Reinforcement learning control for indoor comfort: a survey

To cite this article: Ross May et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 609 062011

View the article online for updates and enhancements.
Reinforcement learning control for indoor comfort: a survey

Ross May¹, Xingxing Zhang¹, Jinshun Wu¹, Mengjie Han¹

¹School of Industrial Technology and Business Studies, Dalarna University, Falun 79188, Sweden
mea@du.se

Abstract Building control systems are prone to fail in complex and dynamic environments. The reinforcement learning (RL) method is becoming more and more attractive in automatic control. The success of the reinforcement learning method in many artificial intelligence applications has resulted in an open question on how to implement the method in building control systems. This paper therefore conducts a comprehensive review of the RL methods applied in control systems for indoor comfort and environment. The empirical applications of RL-based control systems are then presented, depending on optimisation objectives and the measurement of energy use. This paper illustrates the class of algorithms and implementation details regarding how the value functions have been represented and how the policies are improved. This paper is expected to clarify the feasible theory and functions of RL for building control systems, which would promote their wider-spread application and thus contribute to the social economic benefits in the energy and built environments.

1. Introduction
People spend most of their time in buildings [1], and they spend more than 80% of the day indoors. Maintenance of indoor comfort parameters is therefore significant for improving occupants’ feeling of comfort, health, morale, working efficiency as well as productivity [2]. Thermal comfort, visual comfort and indoor air quality (IAQ) seem to be the key parameters that jointly influence indoor comfort [3,4]. Building design and the building management system (BMS) are direct key factors that affect building indoor comfort. The design of buildings relates to occupancy level, ventilation, use of natural resource etc., which remains critical for indoor comfort in future building development [5]. Compared to the design of buildings, BMS considers both the maintenance and the improvement of the indoor comfort level through the diversity of control methods. A BMS generally refers to the integrated monitoring, transmitting and control of the indoor environment based on various protocols and communication interfaces. Such a characteristic enables the BMS to have a wider application in practice. The advanced control methods not only can take advantage of real-time data to produce the desired comfort level, but also can minimize the operational and maintenance cost. As a result, there is a high demand in the development of advanced control methods for future smart and economic-friendly building environments.

In practice, there are mainly real-time control and advanced control strategies. Existing real-time control uses real-time data to control the building systems, where its related impact on indoor environment is often delayed in such a dynamic environment. The existing advanced control strategies can deliver the advanced control signal to avoid the delayed influence on the indoor environment. They work effectively based on the building models, the building system models, weather forecast models, and energy tariff forecast models etc. However, these models are not as accurate in prediction as expected, e.g. [6], thus leading to potential inappropriate control ahead. Therefore, the existing control approaches are facing great challenges in real-time adaption/influence on indoor comfort and may fail to respond/maintain the indoor environment efficiently.
Reinforcement learning (RL), as one of the model-free control techniques, can be an alternative solution to such challenges when it is applied together with advanced control strategies. Model-free control techniques are able to work independently without having a knowledge of specific models, for instance, a recently realised Markov property-based RL method, which can work in both model-based and model-free environments [7]. RL tries to make optimal decisions through sequential actions in a dynamic environment. This approach is the classic model-free learning algorithm, such as Q-learning and $TD(\lambda)$, that makes RL much more attractive and efficient in artificial intelligence applications [8–10]. The efforts made on solving deep RL problems, e.g. [11], open up the possibility of working on continuous large datasets. The distinctive property of RL is that the learner or agent, via a trial-and-error paradigm, can make optimal actions without having a supervisor, which essentially fits the goal of a control problem. Particularly, in building control systems (BCSs) performances of using RL have not been analyzed from the methodological point of view and the future tasks in this field are still rare. Relative review works examining the advanced model-free control method have not drawn too much attention. Unlike energy demand response [12], this paper considers indoor comfort as the principal optimisation target. Therefore, the aim of this paper is to methodologically review the empirical works on how RL methods have been implemented in indoor comfort control among buildings, and provide instructive directions for future research.

2. Review method
In this paper, we make our search of articles in the Web of Science, ScienceDirect and Google Scholar. We do not limit the publication time. Our searching key words are

\[
\text{(building(s)) AND \left(\text{(reinforcement learning)OR (Markov decision processes) OR Q-learning}\right) \text{ AND } \left(\text{comfort OR (thermal comfort)OR (visual comfort)OR (indoor air quality)OR occupant OR (indoor environment)}\right) \text{ OR \left(\text{model free control OR (intelligent control)}\right).}}
\]

We search the words on both sides of the AND operator and only one word or phrase on either side of the OR operator. We also search Markov decision processes (MDPs) and Q-learning to guarantee that the underlying theory of RL and the most popular algorithms are covered. We also include “model free control” and “intelligent control” as alternative keywords because some articles treat RL as a special case of the control methods. We read every search outcome and exclude irrelevant articles without direct optimisation on comfort. Instead, we only include those articles that have clearly optimised comfort. Other joint optimisation objectives may have also been considered but our main interest lies in those articles containing at least one comfort component in the optimisation objectives. Doing so, we have identified 20 most relevant articles.

3. MDPs and Reinforcement learning
3.1. MDPs
In a dynamic sequential decision-making process, the state $S_t \in S$ refers to a specific condition of the environment at discrete time steps $t = 0,1,...$. By realising and responding to the environment, the agent chooses a deterministic or stochastic action $A_t \in A$ that tries to maximise future returns and receives an instant reward $R_{t+1} \in R$ as the agent transfers to the new state $S_{t+1}$. The reward is usually represented by a quantitative measurement. Figure 1 [13] shows how a sequence of state, action and reward is generated to form an MDP.

The Markov property tells us that the future is independent of the past and depends only on the present. In Fig. 1, $S_t$ and $R_t$ are the outcomes after taking an action and are considered as random variables. Thus, the joint probability density function for $S_t$ and $R_t$ is defined by:

\[
p(s', r|s, a) = \mathbb{P}[S_{t} = s', R_{t} = r| S_{t-1} = s, A_{t-1} = a],
\]
where \( s, s' \in S, r \in R \) and \( a \in A \). It can be seen from Eq. (1) that the distribution of state and reward at time \( t \) depends only on the state and action one step before. A Bellman optimality equation is used for optimizing cumulative future reward.

3.2. Reinforcement learning

Value-based algorithms, e.g. the off-policy Q-learning [14], start with a random value function and updates to an improved value function in an iterative process until reaching the optimal value function \( Q(S, A) \). The optimal policy is made by selecting the optimal value function given a certain state. For some value based methods, e.g. the on-policy SARSA and SARSA(\( \lambda \)) [15], they evaluate policies by constructing their value functions and use these value functions to find improved policies.

Policy-based methods use optimisation techniques to directly search for an optimal policy. The policy-based method gives better convergence, especially for the continuous state-action space. In episodic experiments, the average reward for each time step is used as the objective function. The gradient ascent technique iteratively improves the estimation. The action preference is usually assigned to a probability to avoid the deterministic policy.

4. Applications

Table 1 gives a summary of the reviewed literature pertaining to RL methods applied to comfort controls in buildings. We will show specific learning algorithms and the classes they belong to for each publication. We also investigate the representation of value functions to highlight potential optimisation methods. Pre-training refers to whether or not the agents were implemented with pre-trained policies using existing data or simplified models of the physical system. Unless otherwise stated, any reference to RL methods should be assumed to be model-free methods.

| Ref | Optimisation objectives | Class | Algorithms | Pre-training |
|-----|-------------------------|-------|------------|--------------|
| [16] | IAQ | Energy consumption | Value-based | Q-Learning | No |
| [17] | Thermal | Energy consumption | Value-based | Q-Learning | No |
| [18] | Thermal and IAQ | Energy consumption | Value-based | RLS-TD(\( \lambda \)) | No |
| [19] | Thermal and IAQ | Energy consumption | Value-based | RLS-TD(\( \lambda \)) | No |
| [20] | Thermal | Energy consumption | Value-based | Q-Learning | No |
| [21] | Thermal and IAQ | Energy consumption | Value-based | SARSA | No |
| [22] | Thermal and IAQ | N/A | Actor-Critic | TD(lambda) | No |
| [23] | Thermal and IAQ | Energy consumption | Value-based | Fitted Q-Iteration | N/A |
| [24] | Light | Energy consumption | Value-based | Value iteration | N/A |
| [25] | Thermal | Energy consumption | Value-based | Q-Learning | No |
| [26] | Thermal | Energy consumption | Value-based | Fitted Q-Iteration | No |
| [27] | Thermal | Energy consumption | Value-based | Q-Learning | No |
| [28] | Thermal | Energy consumption | Value-based | Fitted Q-Iteration | No |
| [29] | Thermal and IAQ | Energy cost | Value-based | Q-Learning | N/A |
Table 1: Performance Comparison of Different Control Strategies

| Article | Control Strategy | Energy Consumption | Energy Cost | Value-Based Learning | Q-Learning Based | Pre-Training Required |
|---------|------------------|--------------------|-------------|----------------------|-----------------|----------------------|
| [30]    | Thermal          | Value-based        | N/A         | No                   |                 |                      |
| [31]    | Thermal and IAQ  | Value-based        | Q-Learning  | N/A                  |                 |                      |
| [32]    | Thermal          | Value-based        | Variant of DQN, Q-learning | N/A     |                 |                      |
| [33]    | Thermal          | Value-based        | Q-Learning  | No                   |                 |                      |
| [34]    | Thermal          | Value-based        | Q(λ)        | Yes                  |                 |                      |
| [35]    | Thermal          | Value-based        | Actor-Critic | A2C                 |                 |                      |

Energy consumption is mostly controlled with thermal comfort [25,27]. Dalamagkidis and Kolokotsa implemented an RL control for an HVAC system with the goal of maximising both thermal comfort and energy conservation, with a heavier emphasis on thermal comfort [20]. They compared the performance of the RL control to the performance of a fuzzy-PD and a common on/off control over a 5 year simulated time period. They found that after 4 years of simulation the RL control achieved as good as if not better performance than the other two controls. They also suggested pre-training the control before deploying it in a real environment to mitigate suboptimal performance due to policy exploration. For example, pre-training the RL control was able to improve the performance of the low-energy building system in an acceptable period of time by using RL to tune a fuzzy rule-based supervisory control for an HVAC system [34]. Off-policy training for HVAC is also applicable [17]. Through a comparative analysis they found their RL control outperformed, in terms of energy cost, two common strategies for controlling HVAC, namely, the “Always On” and “Programmable Control” methods. Improved thermal comfort was demonstrated.

Baghaee and Ulusoy use RL for operating an HVAC system [16]. The objective of the control was to maintain CO2 concentration at an acceptable range while minimizing energy consumption. In a simulation study they compared their RL control to an On-Off and set point control. Their RL method outperformed the other two controls regarding energy consumption and CO2 concentration. Studies controlling combined factors are rare in the 1990s. For example, Jouffe used RL to tune a ventilation controller for controlled temperature and relative humidity [23]. The policy obtained from the control was exactly to the experts’ specifications. Mozer used RL to control an HVAC and water heating system [36]. The aim of the control framework was to minimise both discomfort (heating and lighting) and energy cost. In Mozer’s work, RL control was found to be more efficient than explicit model-based control of a setpoint generator.

5. Conclusions
Indoor environment affects not only working efficiency and living standards, but also influences the occupants’ health. Apart from building design, efficient control methods on indoor environment guarantees occupants’ satisfaction. This paper briefly examines and analyzes empirical articles regarding the model-free reinforcement learning control method on indoor comfort in buildings. The cutting-edge RL technique has drawn only limited attention regarding the indoor environment oriented smart building controls, even though some studies have empirically tested its feasibility and comparability to other methods. The promising results lead us to a new frontier of building control. We have identified twenty empirical articles in this field, which is much less than the studies in building energy control and needs to be extended. The value-based Q-learning is easy and straightforward to implement and it dominates among learning algorithms. However, the value-based method fails to work when the action space is large or continuous. This leaves a question of how policy-based or Actor-Critic algorithms perform in a practical building environment. The computation platform and the ways of interaction with BMSs are important for conducting real-time control. Especially in the works with physical tests, the working paradigms are still vague. For example, policy-based and Actor-Critic algorithms require more function approximations and thus the power of computing resources should be updated accordingly. We anticipate practical works in standardizing the measurement of indoor comfort and integrating computation platforms and the ways of interaction with BMSs into the smart building systems in future works.
References

[1] Shaikh P H, Nor N B M, Nallagownden P, Elamvazuthi I and Ibrahim T 2013 Robust Stochastic Control Model for Energy and Comfort Management of Buildings Australian Journal of Basic and Applied Sciences 7 137–44

[2] Li N, Cui H, zhu C, Zhang X and Su L 2016 Grey preference analysis of indoor environmental factors using sub-indexes based on Weber/Fechner’s law and predicted mean vote Indoor and Built Environment 25 1197–208

[3] Boodi A, Beddias K, Benamour M, Amirat Y and Benbouzid M 2018 Intelligent Systems for Building Energy and Occupant Comfort Optimization: A State of the Art Review and Recommendations Energies 11 2604

[4] Park J Y and Nagy Z 2018 Comprehensive analysis of the relationship between thermal comfort and building control research - A data-driven literature review Renewable and Sustainable Energy Reviews 82 2664–79

[5] Wang N, Phelan P E, Harris C, Lanevin J, Nelson B and Sawyer K 2018 Past visions, current trends, and future context: A review of building energy, carbon, and sustainability Renewable and Sustainable Energy Reviews 82 976–93

[6] Andersen R, Fabi V, Toftum J, Corgnati S P and Olesen B W 2013 Window opening behaviour modelled from measurements in Danish dwellings Building and Environment 69 101–13

[7] Kaelbling L P, Littman M L and Moore A W 1996 Reinforcement Learning: A Survey Journal of Artificial Intelligence Research 4 237–85

[8] Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, Graves A, Riedmiller M, Fidjeland A K, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S and Hassabis D 2015 Human-level control through deep reinforcement learning Nature 518 529–33

[9] Silver D, Huang A, Maddison C J, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T and Hassabis D 2016 Mastering the game of Go with deep neural networks and tree search Nature 529 484–9

[10] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T and Hassabis D 2017 Mastering the game of Go without human knowledge Nature 550 354–9

[11] Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, Silver D and Wierstra D 2016 Continuous control with deep reinforcement learning arXiv:1509.02971 [cs, stat]

[12] Vázquez-Canteli J R and Nagy Z 2019 Reinforcement learning for demand response: A review of algorithms and modeling techniques Applied Energy 235 1072–89

[13] Sutton R S and Barto A G 1998 Reinforcement learning: an introduction (Cambridge, Mass: MIT Press)

[14] Watkins C J C H 1989 Learning from Delayed Rewards Ph.D. thesis, University of Cambridge

[15] Rummery G and Niranjan M 1994 On-line Q-learning using connectionist systems (Cambridge University)

[16] Baghaee S and Ulusoy I 2018 User comfort and energy efficiency in HVAC systems by Q-learning 2018 26th Signal Processing and Communications Applications Conference (SIU) 2018 26th Signal Processing and Communications Applications Conference (SIU) (Izmir, Turkey: IEEE) pp 1–4

[17] Barrett E and Linder S 2015 Autonomous HVAC Control, A Reinforcement Learning Approach Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science ed A Bifet, M May, B Zadrozny, R Gavalda, D Pedreschi, F Bonchi, J Cardoso and M Spiliopoulou (Springer International Publishing) pp 3–19

[18] Chen Y, Norford L K, Samuelson H W and Malkawi A 2018 Optimal control of HVAC and window systems for natural ventilation through reinforcement learning Energy and Buildings 169 195–205

[19] Dalamagkidis K, Kolokotsa D, Kalaitzakis K and Stavrakakis G S 2007 Reinforcement learning for energy conservation and comfort in buildings Building and Environment 42 2686–98
[20] Dalamagkidis K and Kolokots D 2008 Reinforcement Learning for Building Environmental Control Reinforcement Learning ed C Weber, M Elshaw and N Michael (I-Tech Education and Publishing)

[21] Eller L, Siafara L C and Sauter T 2018 Adaptive control for building energy management using reinforcement learning 2018 IEEE International Conference on Industrial Technology (ICIT) 2018 IEEE International Conference on Industrial Technology (ICIT) (Lyon: IEEE) pp 1562–7

[22] Fu Q, Hu L, Wu H, Hu F, Hu W and Chen J 2018 A Sarsa-based adaptive controller for building energy conservation Journal of Computational Methods in Sciences and Engineering 18 329–38

[23] Jouffe L 1997 Ventilation control learning with FACL Proceedings of 6th International Fuzzy Systems Conference 6th International Fuzzy Systems Conference vol 3 (Barcelona, Spain: IEEE) pp 1719–24

[24] Park J Y, Dougherty T, Fritz H and Nagy Z 2019 LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning Building and Environment 147 397–414

[25] Pedro F, Kalyan V, Pedro L and Una-May O 2014 Using reinforcement learning to optimize occupant comfort and energy usage in HVAC systems Journal of Ambient Intelligence and Smart Environments 675–690

[26] Ruelens F, Iacovella S, Claessens B J and Belmans R 2015 Learning Agent for a Heat-Pump Thermostat with a Set-Back Strategy Using Model-Free Reinforcement Learning Energies 8 8300–18

[27] Sato K, Samejima M, Akiyoshi M and Komoda N 2012 A scheduling method of air conditioner operation using workers daily action plan towards energy saving and comfort at office Proceedings of 2012 IEEE 17th International Conference on Emerging Technologies & Factory Automation (ETFA 2012) 2012 IEEE 17th Conference on Emerging Technologies & Factory Automation (ETFA 2012) (Krakow, Poland: IEEE) pp 1–6

[28] Schmidt M, Moreno M V, Schülke A, Macek K, Mařík K and Pastor A G 2017 Optimizing legacy building operation: The evolution into data-driven predictive cyber-physical systems Energy and Buildings 148 257–79

[29] Sun B, Luh P B, Jia Q and Yan B 2013 Event-based optimization with non-stationary uncertainties to save energy costs of HVAC systems in buildings 2013 IEEE International Conference on Automation Science and Engineering (CASE) 2013 IEEE International Conference on Automation Science and Engineering (CASE) pp 436–41

[30] Sun Y, Somani A and Carroll T E 2015 Learning based bidding strategy for HVAC systems in double auction retail energy markets 2015 American Control Conference (ACC) 2015 American Control Conference (ACC) pp 2912–7

[31] Sun B, Luh P B, Jia Q and Yan B 2015 Event-Based Optimization Within the Lagrangian Relaxation Framework for Energy Savings in HVAC Systems IEEE Transactions on Automation Science and Engineering 12 1396–406

[32] Wei T, Wang Y and Zhu Q 2017 Deep Reinforcement Learning for Building HVAC Control Proceedings of the 54th Annual Design Automation Conference 2017 on - DAC ‘17 the 54th Annual Design Automation Conference 2017 (Austin, TX, USA: ACM Press) pp 1–6

[33] Yang L, Nagy Z, Goffin P and Schlueter A 2015 Reinforcement learning for optimal control of low exergy buildings Applied Energy 156 577–86

[34] Yu Z and Dexter A 2010 Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning Control Engineering Practice 18 532–9

[35] Wang Y, Velswamy K and Huang B 2017 A Long-Short Term Memory Recurrent Neural Network Based Reinforcement Learning Controller for Office Heating Ventilation and Air Conditioning Systems Processes 5 46

[36] Mozer M C 1998 The Neural Network House: An Environment that Adapts to its Inhabitants 5