Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 00 (2018) 000–000

Energy Procedia

www.elsevier.com/locate/procedia

10th International Conference on Applied Energy (ICAE2018), 22-25 August 2018, Hong Kong, China

Advanced Building Control via Deep Reinforcement Learning

Ruoxi Jia[a], Ming Jin[a], Kaiyu Sun[b], Tianzhen Hong[b], Costas Spanos[a]

[a] *University of California Berkeley, Berkeley, CA 94720, USA*
[b] *Lawrence Berkeley National Lab, Berkeley, CA 94720, USA*
Email: ruoxijia@berkeley.edu, jinming@berkeley.edu, ksun@lbl.gov, thong@lbl.gov, spanos@berkeley.edu

Abstract

Building control is a challenging task, not least because of complex building dynamics ad multiple control objectives that are often conflicting. To tackle this challenge, we explore an end-to-end deep reinforcement learning paradigm, which learns an optimal control strategy to reduce energy consumption and to enhance occupant comfort from the data of building-controller interactions. Because real-world control policies need to be interpretable and efficient in learning, this work makes the following key contributions: **(1)** we investigated a systematic approach to encode expert knowledge in reinforcement learning through "experience replay" and/or "expert policy guidance"; **(2)** we proposed to regulate the smoothness property of the neural network to penalize the erratic behavior, which is found to dramatically stabilize the learning process and lead to interpretable control laws; **(3)** we established a virtual testbed for building control by combining the state-of-the-art building energy simulator EnergyPlus with a python environment to provide a systematic evaluation and comparison platform, which will not only further our understanding of the strengths and weaknesses of existing building control algorithms, but also suggest directions for future research. We experimentally verified our proposed deep reinforcement learning paradigm on the virtual testbed in case studies, which demonstrated promising results.

Selection and peer-review under responsibility of the scientific committee of the 10th International Conference on Applied Energy (ICAE2018).

Keywords: smart building; building control; reinforcement learning; energy efficiency; cyber-physical systems; optimization

1. Introduction

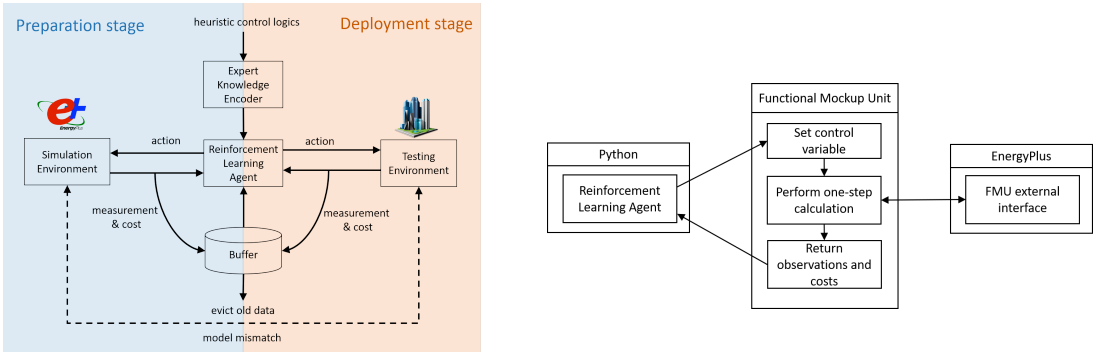
Buildings account for 40% of primary energy usage in the U.S. and are the places where people spend 90% of their time [1]. They are complex energy systems whose dynamics are driven by the interactions among numerous factors including structures, materials, equipment, surrounding environments and occupants. Modeling and analysis of building dynamics is complicated and yet is a prerequisite for the design of optimal controllers. The increasing number of functionalities housed by buildings such as demand response and on-site generation make the optimization of building control even more challenging. In practice, one often trades away controller performance in favor of simplicity of control strategies. Current state-of-the-art building control systems rely on a combination of

Selection and peer-review under responsibility of the scientific committee of the 10th International Conference on Applied Energy (ICAE2018).

Proportional-Integral-Derivative (PID) feedback control and schedule-based setpoint managing without consideration of all of the necessary information to determine optimal performance trajectories for given objectives. Despite the research efforts that attempt to construct intricate building models [2] that help predict and optimize control performance, the expenses and skills required for installation and maintenance have made such building control schemes difficult to deploy in practice.

Reinforcement learning (RL) is a potentially powerful tool for building control, which aims at guiding an agent to perform a task as efficiently and skillfully as possible through interactions with the environment. Significant progress has been made recently by combining advances in deep learning for feature representation [3] with reinforcement learning, aka deep RL. Notable examples include playing Go [4] and Atari games [5] and acquiring advanced manipulation skills using raw sensory inputs [6]. One limitation of current works, however, is that these tasks are typically characterized by the simulated environment (e.g., Atari game), deterministic rules and discrete actions (e.g., Go), and repeatability (e.g., robot manipulations). Real world control tasks like smart building automation are seldom deterministic due to the stochastic nature of the environment (e.g., people, weather), and often involve continuous actions (e.g., temperature setpoint, supply air flow rate). Naïve discretization of the continuous action space quickly becomes intractable due to the curse of dimensionality [7]. In addition, without proper guidance, a great amount of exploration is needed before finding a stable and high-performance control strategy. Thus, while deep RL provides an end-to-end paradigm that directly derives optimal control strategies from data, significant progress needs to be made in order to successfully apply RL to real-world applications.

Furthermore, it is recognized that benchmarks have played a significant role in advancing research in computer vision [8], speech recognition [9], and reinforcement learning [10]. The lack of a standardized and realistic testbed for building control makes it difficult to quantify scientific progress. On the other hand, systematic evaluation and comparison will not only further our understanding of the strengths and weaknesses of existing strategies, but also suggest directions for future research.



(a) (b)

Fig. 1. (a) Diagram of the proposed framework to control buildings via reinforcement learning; (b) Diagram of the building virtual testbed.

In this paper, we propose a framework to learn advance control strategies from data using deep RL, as illustrated by Fig. 1 (a). We build a co-simulation platform can facilitate the development of RL strategies using simulated data from EnergyPlus. We also examine a systematic way to encode expert knowledge to accelerate learning processes. In addition, we propose learning objective functions that promote stability and interpretability in the learning process. The rest of the paper is organized as follows. Related work of heating, ventilation, and air conditioning (HVAC) control strategies and reinforcement learning approaches is surveyed in Sec. 2. Sec. 3 describes the building virtual testbed. The RL framework for HVAC control is discussed in Sec. 4, which also presents expert knowledge encoding methods. Sec. 5 illustrates the approach in a case study for supply air temperature control. Concluding remarks are given in Sec. 6.

2. Virtual testbed

Availability of large datasets has proven to be crucial in many RL application areas, from game mastering to robotic locomotion and manipulation, substantially increasing the capabilities of learning-based control systems. RL

testbeds exist for many problem domains, such as the Arcade Learning Environment for Atari games, MuJoCo for multi-joint dynamics, CISTAR for traffic control, etc. In this section, we present a similarly suitable RL testbed for buildings by making use of an existing building energy simulator.

The proposed building virtual testbed (BVT) enables the study of building control through RL by taking advantage of the data-rich setting of simulation. The BVT simulates building dynamics and evaluates the performance of prospective controllers via EnergyPlus – an open-source and one of the most widely used building energy simulators. Although EnergyPlus is able to run high-fidelity simulations, it has limited capability of algorithm development and optimization. Therefore, it is difficult to implement and simulate RL-based control strategies where control actions are iteratively computed based on the current state or the instantaneous measurements taken from the controlled plant. To tackle this challenge, we adopt a co-operative simulation (co-simulation) approach, which enables the RL algorithms developed in Python to be tested on EnergyPlus building models. Co-simulation is a simulation methodology that allows individual components to be simulated by different tools running simultaneously and exchanging information in a collaborative manner. In recent years, Functional Mock-up Interface (FMI) has emerged as a standard to perform co-simulation. In the BVT, EnergyPlus is converted into a Functional Mock-up Unit (FMU) which allows the communication between Python-based RL agents and EnergyPlus building models via the FMI standard. More specifically, EnergyPlus receives the control actions determined by the RL agent, applies the control action to the current building state, and returns the resulting state and the associated costs to the RL agent (Fig. 1(b)).

3. RL framework for HVAC control

A prerequisite for using RL in building control is to framework the control problem as a Markov decision process (MDP). In this section, we first describe the MDP formulation of HVAC control. Then, we present the learning algorithm and strategies to encode expert knowledge in order to accelerate learning processes.

3.1. Building control RL framework

We can model the task of HVAC control as a MDP, with the following specifications.

State space: State contains information to decide control actions, including room temperature, occupancy, weather, time of the day, energy consumption, etc.

Action space: For HVAC control, a range of parameters can be tuned, such as hot water temperature, room-level terminal fan speed, and system-level supply air temperature set point and flow rate. In practice, the set of controllable parameters depends on the idiosyncrasies of the HVAC system.

Building dynamics: The thermal environment evolution inside a building is a physical process, and can be generally encoded in the form of transition probabilities, e.g., the probability that the room temperature increases 1 degree Celsius given a supply air flow rate.

Reward function: The control goals can be expressed by designing the reward of each state and action pair, e.g., energy savings and occupant comfort.

Building control strategy: An optimal HVAC controller maximizes rewards during building operation, and can be formulated as a (stochastic) policy, which provides a probability distribution over actions given the current building state.

3.2. Learning the optimal RL policy

We adopt the policy gradient algorithm [11] to learn a HVAC controller from the data. In each iteration, we apply the current policy and collect N traces, where each trace consists of T instantiations of (state, action, reward). We then use this batch of samples to estimate the gradient of the expected return, and take an ascent step in the direct of estimated gradient to update the policy. We iterate the process of data collection and policy update until convergence.

However, we noticed that the learning performance is unstable (e.g., the evaluated reward consistently decreases instead of increases), and the learned policy produce results that cannot be interpreted (e.g., the action changes

dramatically at a short time scale). Thus, we investigated a new strategy to stabilize the learning, which penalizes the erratic behavior of a neural network. Detailed formula can be accessed in [12]. This significantly enhanced the stability during training, and the learned policy produces smooth actions that can be interpreted by a domain expert. We also analyzed it from a control-theoretic perspective, which showed that it ensures stability of the closed-loop system [13].

3.3. Expert knowledge encoding

Because people have been operating building HVAC systems for decades, there is a vast knowledge base of experiences and heuristics. While these rules might not necessarily be optimized for individual buildings, they provide informative baselines to help speed up training. We explore two approaches to encoding expert knowledge:

Guidance via experiences replay: When the expert control experiences in the form of state-action pairs are available but the expert policy itself is unknown, we can initialize our neural network controller to clone the behavior of the expert policy by minimizing the mismatch loss, as inspired by the idea of imitation learning [14]. Throughout the training, we continuously use the expert data pool for tuning and stabilizing the algorithm via importance sampling [15].

Guidance via expert policy: When the expert policy is known and can be accessed, we can take advantage of expert knowledge by using the expert policy as a baseline. One challenge in HVAC control is that the environment is changing over time (e.g., the cooling demand during summer is much higher than during winter). This will make the rewards fluctuating dramatically during training. To reduce the variance of rewards, we propose to evaluate the expert reward, which will be used to offset the nominal reward obtained by the current policy. In essence, our goal is transformed from a generic requirement of *learning an optimal policy* to a more specific task of *learning a better policy than the expert*, which has been found highly effective to improve the policy performance. Another mechanism that we investigated is to learn a residual policy that "compliments" or "corrects" the baseline.

5. Case study: supply air temperature control

In this section, we use the BVT to benchmark the relative performance of a commonly used control logic and the RL approach to HVAC control. In addition, we evaluate and compare the potential benefits of the expert knowledge encoding schemes.

3.1. Experimental setup

We will use the supply air temperature (SAT) control in a Variable Air Volume (VAV) system as an example to compare different control strategies. VAV system is one of the most popular types of HVAC systems in the U.S. In a typical VAV system, the outside air is conditioned by a cooling component and then supplied to all zones via the VAV box at each zone. The VAV box contains a damper that can adjust the supply air flow rate to the thermal zone and a re-heating coil that heats up the air in accordance with occupants' comfort needs. The SAT control herein refers to regulating the temperature of the air before it enters the VAV terminal boxes.

Changing the SAT setpoint has different effects on the energy consumption of different components in the VAV system. For instance, reducing the SAT setpoint beyond the point at which comfort needs are met would

- reduce the airflow required by any zones that are currently in cooling mode, which reduces fan energy consumption in a cubic relationship with respect to airflow
- increase the reheat energy used in any zones currently in heating mode in a linear relationship to temperature
- increase cooling energy at the cooling component depending on the status of the economizer, i.e., if the outside air temperature is higher than the SAT
- decrease the heating energy at the heating component when supply air after cooling coil needs to be reheated to SAT setpoint due to overcooling from dehumidification
- reduce the chilled water temperature which reduces chiller efficiency and thus increases chiller electricity use.

Therefore, finding the optimal SAT setpoint is a dynamic optimization as it depends on the relative costs of fan energy, cooling energy, zone reheat energy and comfort needs at that particular moment. Our experiments consider a

single-storied building consisting of five zones and use San Francisco weather data to simulate the outdoor condition of the building.

3.2. Baseline

The current best practice for controlling the SAT, described in a proposed guideline under development by ASHRAE GPC 36 and in operation in many VAV buildings today, is a demand-based reset that constrains the range of possible SAT setpoints based on the outside air temperature. Low outside air temperature allows full reset to the maximum SAT. As the outside air temperature increases, the maximum possible SAT setpoint decreases. The actual SAT setpoint is also adjusted to the number of cooling requests: Every T minutes,

- Increase the SAT setpoint by SP_{trim}
- If there are more than I cooling requests, respond by decreasing the setpoint by $SP_{res} \times (R\text{-}I)$ but no more than $SP_{res\text{-}max}$, where R is the actual number of requests and I is the threshold to ignore.

The SAT control logic described above is also called "trim and respond" (TR), which tailors the treatment to the space based on outside air temperature as well as the comfort feedbacks in each zone. In this paper, we use TR as the baseline control strategy and set the parameters associated with TR according to [16].

3.3. Experimental results

Single-objective control. In order to understand how tuning SAT would affect energy and comfort costs, we first separate two objectives and perform PG to optimize a single objective. One iteration of PG uses a year's simulation data. Fig. 2 (a) shows the change of SAT before and after one iteration of PG when only energy costs are included into the control objective. We can see that an energy-efficient HVAC control favors a higher SAT setpoint. The baseline strategy keeps the highest allowable SAT except a few dips during 10:00-15:00 in response to the cooling requests. Since our learned policy is a combination of baseline policy and a residual policy that is updated at every iteration, the learned SAT trace also exhibits a similar dip during 10:00-15:00. After one iteration of PG, the SAT increases to a higher level in order to save energy costs.

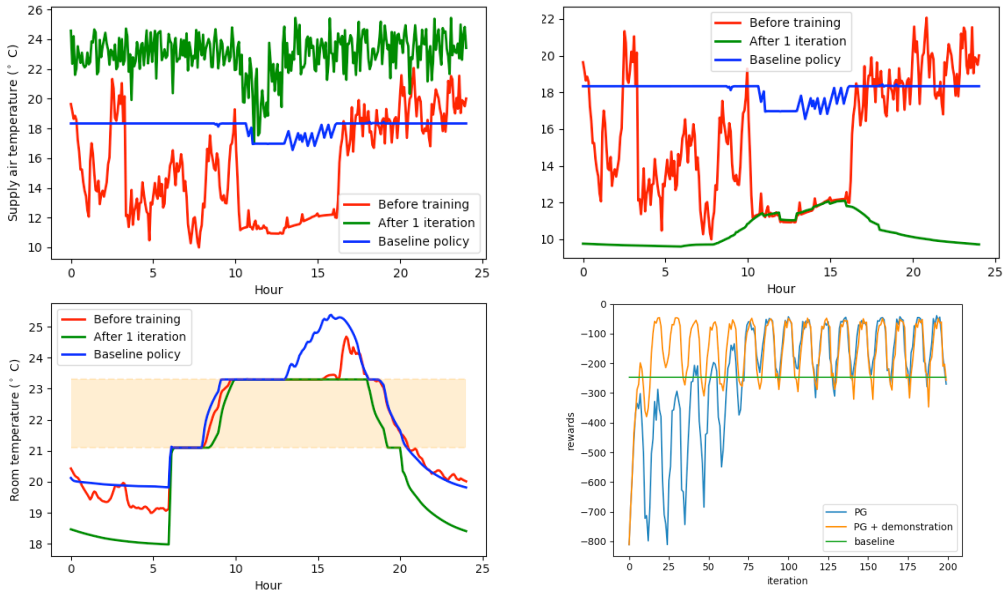


Fig. 2. (a) SAT traces when only the energy objective is optimized. (b) SAT and (c) room temperature traces when only the comfort objective is optimized. (d) Compare the learned policies against the baseline.

By contrast, Fig. 2 (b) illustrates the variation of SAT when only comfort costs are considered. It can be observed that a comfort-oriented HVAC control prefers to reduce SAT. The room temperature is also plotted in Fig. 2 (c)

when SAT obtained from the baseline policy and the learned policies at different training iterations are applied to the building. We can see that after one iteration of PG the learned policy can already maintain the room temperature within the comfort zone.

Multi-objective control. Fig. 2(d) illustrates the relative performance of the learned policy by PG and the TR baseline. After training for 75 iterations (using 6 years' synthetic data), the performance of PG can exceed the TR baseline. In addition, by incorporating the expert demonstrations into training, PG can achieve faster convergence – one-year training can already lead to better control than the baseline.

6. Conclusion

This paper presented a framework to control building systems via reinforcement learning. We introduced the building virtual testbed that allows various RL algorithms to be tested on the state-of-the-art building simulation models. In addition, we presented various methods to incorporate domain knowledge to accelerate the convergence of learning-based control algorithms. We showcased the proposed virtual testbed and domain knowledge encoding strategies via the experiments on a small-scale simulation building. The results demonstrated that the policy gradient with proper expert guidance can achieve better control performance than the current best practice control logic used in buildings. For future work, we intend to deploy the proposed RL framework in real-world HVAC systems.

Acknowledgements

This work is supported by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore–Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) program.

References

[1] Klepeis, N. E., et al. "The National Human Activity Pattern Survey: a resource for assessing exposure to environmental pollutants." Journal of Exposure Science and Environmental Epidemiology 11.3: 231, 2001.
[2] Kim, D. and Braun, J.E. "Reduced-order building modeling for application to model-based predictive control." Proceedings of SimBuild, 5(1), 554-561, 2012.
[3] LeCun, Y., Bengio, Y., and Hinton, G. "Deep learning." Nature 521.7553: 436, 2015.
[4] Silver, D., et al. "Mastering the game of Go with deep neural networks and tree search." Nature 529.7587: 484-489, 2016.
[5] Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." Nature 518.7540: 529, 2015.
[6] Lillicrap, T. P., et al. "Continuous control with deep reinforcement learning." arXiv preprint arXiv:1509.02971, 2015.
[7] Bertsekas, D. P., et al. "Dynamic programming and optimal control." Vol. 1. No. 2. Belmont, MA: Athena scientific, 1995.
[8] Deng, J., et al. "ImageNet: A large-scale hierarchical image database." IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[9] Hirsch, H. G., and Pearce, D. "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop, 2000.
[10] Duan, Y., et al. "Benchmarking deep reinforcement learning for continuous control." In Proceedings of the International Conference on Machine Learning, pp. 1329-1338, 2016.
[11] Williams, R. J. "Simple statistical gradient-following algorithms for connectionist reinforcement learning." Reinforcement Learning. Springer, Boston, MA, 5-32, 1992.
[12] Jia, R., Jin, M., and Spanos, C. "Building Automation via Deep Reinforcement Learning." In submission, 2018. URL: http://www.jinming.tech/papers/BuildingRL_appendix.pdf
[13] Jin, M. and Lavaei, J. "Control-Theoretic Analysis of Smoothness for Stability-Certified Reinforcement Learning," Preprint, 2018. Online: http://www.ieor.berkeley.edu/~lavaei/RL_1_2018.pdf
[14] Schaal, S. "Is imitation learning the route to humanoid robots?" Trends in cognitive sciences 3.6: 233-242, 1999.
[15] Siegmund, D. "Importance sampling in the Monte Carlo study of sequential tests." The Annals of Statistics: 673-684, 1976
[16] Raftery, P., Li, S., Jin, B., Ting, M., Paliaga, G., Cheng, H. "Evaluation of a cost-responsive supply air temperature reset strategy in an office building." Energy and Buildings, 158, 356-370. 2018.