Comparison of two deep reinforcement learning algorithms towards an optimal policy for smart building thermal control

Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Comparison of two deep reinforcement learning algorithms towards an optimal policy for smart building thermal control

To cite this article: Alberto Silvestri *et al* 2023 *J. Phys.: Conf. Ser.* **2600** 072011

Comparison of two deep reinforcement learning algorithms towards an optimal policy for smart building thermal control

Alberto Silvestri[1], **Davide Coraci**[2], **Duan Wu**[3], **Esther Borkowski**[1], **Arno Schlueter**[1]

[1] Architecture and Building Systems, ETH Zurich, Switzerland
[2] Department of Energy, TEBE research group, BAEDA Lab, Politecnico di Torino, Italy
[3] Mitsubishi Electric R&D Centre Europe B.V., Livingston, UK

E-mail: silvestri@arch.ethz.ch

Abstract. Heating, Ventilation, and Air Conditioning (HVAC) systems are the main providers of occupant comfort, and at the same time, they represent a significant source of energy consumption. Improving their efficiency is essential for reducing the environmental impact of buildings. However, traditional rule-based and model-based strategies are often inefficient in real-world applications due to the complex building thermal dynamics and the influence of heterogeneous disturbances, such as unpredictable occupant behavior. In order to address this issue, the performance of two state-of-the-art model-free Deep Reinforcement Learning (DRL) algorithms, Proximal Policy Optimization (PPO) and Soft Actor-Critic (SAC), has been compared when the percentage valve opening is managed in a thermally activated building system, modeled in a simulated environment from data collected in an existing office building in Switzerland. Results show that PPO reduced energy costs by 18% and decreased temperature violations by 33%, while SAC achieved a 14% reduction in energy costs and 64% fewer temperature violations compared to the onsite Rule-Based Controller (RBC).

1. Introduction

Thermally Activated Building Systems (TABS) are becoming increasingly popular due to their energy efficiency and cost savings. TABS use the thermal mass of a building's structure to store and release heat, reducing the need for mechanical heating and cooling. However, designing and optimizing TABS can be challenging due to the complex interactions between various parameters, such as the building's orientation, size, and occupancy patterns [1]. One of the most commonly used control strategy for TABS is Rule-Based Control (RBC), which relies on pre-determined rules, expert knowledge, and assumptions related to weather, occupancy, and other exogenous factors. While RBC is a simple and effective control strategy, it is unable to adapt to changing conditions, account for thermal inertia, or optimize multi-objective control problems which can include predictions about exogenous and endogenous disturbances influencing building behavior, leading to sub-optimal control policies [2]. The limitations related to the RBC operation can be addressed by using advanced control strategies such as Deep Reinforcement Learning (DRL) [3]. DRL is a category of machine learning algorithms that uses trial-and-error interactions with a system to learn a control policy that maximizes a predefined reward function. In the case of TABS control, the reward function can be designed so that the agent minimizes energy consumption while maintaining thermal comfort for occupants. Recent advances in DRL have shown promising results in optimizing building systems, including TABS [4]. However, there are still limitations to using DRL for TABS optimization. DRL algorithms generally require many iterations to converge to an optimal policy [3]. Additionally, DRL algorithms can be computationally costly and may require powerful hardware to train and run [5]. Despite these limitations, the potential benefits of using DRL for TABS optimization are significant. DRL can leverage the large amount of data that is typically available from modern building management systems, enabling more effective control strategies without the need for detailed knowledge of the underlying physics of the system, which can reduce engineering efforts and costs. Once installed, DRL can help reduce carbon emissions and improve the sustainability of buildings by reducing energy consumption and improving thermal comfort. Furthermore, DRL can help reduce operating costs for building owners and improve the overall value of the building.

1.1. Main contributions

This paper compares the performance of two state-of-the-art DRL algorithms, Proximal Policy Optimization (PPO) [6] and Soft Actor-Critic (SAC) [7], in optimizing a simulated TABS. The main contributions of this work are:

- The application of two DRL algorithms to a simulated TABS managing the valve opening percentage in the supply water heating loop, with a focus on energy efficiency and comfort.
- The evaluation of the performance of the algorithms in terms of energy consumption and indoor temperature control after being trained for a limited number of time steps.
- The discussion of the implications of the results for real-world applications of DRL in building systems optimization.

The performance comparison between PPO and SAC was performed to provide insights into the strengths and weaknesses of each algorithm and their potential for optimizing TABS. The study demonstrates the benefits of using advanced control strategies like PPO and SAC over RBC, promoting the adoption of DRL in buildings for better energy efficiency and comfort.

2. Control problem formulation

In this study, the simulation environment consists of a grey-box Resistor-Capacitor (RC) model emulating the building dynamics identified on real data. The objective of the agent is to control the water flow rate of the TABS by choosing the percentage of valve opening at each control time step k of 5 minutes in order to maintain a comfortable indoor temperature while minimizing energy cost. To compare the performance of the DRL agents against a baseline, a RBC was considered. The RBC mimics the behavior of the controller used on the real site and was designed with a three-degree hysteresis, which activates the heating at full power (i.e., 100% of valve opening) once the temperature hits the lower limit at 21 °C and maintains it on until the temperature reaches 24 °C. The RBC implemented in the simulation did not include a night setback feature, as it was not part of the original controller. The considered building is occupied from Monday to Saturday from 7:00 to 21:00. The reward function is designed to penalize high energy consumption and temperature fluctuations that may cause occupant discomfort.

2.1. Design of DRL controllers

Before training the algorithm, the main features of the Deep Reinforcement Learning (DRL) controllers, including the action-space, state-space, and reward function, need to be defined.

2.1.1. Action-space and state-space The action-space A_k consists of all possible actions that can be performed by the agent. In this paper, the action space A_k is continuous and defined as $0 \leq u_{valve}(k) \leq 1$. Both PPO and SAC can handle control problems with this kind of action space. At each control time step k, the agent selects the percentage of valve opening u_{valve}, which is directly proportional to the fraction of the nominal heating power (i.e., 1.5 kW) supplied by the TABS. The state-space consists of an array of observations provided as input to the agent. The state variables analyzed in this work are listed in Table 1, with the reference time step and their lower and upper bounds used to re-scale the state space through a min-max normalization before providing the variables to the DRL controllers.

Table 1. Variables included in the DRL state-space.

Variable	Min value	Max value	Unit	Time step
Indoor air temperature	-10	40	°C	k, k-1, .., k-6
Outdoor air temperature	-10	20	°C	k-6, k-5, ..., k, k+1, ..., k+48
Lower temperature bound	15	21	°C	k, k+1, ..., k+48
Upper temperature bound	24	30	°C	k, k+1, ..., k+48

Outdoor air temperature is included in the state-space as it is the most influencing ambient variable affecting building energy consumption and indoor temperature evolution [2]. In addition to the current outdoor temperature value, the agent receives 6 lagged values and 48 future time steps, corresponding to a time window of the past 30 minutes and the following 4 hours. The information related to the *Indoor air temperature* is integrated into the state-space at the current control time step k and for 6 lagged values to assess the temperature progression in the building over time, accounting for the influence of building thermal dynamics [5]. Moreover, information related to the presence of occupants for the future 48 time steps is included in the state-space as a function of *Lower and Upper temperature bound* values. During occupancy periods, the value of lower and upper temperature bounds are equal to 21 °C and 24 °C, while during unoccupied periods, the acceptability temperature range is relaxed to 15 °C and 30 °C. Overall, the state-space results in a vector of 158 continuous values.

2.1.2. Reward function The reward function is characterized by the weighted sum of two factors: comfort violations, which measure the squared deviation of the temperature of the zone from the desired temperature bounds, and energy consumption. The reward function J is:

$$J = -\max(0, T_i - \overline{T_i})^2 - \max(0, \underline{T_i} - T_i)^2 - \lambda E_c \tag{1}$$

where $\overline{T_i}$ and $\underline{T_i}$ are the upper and lower temperature bounds, λ is a weighting factor, and E_c is the energy cost, proportional to the control action. We set λ to equally penalize 1 kW power use or a 1 °C deviation from temperature bounds.

3. Implementation

Two state-of-the-art model-free DRL algorithms, PPO and SAC, are implemented and compared in this study. PPO aims to optimize the agent's policy while ensuring that the policy updates are stable and sample-efficient by constraining the size of the policy update [6]. On the other hand, SAC uses an entropy-regularized objective function to encourage exploration and prevent premature convergence [7]. The main difference between the two algorithms is that PPO is an on-policy method, which means it only uses data collected from the current policy, while SAC is an off-policy method, which means it can learn from data collected by any policy. The choice between on-policy and off-policy learning involves a trade-off between stability and data efficiency. On-policy algorithms generally offer greater stability and less sensitivity to hyperparameters but require more data to learn effectively, whereas off-policy algorithms tend to be more efficient with data, but may be less stable. The theoretical foundation regarding PPO and SAC can be found in [6, 7] respectively. This study employs the Stable Baselines 3 [8] implementation of PPO and SAC with default hyperparameters, except for the learning rate, which is tuned using grid search. The learning rate is one of the hyperparameters that has a significant impact on the performance of a DRL algorithm, affecting how quickly the algorithm converges to a control policy. The same neural network architecture is employed to represent the policy and value functions for both algorithms, consisting of four fully connected hidden layers with 64 units each and a rectified linear activation function (ReLU). Both algorithms are trained using the same reward function, as described in Section 2.1.2, for 500 weekly episodes (i.e., 2048 training steps per episode), corresponding to about one million steps. A discount factor of 0.99 and a batch size of 2048 are chosen for both algorithms. The DRL algorithms are trained on data collected from the real office building where the RBC is implemented as the actual control strategy. PPO and SAC are compared to the onsite RBC in terms of energy cost and indoor temperature performance during a five-month static deployment from November 1, 2021, to March 31, 2022. The total energy cost is calculated as the sum of the total cost over the whole testing period, as defined in the following equation:

$$C_E = \sum_{k=0}^{k_{end}} c_E \cdot E_{cons,k} \tag{2}$$

where c_E is the electricity price and is defined as a constant value equal to 0.3 CHF/kWh over the whole period, while $E_{cons,k}$ is the TABS energy consumption, evaluated in kWh. The indoor temperature performance was evaluated in terms of the cumulative sum of temperature violations during the whole testing period and defined as follows:

$$T_{viol} = \sum_{k=0}^{k_{end}} b_{occ,k} \cdot T_{viol,k} \tag{3}$$

where $b_{occ,k}$ is a boolean variable being 1 when occupants are present or 0 otherwise. A temperature violation $T_{viol,k}$ is calculated as the absolute temperature difference between the indoor temperature and the upper $\overline{T_i}$ or lower limit $\underline{T_i}$, and can have different expressions depending on the value of the indoor temperature T_i [9]:

$$T_{viol,k} = \begin{cases} \underline{T_i} - T_i & \text{if } T_i < \underline{T_i} \\ 0 & \text{if } \underline{T_i} \leq T_i \leq \overline{T_i} \\ T_i - \overline{T_i} & \text{if } T_i > \overline{T_i} \end{cases} \tag{4}$$

4. Results and discussion

Figure 1 shows a typical week during the heating season. Both controllers successfully learned the system's dynamics and captured the unoccupied period setback with the corresponding preheating phase, anticipating constraint tightening. Although the two DRL controllers behave similarly, with the selected hyperparameters, the PPO controller is more aggressive and tries to minimize as much as possible the energy consumption, keeping the temperature close to the lower bound, which results in some violations. However, the majority of temperature violations (below the lower temperature bound) primarily occur during the night setback period when the building is unoccupied, thereby posing no comfort-related issues for the occupants. On the other hand, the SAC controller maintains the zone at a slightly higher temperature, ensuring that the indoor temperature is always included in the desired bounds. Following the same logic, the SAC agent starts the preheating 2 or 3 hours before the PPO controller. The different behavior can also be noticed in managing the requested thermal power in the bottom plot of Figure 1, especially during the sixth day of the reported week. The PPO shows a smoother behavior than SAC, resulting in less strain on the control valve. Overall a similar trend is highlighted,

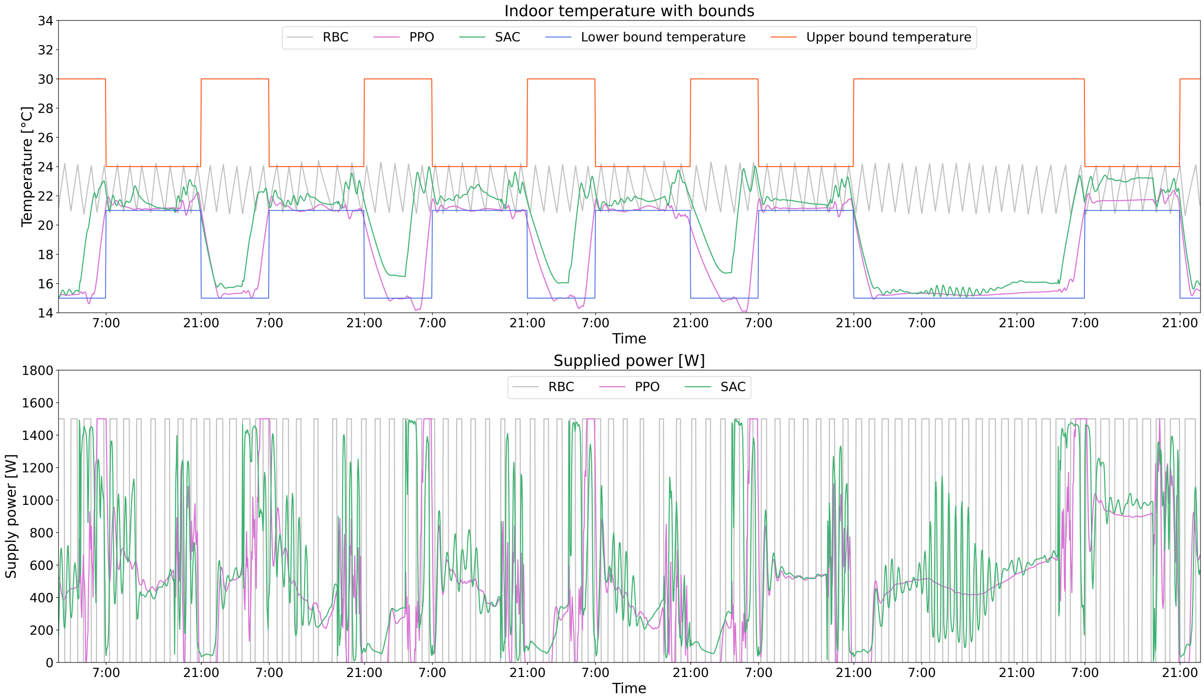

Figure 1. Sample of the behavior of the PPO and SAC agents versus the baseline RBC over a week during the heating season. The top plot shows the controlled zone temperature with the corresponding bounds, and the bottom plot depicts the power input.

confirming that both controllers try to approximate the optimal policy for this particular system and cost function. The performance of the three controllers is shown in Figure 2. In the left

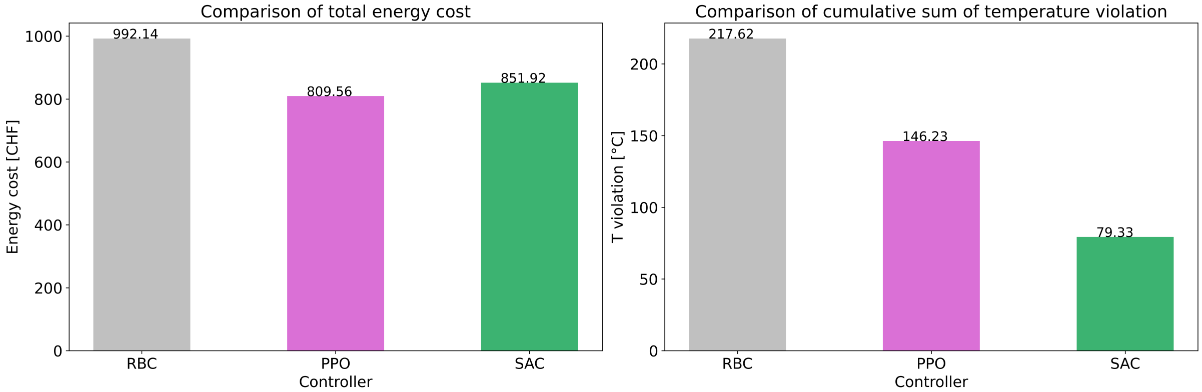

Figure 2. Performance comparison of the DRL agents and the baseline RBC. The total energy cost over the whole period is shown on the left side and the cumulative temperature violations during occupied periods are shown on the right side.

plot, it can be observed that both DRL controllers outperform the RBC in terms of total energy cost. PPO can lead to more monetary savings in the long run, reducing the expenses by almost 200 CHF compared to the RBC and 43 CHF compared to the SAC agent over the analysed period, as the PPO controller is more aggressive. Conversely, as shown in the right plot, SAC achieves the best comfort, being able to minimize the temperature violations during occupied periods by 64% and 45% with respect to the baselines and the PPO, respectively. Considering this metric, PPO is slightly worse than SAC but much better than the baselines, with about 33% fewer temperature violations. Moreover, the main advantage of using DRL-based controllers for reducing energy cost is primarily linked to optimizing the night setback period. Excluding the night setback period, both PPO and SAC achieved a reduction of approximately 5% in energy costs over the analyzed five months when compared to the RBC. The energy costs were 445 CHF for PPO and 450 CHF for SAC, whereas the RBC resulted in a cost of 472 CHF. Furthermore, both PPO and SAC demonstrated similar temperature control performance.

5. Conclusions

In this paper, two DRL algorithms were compared to a baseline RBC in terms of energy cost and temperature control during a five-month heating season in a simulation of an office building in Switzerland by managing the percentage valve opening in a TABS. The DRL controllers outperformed the RBC, with PPO reducing energy costs by 18% and temperature violations by 33%, and SAC reducing energy costs by 14% and temperature violations by 64%. Therefore, the use of DRL agents can lead to substantial energy cost reductions and improved indoor temperature management in buildings, making them more energy-efficient and comfortable for occupants. Nevertheless, this study has some limitations that future work should address. To assess the generalizability of the proposed approach, further research could focus on evaluating the performance of PPO and SAC in real-world conditions not only in the case study presented here but also in other buildings investigating a transfer learning approach. In addition, the study used default hyperparameters for the DRL algorithms, except for the learning rate, which was tuned using grid search. While this approach is common in the literature [2], other hyperparameters could have a significant impact on the performance of the algorithms and should be investigated in future studies. Furthermore, the performance of PPO and SAC could be compared when implemented in more complex HVAC systems, where multiple subsystems need to be controlled simultaneously, or integrated energy systems, with photovoltaic, thermal and electrical storage. In this case, the performance of model-based controllers, such as model predictive control, could be evaluated in addition to the comparison with RBC.

Acknowledgments

The data used in this study were collected and made available by Empa with the support of the Swiss Federal Office of Energy and the Swiss National Science Foundation. This study was also partly financed by Mitsubishi Electric R&D Centre Europe B.V. [ETH ID No 16234].

References

[1] Lydon G P, Caranovic S, Hischier I and Schlueter A 2019 *Energy and Buildings* **202** 109298 ISSN 0378-7788 URL `https://www.sciencedirect.com/science/article/pii/S0378778819305201`

[2] Brandi S, Piscitelli M S, Martellacci M and Capozzoli A 2020 *Energy and Buildings* **224** 110225 ISSN 03787788 URL `https://linkinghub.elsevier.com/retrieve/pii/S0378778820308963`

[3] Sutton R S and Barto A G 2018 *Reinforcement Learning: An Introduction* 2nd ed (The MIT Press) URL `http://incompleteideas.net/book/the-book-2nd.html`

[4] Wang Z and Hong T 2020 *Applied Energy* **269** 115036 ISSN 0306-2619 URL `https://www.sciencedirect.com/science/article/pii/S0306261920305481`

[5] Brandi S, Fiorentini M and Capozzoli A 2022 *Automation in Construction* **135** 104128 ISSN 0926-5805 URL `https://www.sciencedirect.com/science/article/pii/S0926580522000012`

[6] Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O 2017 Proximal Policy Optimization Algorithms arXiv:1707.06347 [cs] URL `http://arxiv.org/abs/1707.06347`

[7] Haarnoja T, Zhou A, Abbeel P and Levine S 2018 Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor arXiv:1801.01290 [cs, stat]

[8] Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M and Dormann N 2021 *Journal of Machine Learning Research* **22** 1–8 URL `http://jmlr.org/papers/v22/20-1364.html`

[9] Coraci D, Brandi S, Hong T and Capozzoli A 2023 *Applied Energy* **333** 120598 ISSN 0306-2619 URL `https://www.sciencedirect.com/science/article/pii/S0306261922018554`