - y(k))$$

for $k \in \{0, \ldots, N-1\}$, where $L_{\mathrm{Kal}}(k), k \in \{0, \ldots, N-1\}$, are the Kalman filter gains for (A, C) and $\hat{x}(k), k \in \{0, \ldots, N\}$, are the estimated states. An output-feedback controller is $u(k) = F_{\mathrm{LQR}}(k)\hat{x}(k), k \in \{0, \ldots, N-1\}$. We consider a modified Kalman filter of the form

$$\hat{x}(k+1) = A\hat{x}(k) + L_{\mathrm{Kal}}(k)(C\hat{x}(k) - y(k)) + \omega_k(e(k); \theta_o)$$

for $k \in \{0, \ldots, N-1\}$, where $e(k) := C\hat{x}(k) - y(k)$, $\omega_k(e(k); \theta_o)$ is an additive term that compensates the original Kalman filter, which is parameterized as follows:

$$\begin{aligned}\omega_k(e(k); \theta_o) = {} & L_1 e(k) + \phi(k-40)L_2 e(k) + \phi(k-80)L_3 e(k) \\ & + \phi(k-120)L_4 e(k) + \phi(e_1(k)-20)L_5 e(k) \\ & + \phi(e_1(k)-25)L_6 e(k) + \phi(e_1(k)-30)L_7 e(k) \\ & + \phi(e_2(k)-20)L_8 e(k) + \phi(e_2(k)-25)L_9 e(k) \\ & + \phi(e_2(k)-30)L_{10} e(k),\end{aligned}$$

and θ_o is any vectorization of the matrices $\{L_1, \ldots, L_{10}\}$. With the state estimate, the control policy $u_k(\hat{x}(k); \theta_c)$ in (9) is used. To guarantee the convergence, the convex set $\mathcal{C}$ is set to be $\mathcal{C} = [-100, 100]^q$. We can prove that J is Lipschitz continuous on $\mathcal{C}$.

Proposition 4. *Assume that the probability mass function $p_{\mathbf{z}}(z; \cdot)$ is Lipschitz continuous on $\mathbb{R}$. Then, J is Lipschitz continuous on $\mathcal{C}$.*

Proof. Since the building thermal dynamics is modelled as a stochastic LTI system, we will apply Proposition 3. In particular, the stage cost function c_k is quadratic (for any $\mathbf{w_1}$ and $\mathbf{w_2}$), which is continuously differentiable. Therefore, it is Lipschitz on any compact subset by [50, Corollary 6.4.20]. Due to the projection in (8), we have $U = [-1000, 1000]$. Therefore, c_k is Lipschitz continuous on $X \times U$ by [49, Corollary 12.2]. Thus, 1) of Proposition 3 is satisfied. The statement 2) Proposition 3 is satisfied by hypothesis. To prove 3), note that (9) is continuously differentiable w.r.t. (x, θ). By [50, Corollary 6.4.20], it is Lipschitz continuous on every compact set on $X \times \mathcal{C}$. On the other hand, it can be proved that the projection $\Pi_{\mathcal{C}}$ is Lipschitz continuous on $\mathcal{C}$ because it is a non-expansive map. By [49, Theorem 12.6], the composition function (8) is Lipschitz on $\mathcal{C}$. Therefore, $u_k(x; \theta)$ in (8) with (9) is Lipschitz continuous w.r.t. (x, θ) on $X \times \mathcal{C}$. Thus, 3) of Proposition 3 is satisfied. Finally, since the disturbance and control input are bounded for all $\theta \in \mathcal{C}$, and the sum in (3) is finite, $J(\theta)$ is also bounded. Thus, 4) of Proposition 3 is satisfied. The proof is completed by applying Proposition 3. □

Recall the set of stationary points of $\hat{J}_\beta$ on $\mathcal{C}$ defined as

$$\mathcal{L} := \{\xi \in \mathcal{C} : -\nabla_\theta \hat{J}(\xi) \in \mathcal{N}_C(\xi)\}.$$

Below, we establish the convergence of Algorithm 1 for the system under our consideration.

Proposition 5. *Assume that the probability mass function $p_{\mathbf{z}}(z; \cdot)$ is Lipschitz continuous on $\mathbb{R}$ and $\hat{J}_\beta(\mathcal{L})$ has an empty interior. Moreover, let $(\theta_t)_{t=0}^{\infty}$ Be the stochastic process generated by Algorithm 1 with the step-size rule in (6). Then, with probability one, $\lim_{k\to\infty} \operatorname{dist}(\theta_k, \mathcal{L}) = 0$, where $\operatorname{dist}(x, \mathcal{L}) := \inf_{y\in\mathcal{L}} \|x - y\|$.*

Proof. We will check that all the conditions in Proposition 2 are satisfied. By Proposition 4, J is Lipschitz continuous on $\mathcal{C}$. In addition, $\mathcal{C}$ is described by $\mathcal{C} = \{\theta \in \mathbb{R}^q : e_j^T\theta - 100 \le 0, -e_j^T - 100\theta \le 0, j \in \{1, \dots, q\}\}$, where $e_j \in \mathbb{R}^q$ is the vector whose elements are zeros except for its j-th element which is one. The left-hand side of each inequality consisting of $\mathcal{C}$ is affine, and all the conditions in Assumption 2 are satisfied. Therefore, the required conditions in Proposition 1 hold, and the proof is completed. □

We applied Algorithm 1 with $\theta = \begin{bmatrix}\theta_c^T & \theta_o^T\end{bmatrix}^T$, $\beta = 0.001$, $N_J = 20$, $N_\eta = 1$, 10^4 iterations, and the initial state $x(0) = \begin{bmatrix}25 & 25 & 25 & 17\end{bmatrix}^T$. Simulation results of the proposed control policy and the LQR policy are given in Figure 4 and Figure 5, respectively. Histograms of the costs of the two methods are compared in Figure 6 with a total of 1500 simulations. The average cost of the proposed control is 508.9, while 1361.7 for the LQR control. The result suggests that Algorithm 1 can potentially improve existing approaches.

Conclusion

In this paper, we studied an approximate policy gradient RL for stochastic optimal control design algorithm. It can be applied to complicated stochastic systems that cannot be easily solved by existing methods. Through simulation studies of building control with occupant interactions, we demonstrated its applicability and performance. Potential future research agendas are summarized as follows. 1) Establishing bounds on the loss of accuracy/suboptimality because of the convolution smoothing approach can be established by using [40, Theorem 1]. Comprehensive analysis is an

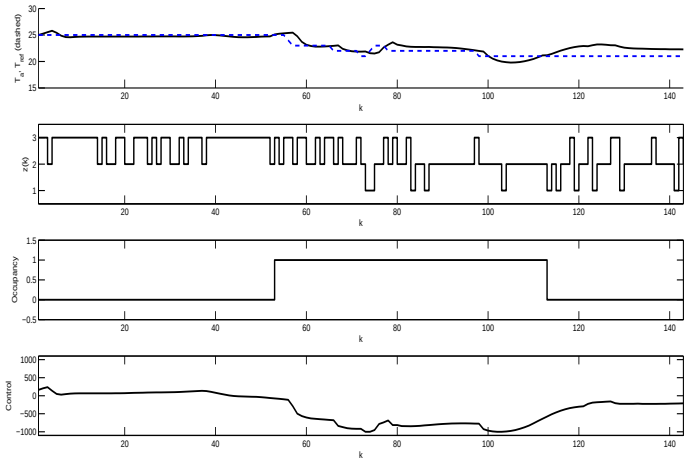

Figure 4: Simulation result with proposed output-feedback control.

avenue for future work. 2) The approach can scale to multi-zone building spaces with occupants movement between zones. In this case, statistical models that can simulate the movement of occupants are required. For computationally scalable algorithms, distributed optimization techniques, e.g., [53], can be considered. 3) An extension of the proposed method to constrained optimizations is useful in building control applications, where one may want to control the zone temperature to lie within an interval. For constrained stochastic optimizations, the barrier or augmented Lagrangian method [47] can be readily applied, while further attentions need to be payed for rigorous convergence analysis.

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] 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.

[3] 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.

[4] F. Oldewurtel, D. Sturzenegger, and M. Morari, "Importance of occupancy information for building climate control," *Applied energy*, vol. 101, pp. 521–532, 2013.

[5] J. R. Dobbs and B. M. Hencey, "Model predictive HVAC control with online occupancy model," *Energy and Buildings*, vol. 82, pp. 675–684, 2014.

[6] S. A. Sadeghi, P. Karava, I. Konstantzos, and A. Tzempelikos, "Occupant interactions with shading and lighting systems using different control interfaces: a pilot field study," *Building and Environment*, vol. 97, pp. 177–195, 2016.

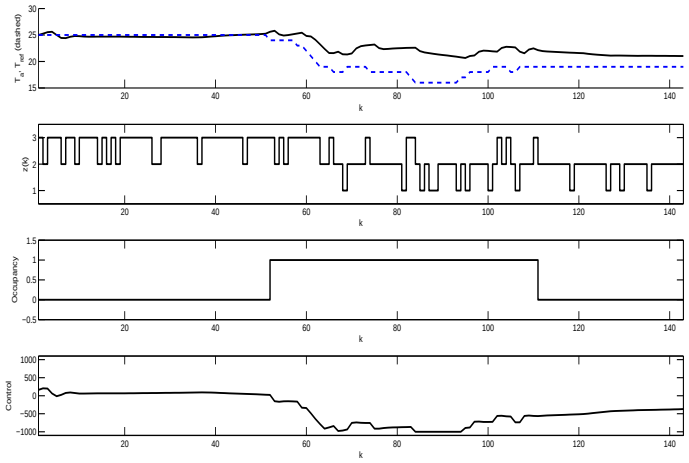

Figure 5: Simulation result with LQR control and Kalman filter.

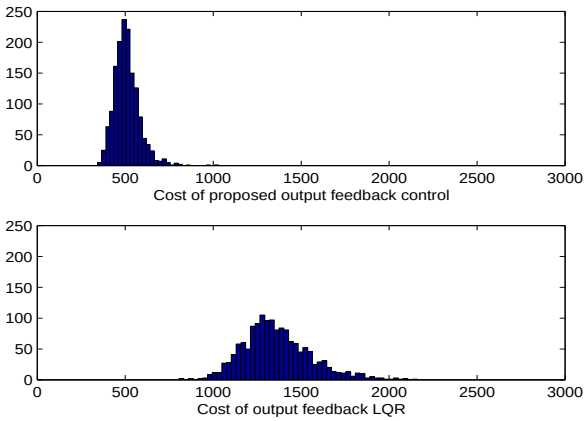

Figure 6: Cost histograms of LQR and proposed output-feedback control.

[7] W. Liu, Z. Lian, and B. Zhao, "A neural network evaluation model for individual thermal comfort," *Energy and Buildings*, vol. 39, no. 10, pp. 1115–1122, 2007.

[8] D. Daum, F. Haldi, and N. Morel, "A personalized measure of thermal comfort for building controls," *Building and Environment*, vol. 46, no. 1, pp. 3–11, 2011.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] P. Patrinos, P. Sopasakis, H. Sarimveis, and A. Bemporad, "Stochastic model predictive control for constrained discrete-time Markovian switching systems," *Automatica*, vol. 50, no. 10, pp. 2504–2514, 2014.

[18] O. L. V. Costa, E. Assumpção Filho, E. K. Boukas, and R. Marques, "Constrained quadratic state feedback control of discrete-time Markovian jump linear systems," *Automatica*, vol. 35, no. 4, pp. 617–626, 1999.

[19] M. Bichi, G. Ripaccioli, S. Di Cairano, D. Bernardini, A. Bemporad, and I. V. Kolmanovsky, "Stochastic model predictive control with driver behavior learning for improved powertrain control," in *49th IEEE Conference on Decision and Control (CDC)*, 2010, pp. 6077–6082.

[20] S. Di Cairano, D. Bernardini, A. Bemporad, and I. V. Kolmanovsky, "Stochastic MPC with learning for driver-predictive vehicle control and its application to HEV energy management," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 3, pp. 1018–1031, 2014.

[21] 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.

[22] A. E.-D. Mady, G. M. Provan, C. Ryan, and K. N. Brown, "Stochastic model predictive controller for the integration of building use and temperature regulation." in *AAAI*, 2011.

[23] B. Dong, K. P. Lam, and C. Neuman, "Integrated building control based on occupant behavior pattern detection and local weather forecasting," in *Twelfth International IBPSA Conference. Sydney: IBPSA Australia*, 2011, pp. 14–17.

[24] M. Prandini, J. Hu, J. Lygeros, and S. Sastry, "A probabilistic approach to aircraft conflict detection," *IEEE Transactions on intelligent transportation systems*, vol. 1, no. 4, pp. 199–220, 2000.

[25] 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.

[26] 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.

[27] 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.

[28] A. Parisio, M. Molinari, D. Varagnolo, and K. H. Johansson, "A scenario-based predictive control approach to building HVAC management systems," in *Automation Science and Engineering (CASE), 2013 IEEE International Conference on*, 2013, pp. 428–435.

[29] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific Belmont, MA, 1996.

[30] 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.

[31] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT Press, 1998.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] J. Kreimer and R. Y. Rubinstein, "Nondifferentiable optimization via smooth approximation: General analytical approach," *Annals of Operations Research*, vol. 39, no. 1, pp. 97–119, 1992.

[38] M. Styblinski and A. Ruszczynski, "Stochastic approximation approach to statistical circuit design," *Electronics Letters*, vol. 19, no. 8, pp. 300–302, 1983.

[39] M. Styblinski and T.-S. Tang, "Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing," *Neural Networks*, vol. 3, no. 4, pp. 467–483, 1990.

[40] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[41] F. Yousefian, A. Nedić, and U. V. Shanbhag, "On stochastic gradient and subgradient methods with adaptive steplength sequences," *Automatica*, vol. 48, no. 1, pp. 56–67, 2012.

[42] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[43] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 3, 2004, pp. 2619–2624.

[44] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural networks*, vol. 21, no. 4, pp. 682–697, 2008.

[45] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *ICML*, 2014.

[46] D. Lee, S. Lee, P. Karava, and J. Hu, "Simulation-based policy gradient and its building control application," in *American control conference (ACC2018) (in press)*, 2018.

[47] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[48] P. Sadegh and J. C. Spall, "Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 43, no. 10, pp. 1480–1484, 1998.

[49] K. Eriksson, D. Estep, and C. Johnson, *Applied mathematics: Body and soul: Volume 1: Derivatives and geometry in IR3*. Springer Science & Business Media, 2013.

[50] H. H. Sohrab, *Basic real analysis*. Springer, 2003, vol. 231.

[51] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 391–405, 2013.

[52] U. D. of Energy, "Energyplustm 8.7.0 documentation," https://energyplus.net/documentation, [Online Available].

[53] J. Cai, D. Kim, R. Jaramillo, J. E. Braun, and J. Hu, "A general multi-agent control approach for building energy system optimization," *Energy and Buildings*, vol. 127, pp. 337–351, 2016.

[54] S. Bhatnagar, H. L. Prasad, and L. A. Prashanth, *Stochastic recursive algorithms for optimization: simultaneous perturbation methods*. Springer, 2012, vol. 434.

Appendices

6 Useful Lemmas

In this section, some important results for the proofs of this paper are presented.

Lemma 1 ([40]). *Under Assumption 1, the following properties hold:*

1. *$\hat{J}_\beta$ is Lipschitz continuous on $\mathcal{C}$ with constant $L_0(\hat{J}_\beta) > 0$ with $L_0(\hat{J}_\beta) \leq L_0(J)$.*
2. *$|J(\theta) - \hat{J}_\beta(\theta)| \leq \beta L_0(J) q^{1/2}$, $\forall \theta \in \mathcal{C}$.*
3. *$\nabla_\theta \hat{J}_\beta$ is Lipschitz continuous on $\mathcal{C}$ with constant $L_1(\hat{J}_\beta) = (2q^{1/2}/\beta)L_0(J)$.*

Proof. The proof of 1) follows from [40, pp. 533], proof of 2) is given in [40, Theorem 1], and 3) is proved in [40, Lemma 1]. □

Another important property is that $\mathbb{E}\{\|g(x)\|^2\}$ is bounded, where $g(x)$ is the following gradient estimate defined in (5):

$$\nabla_\theta \hat{J}_\beta(\theta) \cong g(\theta) := \frac{1}{2\beta}\frac{1}{N_\eta}\sum_{i=1}^{N_\eta}[\tilde{J}(\theta + \beta\eta^{(i)}) - \tilde{J}(\theta - \beta\eta^{(i)})]\eta^{(i)}.$$

The boundedness can be proved by using the following fact: since U, X, and W are bounded, there exists a real number $G > 0$ such that $\tilde{J}(\theta) \leq G$ for all $\theta \in \mathcal{C}$.

Lemma 2. *There exists a real number $G > 0$ such that $\tilde{J}(\theta) \leq G$ for all $\theta \in \mathcal{C}$.*

Using the above lemma, we can prove that $\mathbb{E}\{\|g(\theta)\|^2\}$ is bounded for all $\theta \in \mathcal{C}$.

Lemma 3. *We have $\mathbb{E}\{\|g(\theta)\|^2\} \leq \frac{1}{N_\eta} G^2 q$ for all $\theta \in \mathcal{C}$.*

Proof. We have

$$\begin{aligned}
&\mathbb{E}\{\|g(x)\|^2\} = \mathbb{E}\{g(x)^T g(x)\} \\
&= \left(\frac{1}{2\beta}\frac{1}{N_\eta}\right)^2 \mathbb{E}\left\{\sum_{i=1}^{N_\eta}\sum_{j=1}^{N_\eta}(\tilde{J}(\theta + \beta\eta^{(i)}) - \tilde{J}(\theta - \beta\eta^{(i)}))(\tilde{J}(\theta + \beta\eta^{(j)}) - \tilde{J}(\theta - \beta\eta^{(j)}))(\eta^{(i)})^T\eta^{(j)}\right\} \\
&\leq \left(\frac{1}{2\beta}\frac{1}{N_\eta}\right)^2 \mathbb{E}\left\{\sum_{i=1}^{N_\eta}\sum_{j=1}^{N_\eta} 4G^2(\eta^{(i)})^T\eta^{(j)}\right\} \\
&= \left(\frac{1}{2\beta}\frac{1}{N_\eta}\right)^2 \mathbb{E}\left\{\sum_{i=1}^{N_\eta} 4G^2(\eta^{(i)})^T\eta^{(i)}\right\} \\
&= \left(\frac{1}{2\beta}\frac{1}{N_\eta}\right)^2 4G^2 \sum_{i=1}^{N_\eta}\mathbb{E}\{\mathrm{tr}(\eta^{(i)}(\eta^{(i)})^T)\}
\end{aligned}$$

$$= \left(\frac{1}{2\beta}\frac{1}{N_\eta}\right)^2 4G^2\beta^2 q N_\eta$$
$$= \frac{1}{N_\eta} G^2 q,$$

where in the first inequality, we used Lemma 2 and in the fourth equation, we used $\mathbb{E}\{\mathrm{tr}(\eta^{(i)}(\eta^{(i)})^T)\} = \mathrm{tr}(\beta^2 I_q)$. □

7 Proof of Proposition 1

To prove Proposition 1, we use results in [29, Prop. 4.1]. Consider an algorithm of the form $\theta_{t+1} = \theta_t + \gamma_t s_t$, and denote the history of the algorithm until time t by $\mathcal{I}_t := \{\theta_0, \ldots, \theta_t, \gamma_0, \ldots, \gamma_t, s_0, \ldots, s_{t-1}\}$. For convergence of this algorithm, some assumptions are needed.

Assumption 3 ([29, Assumption 4.2]). *We assume that there exists a function $f : \mathbb{R}^q \to \mathbb{R}$ with the following properties:*

1. *$f(x) \geq 0$ for all $x \in \mathbb{R}^q$.*
2. *f is differentialble, and its gradient ∇f is Lipschitz continuous with constant $L > 0$.*
3. *There exists a positive constant $c > 0$ such that $c\|\nabla f(\theta_t)\|^2 \leq -\nabla f(\theta_t)^T \mathbb{E}\{s_t|\mathcal{I}_t\}$ for all t.*
4. *There exist positive constants $K_1 > 0$ and $K_2 > 0$ such that $\mathbb{E}\{\|s_t\|^2|\mathcal{I}_t\} \leq K_1 + K_2\|\nabla f(\theta_t)\|^2$ for all t.*

Recall the step-size rule in (6)

$$\lim_{t\to\infty} \gamma_t = 0, \quad \sum_{t=0}^{\infty} \gamma_t = \infty, \quad \sum_{t=0}^{\infty} \gamma_t^2 < \infty.$$

Under Assumption 3 and the above step-size rule, convergence of θ_t to a stationary point of f is guaranteed with probability one.

Lemma 4 ([29, Prop. 4.1]). *Consider the algorithm $\theta_{t+1} = \theta_t + \gamma_t s_t$, where $(\gamma_t)_{t=0}^{\infty}$ are nonnegative and satisfy (6). Under Assumption 3, the following properties hold with probability one:*

1. *The sequence $(f(\theta_t))_{t=0}^{\infty}$ converges.*
2. *$\lim_{t\to\infty} \nabla f(\theta_t) = 0$.*
3. *Every limit point of $(\theta_t)_{t=0}^{\infty}$ is a stationary point of f.*

Now, one can easily prove that $g(\theta_t)$ in (5) satisfies Assumption 3 for $\hat{J}_\beta$. Therefore, by Lemma 4, $(\theta_t)_{t=0}^{\infty}$ converges to its stationary point θ^* with probability one.

Proof of Proposition 1. All we need to do is to prove that Assumption 3 is satisfied. Since the stage cost function satisfies $c_k \geq 0$ by assumption, $J \geq 0$ and $\hat{J}_\beta \geq 0$. Thus, the first statement of Assumption 3 holds. The second and forth statements are proved by 3) of Lemma 1 and Lemma 3, respectively, with $s_t = -g(\theta_t)$. Moreover, since $\mathbb{E}\{-g(\theta_t)|\mathcal{I}_t\} = -\nabla\hat{J}_\beta(\theta_t)$, the third statement holds with $C = 1$. The proof is completed by Lemma 4. □

8 Proof of Proposition 2

To prove Proposition 2, we can use the Kushner-Clark theorem for convergence of projected stochastic approximation in [54] or follow the existing work [51] for multi-agent projected stochastic gradient algorithm. In this section, we will apply single-agent version of [51, Theorem 1]. Consider the optimization problem $\min_{\theta\in\mathcal{C}} \hat{J}_\beta(\theta)$, where $\hat{J}_\beta : \mathbb{R}^q \to \mathbb{R}$ is continuously differentiable and $\mathcal{C}$ is a nonempty compact convex subset of $\mathbb{R}^q$. In addition, let $(\theta_t)_{t=0}^\infty$ be a stochastic process generated by $\theta_{t+1} = \Pi_{\mathcal{C}}(\theta_t - \gamma_t Y_t)$, where $(Y_t)_{t=0}^\infty$ is another process and $(\gamma_t)_{t=0}^\infty$ is a deterministic step-size rule. Suppose that $(Y_t)_{t=0}^\infty$ is defined on a measurable space equipped with a probability $\mathbb{P}$. We introduce the σ-filed $\mathcal{F}_t := \sigma(\theta_0, Y_0, \ldots, Y_t)$. Recall the set of stationary points of $\hat{J}_\beta$ on $\mathcal{C}$ defined as

$$\mathcal{L} := \{\xi \in \mathcal{C} : -\nabla_\theta \hat{J}(\xi) \in \mathcal{N}_C(\xi)\}.$$

The convergence result in [51, Theorem 1] is summarized in the following lemma.

Lemma 5 ([51, Theorem 1]). *Suppose that Assumption 2 holds, $\hat{J}_\beta(\mathcal{L})$ has an empty interior, and $(\gamma_t)_{t=0}^\infty$ are nonnegative and satisfy (6). In addition, assume that $\mathbb{E}\{Y_t|\mathcal{F}_{t-1}\} = \nabla \hat{J}_\beta(\theta_t), t \in \{1, \ldots\}$, and $\sup_{\theta\in\mathcal{C}} \int \|\mathbf{y}\|^2 \, d\mu_\theta(\mathbf{y}) < \infty$, where $\mu_{\theta_{n-1}}(S) := \mathbb{P}(Y_t \in S|\mathcal{F}_{t-1})$ for any measurable set S and $(\mu_\theta)_{\theta\in\mathbb{R}^q}$ is a given family of probability measures on $\mathbb{R}^q$. Then, with probability one, $\lim_{t\to\infty} \operatorname{dist}(\theta_t, \mathcal{L}) = 0$, where $\operatorname{dist}(x, \mathcal{L}) := \inf_{y\in\mathcal{L}} \|x - y\|$.*

Sketch of the proof of Proposition 2. We first replace Y_t with $g(\theta_t)$. By 3) of Lemma 1, $\hat{J}_\beta$ is continuously differentiable. Since $Y_t = g(\theta_t)$ and the random variable $g(\theta)$ only depends on $\theta \in \mathbb{R}^q$, there exists a family of probability measures $(\mu_\theta)_{\theta\in\mathbb{R}^q}$ such that $\mu_{\theta_{n-1}}(S) := \mathbb{P}(Y_t \in S|\mathcal{F}_{t-1})$. To prove $\sup_{\theta\in\mathcal{C}} \int \|\mathbf{y}\|^2 \, d\mu_\theta(\mathbf{y}) < \infty$, set $\mathbf{y} = g(\theta)$. Then, it is equivalent to $\sup_{\theta\in\mathcal{C}} \mathbb{E}\{g(\theta)\} < \infty$, which is proved in Lemma 3. This completes the proof. □

9 Proof of Proposition 3

To prove Proposition 3, we need an intermediate result which states that the state vector at time k is Lipschitz with respect to θ on $\mathcal{C}$.

Lemma 6. *Suppose that the assumptions in Proposition 3 hold. For any fixed $\bar{w} \in W^N$ and $k \in \{1, \ldots, N\}$, $x(k; \cdot, \bar{w})$ is Lipschitz continuous on $\mathcal{C}$.*

Proof. The proof is completed by induction. First, we prove that $x(1; \cdot, \bar{w})$ is Lipschitz continuous on $\mathbb{R}^q$. For any given $x(0) = z$, we have $x(1; \cdot, \bar{w}) = Az + Bu_0(z, \cdot) + Dw(0)$, which is an affine transformation of the Lipschitz continuous function $u_0(z, \cdot)$. By [49, Theorem 12.6], the composition of two Lipschitz functions is Lipschitz. Therefore, $x(1; \cdot, \bar{w})$ is Lipschitz on $\mathbb{R}^q$. Assume that $x(k-1; \cdot, \bar{w})$ is Lipschitz on $\mathbb{R}^q$. Then, $x(k; \cdot, \bar{w})$ is described by $x(k; \cdot, \bar{w}) = Ax(k-1; \cdot, \bar{w}) + Bu_{k-1}(x(k-1; \cdot, \bar{w}), \cdot) + Dw(k-1)$, which is an affine function of $x(k-1; \cdot, \bar{w})$ and $u_{k-1}(x(k-1; \cdot, \bar{w}), \cdot)$. By assumption, $x(k-1; \cdot, \bar{w})$ and $u_{k-1}(\cdot, \cdot)$ are Lipschitz. By composition of functions, $x(k; \cdot, \bar{w})$ is also Lipschitz continuous. In particular, we have

$$\begin{aligned}
&\|x(k; \theta, \bar{w}) - x(k; \theta', \bar{w})\| \\
&= \|Ax(k-1; \theta, \bar{w}) + Bu_{k-1}(x(k-1; \theta, \bar{w}), \theta) - Ax(k-1; \theta', \bar{w}) - Bu_{k-1}(x(k-1; \theta', \bar{w}), \theta')\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \|A\|\|x(k-1;\theta,\bar{w}) - x(k-1;\theta',\bar{w})\| \\
&+ \|B\|\|u_{k-1}(x(k-1;\theta,\bar{w}),\theta) - u_{k-1}(x(k-1;\theta',\bar{w}),\theta')\| \\
&\leq \|A\|L(x(k-1;\cdot,\bar{w}))\|\theta-\theta'\| \\
&+ \|B\|L(u_{k-1}(\cdot,\cdot))\left\|\begin{bmatrix} x(k-1;\theta,\bar{w}) \\ \theta \end{bmatrix} - \begin{bmatrix} x(k-1;\theta',\bar{w}) \\ \theta' \end{bmatrix}\right\| \\
&\leq \|A\|L(x(k-1;\cdot,\bar{w}))\|\theta-\theta'\| \\
&+ \|B\|\, L(u_{k-1}(\cdot,\cdot))(\left\|x(k-1;\theta,\bar{w}) - x(k-1;\theta',\bar{w})\right\| + \left\|\theta-\theta'\right\|) \\
&\leq \underbrace{[\|A\|L(x(k-1;\cdot,\bar{w})) + \|B\|L(u_{k-1}(\cdot,\cdot)) + \|B\|L(u_{k-1}(\cdot,\cdot))L(x(k-1;\cdot,\bar{w}))]}_{L(x(k;\cdot,\bar{w}))}\|\theta-\theta'\|
\end{aligned}
$$

By the induction argument, $x(k;\cdot,\bar{w})$ is Lipschitz in $\mathcal{C}$ for all $k \in \{1,\dots,N\}$. This completes the proof. □

As a next step, we prove that J is also Lipschitz continuous on $\mathcal{C}$.

Lemma 7. *For any fixed $\bar{w} \in W^N$, $J(\theta,\bar{w})$ defined as*

$$
J(\theta,\bar{w}) := \sum_{k=0}^{N} c_k(x(k;\theta,\bar{w}), u_k(x(k;\theta,\bar{w}),\theta)),
$$

is Lipschitz continuous w.r.t θ on $\mathcal{C}$.

Proof. We have

$$
\begin{aligned}
|J(\theta,\bar{w}) - J(\theta',\bar{w})| &\leq \sum_{k=0}^{N} \left| \begin{pmatrix} c_k(x(k;\theta,\bar{w}), u_k(x(k;\theta,\bar{w}),\theta)) \\ -c_k(x(k;\theta',\bar{w}), u_k(x(k;\theta',\bar{w}),\theta')) \end{pmatrix} \right| \\
&\leq \sum_{k=0}^{N} L_0(c_k)\|x(k;\theta,\bar{w}) - x(k;\theta',\bar{w})\| \\
&+ \sum_{k=0}^{N} L_0(c_k)\|u_k(x(k;\theta,\bar{w}),\theta) - u_k(x(k;\theta',\bar{w}),\theta')\| \\
&\leq \sum_{k=0}^{N} L_0(c_k)\|x(k;\theta,\bar{w}) - x(k;\theta',\bar{w})\| \\
&+ \sum_{k=0}^{N} L_0(c_k)L_0(u_k)\|x(k;\theta,\bar{w}) - x(k;\theta',\bar{w})\| + \sum_{k=0}^{N} L_0(c_k)L_0(u_k)\|\theta-\theta'\|,
\end{aligned}
$$

where the second inequality follows from the Lipschitz continuity of c_k, and the last inequality from the Lipschitz continuity of u_k. The proof is completed by using Lemma 6. □

Proof of Proposition 3. We have

$$
\begin{aligned}
&|J(\theta) - J(\theta')| \\
&\leq \sum_{\bar{w}\in W^N} |J(\theta;\bar{w})p_{\bar{w}}(\bar{w};\theta) - J(\theta';\bar{w})p_{\bar{w}}(\bar{w};\theta')|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{\bar{w}\in W^N} |J(\theta;\bar{w})p_{\bar{w}}(\bar{w};\theta) - J(\theta';\bar{w})p_{\bar{w}}(\bar{w};\theta)| + \sum_{\bar{w}\in W^N} |J(\theta';\bar{w})p_{\bar{w}}(\bar{w};\theta) - J(\theta';\bar{w})p_{\bar{w}}(\bar{w};\theta')| \\
&\leq \|\theta-\theta'\| \sum_{\bar{w}\in W^N} L_0(J(\cdot;\bar{w}))p_{\bar{w}}(\bar{w};\theta) + \|\theta-\theta'\| \sum_{\bar{w}\in W^N} J(\theta';\bar{w})L_0(p_{\bar{w}}(\bar{w};\cdot)) \\
&\leq L_0(J)\|\theta-\theta'\|,
\end{aligned}
$$

where $L_0(J) = G\max_{\bar{w}\in W} L_0(p_{\bar{w}}(\bar{w};\cdot)) + \max_{\bar{w}\in W} L_0(J(\cdot;\bar{w}))$, the first inequality follows from the triangle inequality, the third inequality follows from Lemma 7 and the Lipschitz continuity of $p_{\bar{w}}(\bar{w};\cdot)$, and the last inequality from the boundedness of $J(\cdot;\bar{w})$. This completes the proof. □