Working papers in transport, tourism, information technology and microdata analysis

A review of reinforcement learning methodologies on control systems for building energy

Mengjie Han

Xingxing Zhang

Liguo Xu

Ross May

Song Pan

Jinshun Wu

Editor: Hasan Fleyeh

Nr: 2018:02

Working papers in transport, tourism, information technology and microdata analysis
ISSN: 1650-5581

A review of reinforcement learning methodologies on control systems for building energy

Mengjie Han*[a], Xingxing Zhang*[a], Liguo Xu[b], Ross May[a], Song Pan[c], Jinshun Wu[c]

Abstract:

The usage of energy directly leads to a great amount of consumption of the non-renewable fossil resources. Exploiting fossil resources energy can influence both climate and health via ineluctable emissions. Raising awareness, choosing alternative energy and developing energy efficient equipment contributes to reducing the demand for fossil resources energy, but the implementation of them usually takes a long time. Since building energy amounts to around one-third of global energy consumption, and systems in buildings, e.g. HVAC, can be intervened by individual building management, advanced and reliable control techniques for buildings are expected to have a substantial contribution to reducing global energy consumptions. Among those control techniques, the model-free, data-driven reinforcement learning method seems distinctive and applicable. The success of the reinforcement learning method in many artificial intelligence applications has brought us an explicit indication of implementing the method on building energy control. Fruitful algorithms complement each other and guarantee the quality of the optimisation. As a central brain of smart building automation systems, the control technique directly affects the performance of buildings. However, the examination of previous works based on reinforcement learning methodologies are not available and, moreover, how the algorithms can be developed is still vague. Therefore, this paper briefly analyses the empirical applications from the methodology point of view and proposes the future research direction.

Keywords:

Reinforcement learning; Markov decision processes; building energy; control; multi-agent system

[*] Corresponding authors: Mengjie Han (mea@du.se); Xingxing Zhang (xza@du.se).

[a] School of Industrial Technology and Business Studies, Dalarna University, Falun 79188, Sweden

[b] School of Management, Xi'an Jiaotong University, No.28 Xianning West Road, Xi'an, China

[c] Beijing Key Laboratory of Green Built Environment and Energy Efficient Technology, Beijing University of Technology, Beijing 100124, China

Nomenclature			
ANN	artificial neural network	MPC	model predictive control
BCS	building control system	PCM	phase change material
FL	Fuzzy logic	PI	policy iteration
FQI	fitted Q-iteration	PID	proportional-integral-derivative
HVAC	heating, ventilation, and air conditioning	PMV	predictive mean vote
		PSO	particle swarm optimization
JAL	extended joint action learning	RL	reinforcement learning
MACS	multi-agent control system	RLS	recursive least-squares
MAS	multi-agent system	TD	temporal difference
MARL	multi-agent reinforcement learning	TES	thermal energy storage
MDPs	Markov decision processes	VI	value iteration
Notations			
t	discrete time steps	$\mathbf{w}$	weight vector
S_t	state at time t stochastically	$\|\{s,a\}\|$	number of state-action pairs
$\mathcal{S}$	set of states	$\pi(a\|s)$	prob. of taking action a in state s
A_t	action at time t, stochastically	$v_\pi(s)$	state-value function
$\mathcal{A}$	set of actions	$q_\pi(s,a)$	action-value function
R_t	reward at time t, stochastically	$p(s'\|s,a)$	transition prob.
$\mathcal{R}$	set of rewards	$r(s,a)$	expected reward
π	policy	$v_*(s)$	optimal state-value function
γ	discount parameter	$q_*(s,a)$	optimal action-value function
λ	trace-decay parameter	Π_i	policy set for multi-agent
s, s'	States	ϵ	prob. of selecting random actions
a	action	$Q(S,A)$	approx. of $q(s,a)$ from data
r	reward	α, β	step size parameter
$q(s,a;\mathbf{w})$	mapping rule for the q-function	$\pi_\theta(s\|a;\boldsymbol{\theta})$	parametrized policy
$\hat{q}(s,a;\mathbf{w})$	estimate of the q-function		

1 Introduction

Building energy consumption amounts to around 30%-40% of all energies consumed in both developed and developing countries [1-4]. The tendency of power demand is still increasing. Therefore, not only does this increase the operating cost of energy consumption, it also induces growing emissions of greenhouse gases. Since buildings are also responsible for one-third of global energy-related greenhouse gas emissions [5], efficient strategies for reducing the consumption of building energy are indispensable in the future.

The building design and management systems are direct key factors that affect building energy. The design of building relates to population growth, climate change, and resource consumption trends that remain critical for future building development [6]. It is a difficult task to find better design alternatives satisfying several conflicting criteria, especially, economic and environmental performance [7]. Compared to the design of building, the building management system or building control system (BCS), which generally refers to centralised and integrated hardware and software networks [8], considers the improvement of energy utilisation efficiency, energy cost reduction, and renewable energy technology utilisation in order to serve local energy loads [9]. The diversity of control methods enables the BCS to be further developed in wider disciplines.

For buildings, around 80% of used energy comes from heating, ventilation, and air conditioning (HVAC) systems, lighting, and equipment (appliances) [2]. Moreover, the occupants' factors, e.g. human comfort and window opening, also have substantial impacts on energy consumptions. Therefore, control strategies regarding both devices and human factors have become the main target in BCSs.

Review works [8,10-15] have extensively summarised and compared conventional methods and applications for control strategies (see also Section 3). For almost all of them, a building model or optimisation model is required. An accurate model representation is a necessity for finding efficient and robust control strategies to reducing energy consumption. For complex building systems, the performance of the models becomes inferior to those well-established models in simpler systems. As an alternative solution, model-free control techniques are able to work independently without having the knowledge for specific models. Nevertheless, they have not drawn much attention in empirical studies.

A recently realised Markov property based machine learning method, reinforcement learning (RL), can work in both model-based and model-free environments [16]. However, it is the classic model-free learning algorithms, such as Q-learning and $TD(\lambda)$, that make RL much more attractive and efficient in artificial intelligence applications [17-21]. The efforts made on solving deep RL problems, e.g. [22,23], open up the possibility of working on continuous large datasets. The distinctive property of RL is that the learner or agent, via trial-and-error paradigm, can make optimal actions without having a supervisor, which essentially fits the goal of a control problem. For BCSs, performances of using RL have not been analysed from the methodological point of view and the indication for future tasks is still unclear. Therefore, the aim of this paper is to methodologically review the empirical works on how RL methods have implemented energy control for buildings, and provide instructive directions for future research.

The contributions of this paper are threefold. Firstly, this paper provides a comprehensive understanding of how RL works for building energy. Secondly, we identify the current research gap and propose future research from the methodology point of view. Thirdly, we summarise the application of RL on building energy in the multi-agent context.

In the second section of this paper, we briefly introduce the philosophy of RL and its corresponding algorithms. In Section 3 we examine the conventional control methods used on building energy. Section 4 and 5 then analyse the current publications for both single agent and multi-agent systems. Finally, section 6 concludes the findings and proposes research directions.

2 The reinforcement learning method

The idea of reinforcement learning originated from the term "optimal control" which emerged in the late 1950s, where a problem was formulated by designing a controller to minimise a measure of the behavior of a system over time [24,25]. Bellman [26] came up with the concept of Markov decision processes (MDPs) or finite MDPs, a fundamental theory of RL, to formulate optimal control problems.

The learner or *agent* of RL learns how to map situations to actions to maximise a numerical delayed reward signal. It does not have to have a "teacher" telling it how to take an action but, rather, makes decisions via implementing trial-and-error search, and recognizing the delayed reward from the environment that the agent interacts with [24,25]. RL is neither supervised learning nor unsupervised learning, it is a third category of machine learning. Whereas supervised learning gets signals of correct actions, RL gets signals from the reward of an action without knowing if the action was correct or not. RL, in a sense, is the core of machine learning techniques. In the context of artificial intelligence, RL allows the agent to automatically determine behaviors, which cannot be achieved by supervised learning or unsupervised learning.

2.1 Elements of reinforcement learning and MDPs

2.1.1 Elements

In a dynamic sequential decision-making process, the *state* $S_t \in \mathcal{S}$ refers to a specific condition of the environment at discrete time steps $t = 0,1, \ldots$. By realising and responding to the environment, the agent chooses a deterministic or stochastic *action* $A_t \in \mathcal{A}$ that tries to maximise future returns and receives an instant *reward* $R_{t+1} \in \mathcal{R}$ as the agent transfers to the new state S_{t+1}. The reward is usually represented by a quantitative measurement. A sequence of state, action and reward is generated to form an MDP (Fig. 1 [24,25]).

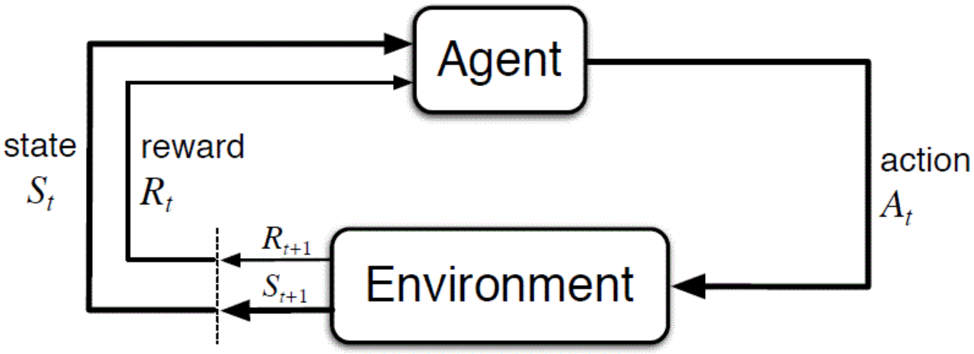

Fig. 1 The interaction between agent and environment in an MDP

2.1.2 Markov decision processes

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:

$$p(s', r|s, a) = \mathbb{P}[S_t = s', R_t = r|\, S_{t-1} = s, A_{t-1} = a], \tag{1}$$

where $s, s' \in \mathcal{S}, r \in \mathcal{R}$ and $a \in \mathcal{A}$. It can be seen from (1) that the distribution of state and reward at time t depends only on the state and action one step before. Eq. (1) implies the basic rule of how the MDP works and one can easily determine the marginal transition probabilities $p(s'|s, a)$:

$$p(s'|s, a) = \mathbb{P}[S_t = s'|S_{t-1} = s, A_{t-1} = a] = \sum\nolimits_{r \in \mathcal{R}} p(s', r|s, a). \tag{2}$$

Eq (3) gives the expected reward by using the marginal distribution of R_t:

$$r(s, a) = \mathbb{E}[R_t|S_{t-1} = s, A_{t-1} = a] = \sum\nolimits_{r \in \mathcal{R}} r \sum\nolimits_{s' \in \mathcal{S}} p(s', r|s, a). \tag{3}$$

Both Eq. (2) and (3) are used for solving the optimal value functions presented in Section 2.3.

2.2 Policies and value functions

A *policy* π is a distribution over actions given states. It fully defines the behavior of an agent by telling the agent how to act when it is in different states. The policy itself is stochastic [25] and the probability of taking an action, a, in state s is:

$$\pi(a|s) = \mathbb{P}[A_t = a|S_t = s]. \tag{4}$$

The policy can be considered as a function of actions. It acts either as a look-up table or in an approximation form (see Section 4 for the discussion). The overall goal of RL is to find the optimal policy given a state.

An optimal policy tries to maximise the expected future return from time t: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$, where $0 \leq \gamma \leq 1$ is the discount parameter. The *state-value function,* $v_\pi(s)$, and the *action-value function,* $q_\pi(s, a)$, are two useful measures in RL that can be estimated from the data. The literature defines $v_\pi(s)$, of an MDP, under policy π, as the expectation of the return starting from state s:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s\right], for\ all\ s \in \mathcal{S}. \tag{5}$$

In practical applications, $v_\pi(s)$ is more applicable for model-based problems, whereas the action-value function, $q_\pi(s, a)$, is more useful in the model-free context [25]. When the full environment or the model is unknown, episodic simulations are often used to estimate $q_\pi(s, a)$, that is, under policy π, the expectation of the return starting from state s and taking the action a:

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi[G_t|S_t = s, A_t = a] \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a\right], for\ all\ s \in \mathcal{S}\ and\ a \in \mathcal{A}. \end{aligned} \tag{6}$$

The task of finding the optimal policy, π_*, is achieved by evaluating either the optimal state-value function

$$v_*(s) = \max_\pi v_\pi(s), \tag{7}$$

or the optimal action-value function

$$q_*(s, a) = \max_\pi q_\pi(s, a). \tag{8}$$

2.3 Bellman optimality equation

The way of optimising Eq. (7) or (8) is to make use of the recursive relationships between two states or actions in a sequential order. Since the procedures are similar, we only present the relationship starting from the action-values, i.e. the *Bellman optimality equation* for $q_*(s, a)$ [27].

The backup diagrams in Fig.2 [28] show relationships between the value function and a state or state-action pairs. Fig. 2 (a) considers the optimal state-value function when taking an action. The agent looks at each of the possible actions it might take and selects the action with maximum action-value which tells the agent how good the state is. That is,

$$v_*(s) = \max_a q_*(s, a). \quad (9)$$

Similarly, Fig. 2 (b) evaluates the dynamic and stochastic environment when an action is taken. Each of the states it ends up in has an optimal value. Thus, the optimal action-value counts the immediate expected reward, $r(s, a)$, from Eq. (3), and a discounted optimal state-value:

$$q_*(s, a) = r(s, a) + \gamma \sum\nolimits_{s' \in \mathcal{S}} p(s'|s, a) v_*(s'). \quad (10)$$

Thus, as show in Fig. 2 (c), the Bellman optimality equation for $q_*(s, a)$ is obtained by substituting Eq. (9) in to Eq. (10):

$$q_*(s, a) = r(s, a) + \gamma \sum\nolimits_{s' \in \mathcal{S}} p(s'|s, a) \max_{a'} q_*(s', a'). \quad (11)$$

In a similar way, we can derive the Bellman optimality equation for $v_*(s)$. Both of them are the fundamental expressions for the MDPs. The recursive relationship assists in splitting the current value function into the immediate reward and the value of the next action. The value iteration and policy iteration algorithms, presented in Section 2.4, make use of the Bellman optimality equations to reach optimal policies.

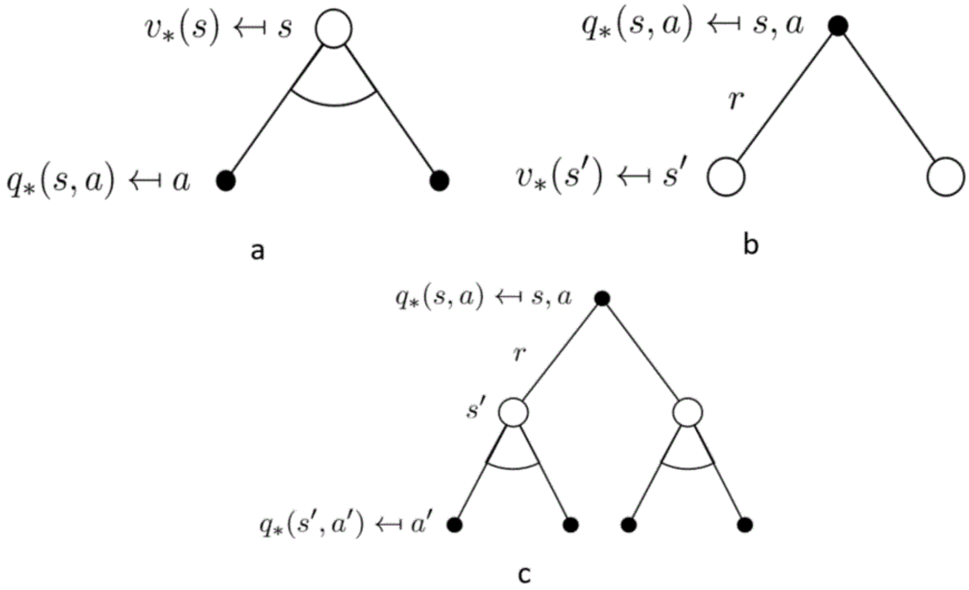
$v_*(s) \leftarrow\!\!\dashv s$
$q_*(s, a) \leftarrow\!\!\dashv a$
a
$q_*(s, a) \leftarrow\!\!\dashv s, a$
r
$v_*(s') \leftarrow\!\!\dashv s'$
b
$q_*(s, a) \leftarrow\!\!\dashv s, a$
r
s'
$q_*(s', a') \leftarrow\!\!\dashv a'$
c

Fig. 2 Backup diagrams for the optimal value functions

2.4 Category of RL algorithms

In this subsection, we briefly summarise RL algorithms and relate works applied on building energy controls (Section 4) to them. The purpose is to examine current research methods and outline future work.

There are many categorisation methods for RL algorithms. In this paper, we investigate it from two perspectives: the way of finding the optimal policy and whether or not the full model is available [29].

Table 1 Categorisation of RL algorithms

	model-based	model-free
value iteration	Q-iteration [29]	Q-learning [30]
policy iteration	policy evaluation for Q-functions [29]	*SARSA* [31]
policy search	policy gradient [32]	greedy updates [33]

In Table 1, three ways, value iteration, policy iteration and policy search, and two model types indicate a combination of six categories where we expect a variety of algorithms bringing efficient solutions to real problems. *Value iteration* (VI) starts with a random value function and updates to an improved value function in an iterative process until reaching the optimal value function. The optimal policy is made based on the optimal value function. *Policy iteration* (PI) or *generalized policy iteration* (GPI) evaluates policies by constructing their value functions and uses these value functions to find improved policies. An integration of VI and PI forms the value-based methods. We further distinguish tabular method and approximation approach as two different strategies for reaching the estimation of value functions for value based methods. *Policy search* uses optimisation techniques to directly search for an optimal policy [29]. For a model-based algorithm, the explicit model, e.g. $p(s', r|s, a)$, is required, whereas this is not needed for a model-free algorithm. The most representative algorithm for each of the model-free category is given. Detailed illustrations are presented in Section 4.

3 Conventional control methods for building energy

Various studies have reviewed the classification of different control methods for building energy [8,10-12,34,35]. Yu et al. [10] reviewed control techniques for thermal energy storage (TES) that is integrated with buildings. Shaikh et al. [8] made a survey on the intelligent control system for building energy and occupant's comfort, whereas Dounis and Caraiscos [11] focused on the agent-based control system. Aste et al. [12] summarised the model-based strategies for building simulation, control and data analytics. The previous surveys provide a framework of how the different methods relate to each other and the pros and cons of each. However, a generic challenge of conventional methods lies in the difficulty of including all unknown environmental factors in the models. Uncertainties of the environment may largely affect the model accuracy and the consequent control quality. On the other hand, the model-free RL technique on building controls has not drawn much attention and the performance of RL algorithms has thus not been evaluated yet. Even though Royapoor et al. [34] realise that RL methods are notable, a framework of implementations and explorations on efficient RL methods needs to be systematically investigated and discussed.

Empirical studies classify control methods into classic control, hard control, soft control and other control. In Fig. 3 [10,13], the RL control resides in the category of other control and explicit applications are still rare in BCS.

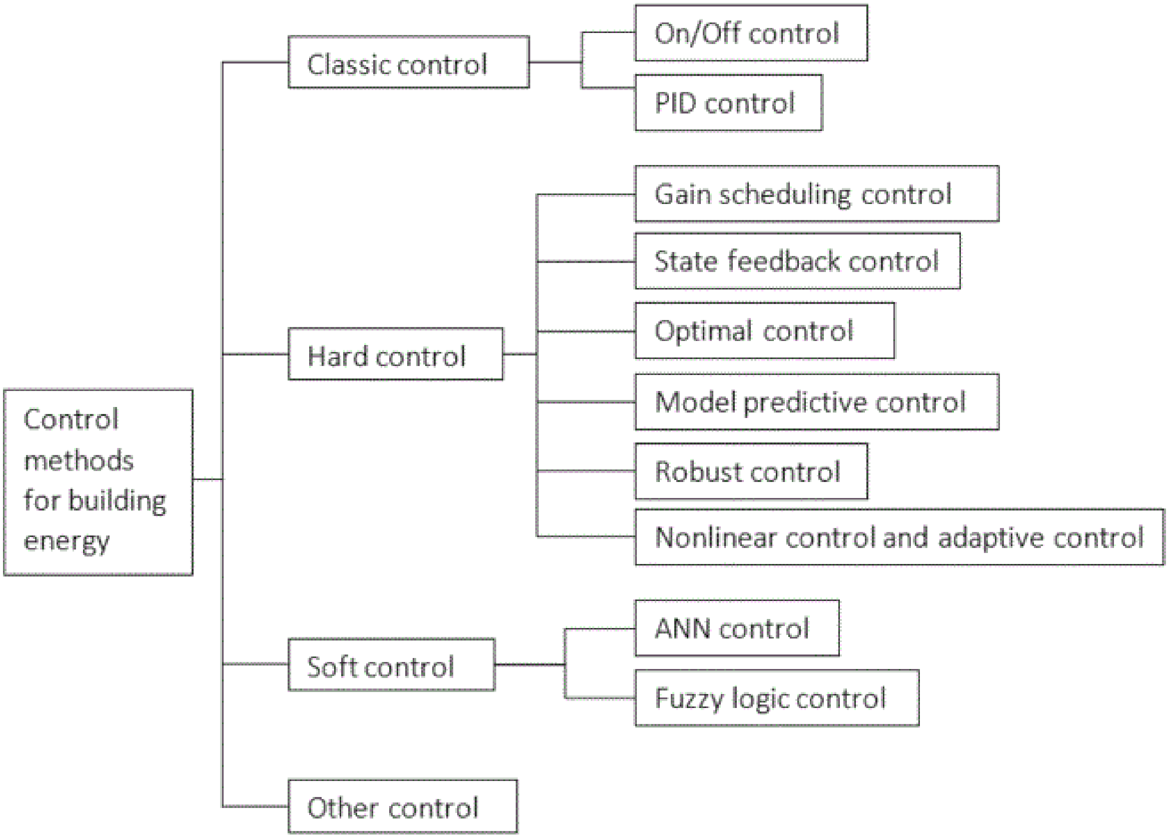

Fig. 3 The categorisation of control methods

3.1 Classic control

The classic control refers to the on-off control and proportional-integral-derivative (PID). Since the on-off control is simple to implement, there are many drawbacks to it. The binary dimension of the action space makes the controller unable to make decisions on multi-value variables, e.g. temperature set point. Moreover, the on-off control forces a compromise between good controls, long warm-up times and increased cost, in particular, the action frequency is high.

The classical local-loop proportional-integral-derivative (PID) control is also easy to implement and test. The task is to convert the error between the output signal and the input signal to an action. PID describes how the error term is handled. The proportional (P) control tries to get the system to reach the output signal as fast as possible, whereas the integral (I) control tries to adjust the summation of errors in order to remove residual errors. The derivative (D) control, on the other hand, tries to restrain the system from adjusting too quickly. Such applications can be found in HVAC systems [36,37]. The tuning and auto-tuning methods still need to be improved [38].

3.2 Hard control

The hard control, also known as hard computing [14], typically includes gain scheduling, state feedback control, model predictive control (MPC), robust control, nonlinear control and adaptive control [10,13,14].

For a system with several operating points, the gain scheduling controller designs linear controllers for each of those points and applies an interpolation strategy to obtain a global control [39]. Bianchi et al. studied the control of wind energy conversion systems by using the gain scheduling technique [39]. Moradi et al. applied gain scheduling to control both indoor temperature and relative humidity, where the controller designed, based on feedback linearisation shows a robust performance [40]. Controllers are used for determining the operating region, which needs to be optimally tuned. An increasing number of controllers bring heavy tuning burdens.

The state feedback control assumes the state of the system is available for measurements. If the action of the control depends only on the values or external input, the control is said to be a state feedback control [41]. An application can be seen for controlling boiler temperature for hot water [42].

The optimal control techniques characterised by mathematical models have been developed since the 1950s [10,43]. An objective function of minimum energy consumption, maximum thermal comfort or maximum saving is usually defined and an optimisation problem arises to be solved. A dynamic optimisation technique is applied on cooling systems to determine the maximum savings [44]; Henze et al. applied an optimal controller for thermal energy storage systems to minimise the cost of operating a cooling plant [45]. Non-linear optimal controllers have also been developed on HVAC systems [46]. For real-time controls, however, the system is usually too complex and thus requires a lot of computing effort.

The MPC has three elements: predictive models, an objective function, and a control law [10,47]. For building energy controls, two typical objectives are energy consumption and cost. A recent work comprehensively reviewed the artificial neural network (ANN) based MPC for HVAC systems [48]. Ma et al. [49] applied an economic MPC in reducing energy demand cost for a building HVAC by solving a min-max linear programming problem. Cole et al. [50] studied the optimal precooling strategy to reduce the peak energy consumption from air conditioning. The study of occupancy recognition developed an automatic HVAC control system [51]. Ascione et al. [52] considered a multi-objective MPC for cost-optimal building design. Even though there is much room to increase model performance, complex model specifications usually bring heavy computations.

The robust control methodology takes care of time-varying model uncertainty and the nonlinearities associated with the system [14,53]. For building applications, for example, a multiple-input-multiple-output robust controller has been applied on HVAC systems [54] and a linear time-invariant model has been applied to an air heating plant [55].

Nonlinear control methods are usually developed through feedback linearisation and gain scheduling approaches [56,57]. Applications on air-handling units aim to reduce the operation through controlling the air-conditioner [56], whereas a study of Variable Air Volume Air Conditioning focuses on reducing the building energy consumption [57]. The adaptive control is considered as a specific type of nonlinear control, where the controller can modify its behavior in response to changes in the dynamics of the process and the character of the disturbances [58]. Different from robust control, adaptive control does not require an unchanged control law. Adaptive control is examined, for example, in HVAC systems [59,60].

3.3 Soft control

The soft control refers to ANN control, fuzzy logic control and other evolutionary techniques [15]. ANN is a data-driven approach that learns optimal weights in a nonlinear model to minimize the errors between the estimated data and the target data. The structure for studying direct thermal comfort control is set up by specifying the error between reference comfort index predictive mean vote (PMV) and feedback PMV as input [61]. In Fig. 4 [10,61], the output value sends the control signal to the HVAC system to form measured data. The network is trained until convergence. Similar works have been extended by applying powerful computing techniques [62,63], whereas a study on control of air-cooled chiller condenser fans applies clustering ANNs [64].

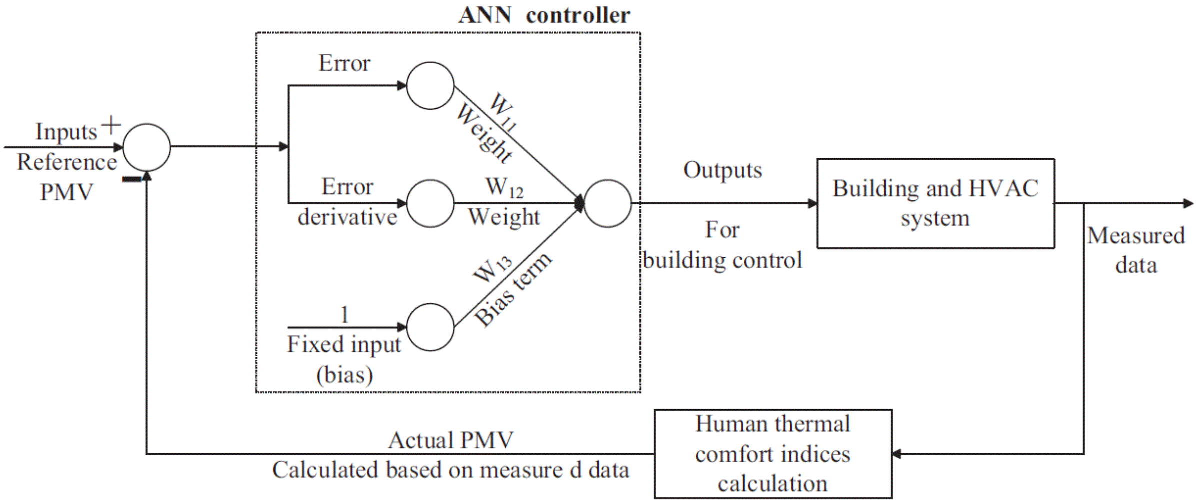

Fig. 4 ANN controller for an HVAC system

Fuzzy logic (FL) control tries to represent human knowledge and reasoning in the form of membership functions and rules to make useful inference actions for the control. The FL control does not require an exact mathematical model of the control process. [15,65]. In many practical HVAC applications, FL fused with ANN shows more effective control [66-68].

4 Applications and overview of publications

Compared to the conventional methods presented in Section 3, the RL technique is still not well developed for building energy control and potential space for exploring the use of RL is large. This can be seen in three ways. First, the shortage of scientific research publications prevents building users, building managers, device controllers, energy agencies, governments and other related parties from being aware of the neglected technique. Despite a number of early studies focusing on controlling HVAC or photovoltaic devices [69-72], an integration with explicit building prototypes has not been comprehensively examined. Second, existing publications (Table 2) primarily try to learn optimal policies for controlling the physical components. The outcome indicating occupant behaviors, e.g. window opening and occupancy rate, may need to be further considered. Third, the curse of dimensionality, the fact that the number of state-action pairs grows exponentially with complex states and actions, is an inherent problem. Approximate solution methods provide the possibility to overcome this. However, deficient consideration of it hinders the development of solutions. Thus, the necessity for investigating current studies and indicating future studies first requires an overview (Table 2).

Unlike conventional control methods, RL does not require a model. A benefit of a model-free approach is that it simplifies the problem when the system is complex. Different from independent and identically distributed (i.i.d.) data that some conventional models need, the RL agent receives subsequent reward signals from its actions. Another benefit is that the trade-off between exploration and exploitation can be balanced. The agent can either exploit past actions or explore unselected actions, which makes it flexible to control the learning process. Furthermore, a rich class of learning algorithms fused with deep neural networks [73] provide a potential for accurate estimation of value functions. Herein, subsections 4.1 and 4.2 illustrate how tabular and approximate methods work on building models. Subsection 4.3 discusses the policy-based algorithm.

Table 2 RL control for building energy

Year	Optimization objectives	States	Actions	Value function	Methods
2007 [74]	Energy consumption, thermal comfort, indoor air quality	Temp.[1], humidity, CO2 concentration	Heat pump setting, air ventilation, window opening	Approx.[2]	$RLS-TD(\lambda)$ with eligibility trace
2018 [75]	Cost	Temp., outside Temp.	Heat extracting, thermal output power	Approx.	Fitted *Q*-iteration
2015 [76]	Energy savings	PCM Temp.	(non-)activation of PCM	Approx.	$SARSA(\lambda)$
2016 [77]	Cost savings, CO_2 savings	PCM Temp.	(non-)activation of PCM	Approx.	$SARSA(\lambda)$
2017 [78]	Energy consumption (cost), thermal comfort	Room Temp., circuits' Temp., time, weather forecast	Temp. set points	Approx.	Fitted *Q*-iteration
2006 [79]	Energy consumption (cost)	Building modes, state of charge	Temp. set-points, TES charging rate	Tabular	*Q*-learning
2006 [80]	Energy consumption (cost)	Building modes, state of charge	Temp. set-points, TES charging rate	Tabular	*Q*-learning
2007 [81]	Energy consumption (cost)	Building modes, state of charge, zone Temp.	Temp. set-points, TES charging rate	Tabular/ Approx.	*Q*-learning
2016 [82]	Human comfort, energy conservation	Comfort status, illuminance, blind status	Turing on/off light, changing blind angel	Tabular	Improved *Q*-learning

[1] Temp. stands for temperature.

[2] Approx. stands for approximation approach of computing the value function.

2016 [83]	Energy consumption	Building energy values	Changes in Smart Grid, building renovation	Tabular	*SARSA* *Q*-learning
2016 [84]	Energy consumption (cost)	Time, indoor Temp., outdoor Temp.	HVAC on/off	Approx.	Fitted *Q*-iteration
2015 [85] (L1)	Net thermal output	Solar irradiation, air Temp.	Mass flow rate	Approx.	Batch *Q*-learning
2015 [85] (L23)	Heat supply, maintaining heat pumps Temp.	Heating demand, month	Mass flow rate	Tabular	*Q*-learning
2008 [86]	Energy consumption, thermal comfort	Temp., CO2 concentration, time	Heat pump, window opening/closing	Approx.	$RLS - TD$
2003 [87]	Energy consumption (cost)	State-of-charge of chiller	TES charging rate	Tabular	*Q*-learning
2010 [88]	Energy cost, discomfort cost	Time, indoor Temp., tank water Temp., Temp. increase	Room Temp. set point, tank water Temp. set point, working mode	Tabular	$Q(\lambda)$ with eligibility trace
2016 [89]	Energy consumption (cost)	Time, Temp. of water heater	On/Off of the heater	Approx.	Fitted *Q*-iteration
2018 [90]	Energy consumption, thermal comfort	Outdoor Temp., solar radiation, wind speed, indoor Temp.	HVAC on/off, window opening/closing	Tabular	*Q*-learning

4.1 Tabular method

In systems with small and discrete state or state-action sets, it is preferable to formulate the estimations using look-up tables with one entry for each state or state-action value. The tabular method is easy to implement and guarantees convergence [25]. A direct application of Eq (11) is the tabular *Q*-learning control algorithm [30] which seems to be the most common of the RL algorithms used in building energy control. In Algorithm 1, the ϵ-greedy policy indicates that the agent chooses an action that has maximal estimated action value with probability $1 - \epsilon$, but with probability ϵ the agent selects an action at random. The update to a new action value is achieved by adding a so called *TD*-error, $\alpha\left[R + \gamma \max_{a'} Q(S', a') - Q(S, A)\right]$, to the old action values. The value function $Q(S, A)$[3] asymptotically converges to $q_*(s, a)$.

[3] We use $Q(S, A)$ to represent an approximate value function from the data and $q(S, A)$ to represent the target of the approximation.

```
Algorithm 1. Tabular Q-learning
Input: discount parameter $\gamma$; step size parameter $\alpha$; $\{s, a\} \in \{\mathcal{S}, \mathcal{A}\}$; $\epsilon > 0$.
    Loop for each episode
    Initialize $Q(s, a)$
        Loop for each time step
            Choose $A$ by $\epsilon$-greedy policy
            Observe the immediate reward $R$ and $S'$
            $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \max_{a'} Q(S', a') - Q(S, A)\right]$
            $S \leftarrow S'$
        Until $S$ is terminal
Output: $Q$-table
```

Studies on thermal energy storage (TES) systems [87], in which a control on the charging rate of chillers used to shift cooling loads from on-peak to off-peak periods, compare load savings. By constraining the exploration, the RL methods are only slightly inferior to model-based methods. Another strategy is to first create correct state-action space by simulation and then apply the learning algorithm [79,80], which significantly reduces the effort of exploration. Similarly, pre-estimation for reducing the state-action space provides a scheme of efficient learning [84]. On the other hand, one does not always need to reach the exact control policy rather than the action patterns [81]. The *Q*-learning extended by adding the eligibility traces (back) [25] leads to efficient computation in that only single vector is required for storing the sum of the cost for the past few hours [88]. The fuzzy state discretization is employed for a reduction of state space. It shows that RL control can deal with inaccurate building models and compensate for incorrect specification of discomfort cost. Benefiting from the development of the computing power, the *Q*-learning policy outperforms over conventional models when occupants' comfort and behaviors are considered [82,90].

4.2 Approximation method

For large MDP problems, we do not always want to separately see the trajectory of each entry of the look-up table. The parameterised value function approximation $\hat{q}(s, a; \mathbf{w}) \approx q_\pi(s, a)$ gives a mapping from the state-action to a function value, for which there are many mapping functions are available, for example, linear combinations, neural networks, etc. It generates the state-actions that we cannot observe. Updating of the weight vector, $\mathbf{w}$, leads to the *incremental method* and the *batch method*.

For the incremental method, $\mathbf{w}$ is updated by gradient descent:

$$\mathbf{w}_{\boldsymbol{t+1}} = \mathbf{w}_{\boldsymbol{t}} + \beta[q_\pi(S_t, A_t) - \hat{q}(S_t, A_t; \mathbf{w}_{\boldsymbol{t}})]\nabla\hat{q}(S_t, A_t; \mathbf{w}_{\boldsymbol{t}}). \tag{12}$$

The learning target $q_\pi(S_t, A_t)$ is iteratively obtained from the Bellman Equation Eq. (11). The $SARSA(\lambda)$ algorithm [25] makes use of the approximation to update the eligibility trace and the value function. In Algorithm 2, the linear case assumes $\hat{q}(s, a; \mathbf{w}) = \mathbf{w}^T\mathbf{x}(s, a)$. The vector, $\boldsymbol{z}$, with the same number of components as $\mathbf{w}$, is the eligibility trace that keeps track of which components of the weight vector have contributed to the recent state values.

```
Algorithm 2. Approximate $SARSA(\lambda)$
Input: discount parameter $\gamma$; trace-decay parameter $\lambda$; step size parameter $\alpha$;
$\{s, a\} \in \{\mathcal{S}, \mathcal{A}\}$; $\epsilon > 0$.
    Loop for each episode
    Initialize $Q_{old}(s, a)$; $\mathbf{z} = \mathbf{0}$; $\mathbf{w} = \mathbf{0}$
        Loop for each time step
            Take $A$; observe the immediate reward $R$ and $S'$
            Choose $A'$ from $S'$ by $\epsilon$-greedy policy
            $\mathbf{x}' \leftarrow \mathbf{x}(S', A')$
            $Q(S, A) \leftarrow \mathbf{w}^T\mathbf{x}$; $Q(S', A') \leftarrow \mathbf{w}^T\mathbf{x}'$
            $\delta \leftarrow R + \gamma Q(S', A') - Q(S, A)$
            $\mathbf{z} \leftarrow \gamma\lambda\mathbf{z} + (1 - \alpha\gamma\lambda\mathbf{z}^T\mathbf{x})\mathbf{x}$
            $\mathbf{w} \leftarrow \mathbf{w} + \alpha[\delta + Q(S, A) - Q_{old}]\mathbf{z} - \alpha[Q(S, A) - Q_{old}]\mathbf{x}$
            $\mathbf{x} \leftarrow \mathbf{x}'$; $A \leftarrow A'$; $Q_{old} \leftarrow Q(S', A')$
        Until $S$ is terminal
Output: $\mathbf{w}^T$
```

Studies on controlling the operational schedule of a ventilated facade with phase change material (PCM) in its air chamber shows a significant electricity saving and robustness to the weather [76]. The learning also works also well in terms of simultaneously reducing CO_2 emissions for the TES [77]. A slightly different algorithm, recursive least-squares temporal-difference (RLS-TD) [91], also uses the eligibility trace but updates the weight vector through a least-squares problem. When it comes to the applications considering both energy consumption and thermal comfort, an early study [74] observed an improved policy and learning ability through a 4-year simulation. This was consolidated by another study [86] in which a fifth year was added. However, the benefit of the incremental control method on building energy is still an attractive issue and hence empirical studies are required.

Whereas the incremental method makes use of the experience once to update the estimate of the value function and then throwing it away before going to the next step, the batch method is sample efficient and tries to find the best fitting to all of the data. It tries to explain all of the reward received at each step. The fitted *Q*-iteration (FQI) algorithm [92] also uses gradient descent optimisation, but for a sample of observations, to update **w** (Algorithm 3 [29]).

```
Algorithm 3. Fitted Q-iteration
Input: discount parameter $\gamma$; mapping rule $q(s, a; \mathbf{w})$; samples $\{s_t, a_t, r_{t+1}, s_t'\}$
for each $\{s, a\} \in \{\mathcal{S}, \mathcal{A}\}$.
    Initialize $\mathbf{w}$ and $Q(s, a; \mathbf{w})$
    Loop at iteration $l = 1,2, \ldots$
        for $l_s = 1,2, \ldots, |\{s, a\}|$ do
            $Q_{l+1,l_s} \leftarrow r_{l_s} + \gamma \max_{a'} Q_l\left(S'_{l_s}, a'; \mathbf{w}_l\right)$
        End for
        Update $\mathbf{w}$ by fitting $Q_{l+1,l_s}$
    Until the change of $Q_{l+1,l_s}$ is trivial
Output: $\hat{q}(s, a; \mathbf{w}) = Q_{l+1,l_s}$
```

The FQI algorithm on building energy control has just been recently noticed and efforts on improving the learning quality have only been examined in limited studies. A study of thermostatically controlled loads connected to a district heating network in a cluster uses FQI to obtain the optimal actions for the entire cluster [75]. The policy is then combined with a market-based multi-agent system for minimising the cost of the loads. Both the energy arbitrage and peak saving scenarios show promising control strategies after 40-60 days' learning. As a strategy for improving the learning efficiency, combining model-free methods with good domain knowledge can be further considered. The implementation of FQI combined with an auto-encoder network has also been shown to be feasible in reconstructing the states [89]. For a single building, the heating system can be optimally controlled by adding virtual support tuples for those states with a low-density experiment sample [84].

4.3 Policy-based method

Both the tabular and approximation method work in value-based paradigm where the value functions have to be approximated and the policy is taken by greedy or ϵ-greedy strategies, whereas the policy-based method [93] directly searches for the parametrised policy:

$$\pi_\theta(a|s;\boldsymbol{\theta}) = \mathbb{P}[A_t = a|S_t = s;\boldsymbol{\theta}_t = \boldsymbol{\theta}]. \tag{13}$$

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 updates $\boldsymbol{\theta}$ for the optimisation. The action preference is usually assigned to a probability to avoid the deterministic policy. For the control of building energy, the application of policy-based methods has not yet been empirically made. Furthermore, a combination of the value-based method and policy-based method, e.g. the Actor-Critic algorithm [94], is also appealing.

5 Multi-agent reinforcement learning

As an agent-based technology, the multi-agent system (MAS) provides promising paradigms in artificial intelligence [95-97]. The decomposition of complex systems facilitates each agent to share a common environment and work independently on a specific sub-problem. MAS extends the single agent system by allowing each agent to interact with other agents – not simply by exchanging data, but also by engaging in analogues of social activity: cooperation and negotiation [98].

5.1 Multi-agent control system

When it comes to a multi-agent control system (MACS) in a building, each agent implements autonomous actions in order to optimally run building models in a dynamic system. In most of the studies, the hierarchical central-local agent structure is embedded in the building models to balance the energy consumption and the occupants' comfort [99-102]. A (multi-object) particle swarm optimisation (PSO) technique is utilized for optimising intelligent management. In Fig. 7 [99], for instance, a central agent communicates with building managers and zone agent to decide the optimal power distribution for each zone by considering the comfort demand. A zone agent, on the other hand, communicates with occupants and decides power demand. The local agents take care of temperature control, illumination control, and air quality control. The objective function aims at maximising total thermal comfort and minimising energy consumption. Similar central-local systems utilise FL controllers for maximising the thermal comfort [103]. An occupant-driven control is studied in MASs including HVAC agents, occupant agents and meeting agents [104], where the MDP based coordination tries to find the optimal policy by considering energy consumption, occupant comfort, and scheduling convenience individually.

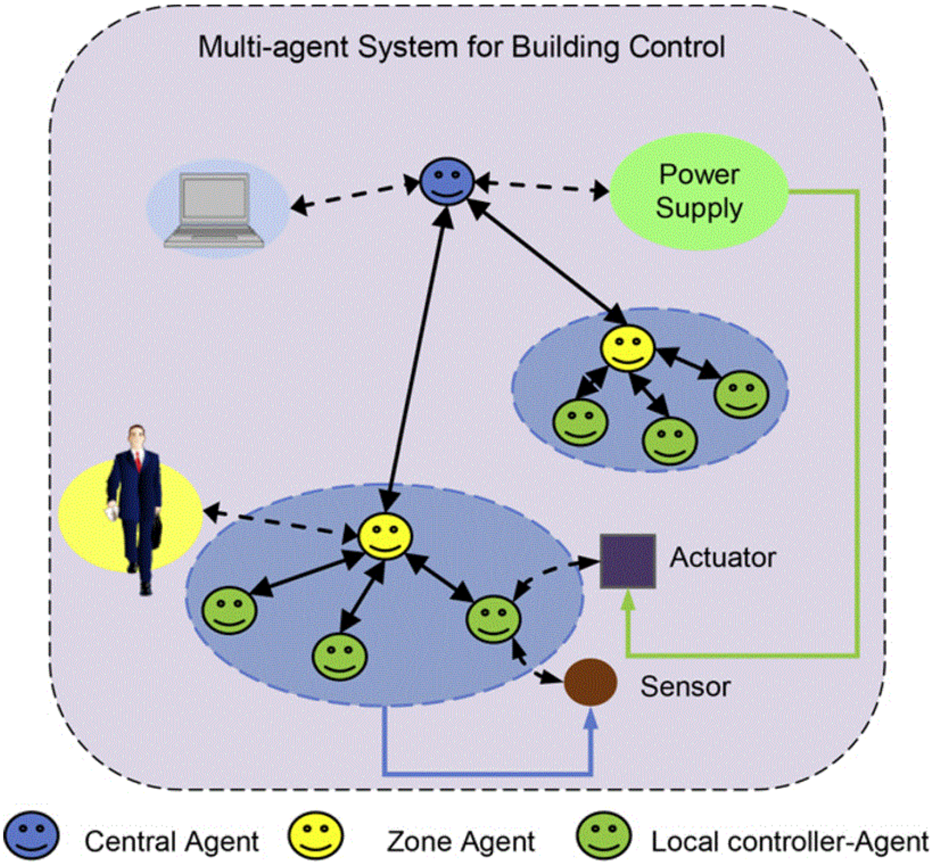

Fig. 7 MAS for building energy control

5.2 Multi-agent reinforcement learning

The MDP property for multi-agent reinforcement learning (MARL) has been extensively studied in matrix game playing [105-107] since both cooperative and competitive as well as a mixed environment can be modeled and simulated. Previous survey works [108,109] have summarised and explored MARL theory, algorithms and applications. The benefit of MARL comes from experience sharing, information exchange, and skill learning among agents. When one or more agents fail to work in the system, the remaining agents are still able to react optimally by learning from the new environment. They also pointed out that MARL faces challenges [108-110]. Firstly, the agents might work independently instead of cooperatively resulting in sub-optimal returns. Secondly, the decision-making process for a single agent in MARL is non-stationary due to the dynamic policy changes from other agents. Thirdly, the exploration strategy becomes complex when the number of agents increases. Despite those challenges, making assumptions assures that MARL algorithms perform efficiently.

Generally, the transition probability for MARL extends Eq. (1) to a multi-action case:

$$p(\boldsymbol{s}', r_i|\boldsymbol{s}, a_i) = \mathbb{P}[\boldsymbol{S_t} = \boldsymbol{s}', \boldsymbol{R}_{t,i} = \{r_1, \dots, r_n\}|\ \boldsymbol{S_{t-1}} = \boldsymbol{s}, \boldsymbol{A}_{t-1,i} = \{a_1, \dots, a_n\}], \qquad (14)$$

where n is the total number of agents, r_i is the reward for agent i and $\boldsymbol{s} = \{s_1, \dots s_n\}$ is the set of individual states. In Eq. (14), the stochastic transition is a probability distribution over the next vector of states, $\boldsymbol{s}'$, given the current vector of states, $\boldsymbol{s}$, and joint action, a_i. For the policy $\pi_i \in \Pi_i$, the optimal policy π_i^* for agent i fulfills the Nash equilibrium:

$$\begin{aligned}&\sum_{a_1,\dots,a_n} q_*(\boldsymbol{s}, a_i)\, \pi_1^*(a_1|\boldsymbol{s}) \cdot \dots \cdot \pi_i^*(a_i|\boldsymbol{s}) \cdot \dots \cdot \pi_n^*(a_n|\boldsymbol{s}) \\ \geq &\sum_{a_1,\dots,a_n} q_*(\boldsymbol{s}, a_i)\, \pi_1^*(a_1|\boldsymbol{s}) \cdot \dots \cdot \pi_i(a_i|\boldsymbol{s}) \cdot \dots \cdot \pi_n^*(a_n|\boldsymbol{s}),\end{aligned} \qquad (15)$$

where $q_*(\boldsymbol{s}, a_i)$ is the optimal action-value function for agent i and $\pi_i^*(a_i|\boldsymbol{s})$ is the individual probability of taking action a_i given the Nash equilibrium policy.

Similar to game players, the MARL control for building energy can be either in a cooperative or non-cooperative system. Hurtado Munoz et al [111] compare the extended joint action learning (JAL) to a decentralized non-cooperative *Q*-learning method and a centralized Nash *n*-player game method to study the energy demand flexibility of a cluster of buildings. The MARL methods show that a range of flexibility requests can be met by providing an optimal energy portfolio of buildings and the proposed extended JAL performs best, considering responsibility and the commitment of allocation values [112]. Bollinger and Evins [113] illustrate the potential of MARL optimisation methods used for distributed multi-energy systems with four agents. The *Q*-learning and continuous Actor-Critic learning automaton algorithms have been tested. As shown in Fig. 8 [113], building agents generate offers and/or requests for electricity and/or heat for each hour within a 24-hour period. The market agent analyses the interactions between buyers and sellers and tries to match the optimal buyers and sellers. The states consist of time of day and state of charge and the actions consist of the prices and quantities of electricity and heat. The sum of electricity and heat profit form the reward signal. The authors conclude that the MARL shows promise in addressing specific classes of energy hub problems, especially those dealing with large scales. The integration of human knowledge and equipment sensors applied in collaborative multi-agent buildings for optimising hot water production can also outperform single agent systems [114].

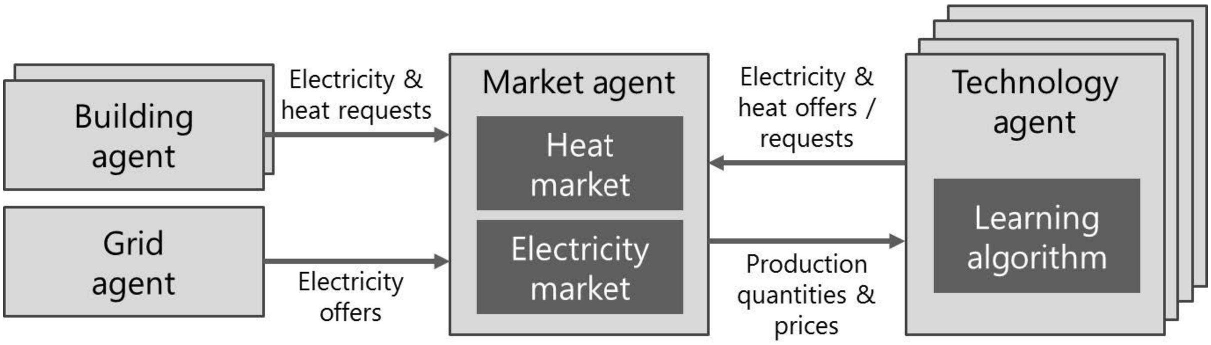

Fig. 8 MARL for building energy control

Even though paradigms of implementing MARL for building energy controls have incomprehensively arisen, all the empirical studies have tried to outline the potentials for developing various methods. Unlike a single agent system, the MAS may require agents being aware of others' actions. Therefore, more efforts may be devoted on how to convey behaviors among agents in cooperative systems.

6 Conclusions

In this paper, we investigated the reinforcement learning control method on building energy. Efforts made on building automation are crucial for making buildings smarter in terms of cutting costs by streamlining building operations like air conditioning and lighting [115]. As a data-driven method, model-free RL can work in a simulation paradigm [116] and has brought broad novelties into artificial intelligence applications. However, it seems that cutting-edge RL techniques have drawn only a few attentions regarding smart building automation systems.

We briefly reviewed the current applications on building energies according to the method of handling the value functions for both value iteration and policy iteration algorithms. It is surprising that both VI and PI dominate the algorithms, which leaves open the question of how the third category, namely, the policy-based search, or a combination of them perform when it is applied to building energy control. Various empirical studies are expected for checking this in the future.

Compared to the tabular method, the development of the approximation method provides the possibility of managing the curse of dimensionality. Algorithms like FIQ and RLS have set feasible paradigms to estimating value functions by linear models. Since linear models guarantee convergence, the implementation becomes meaningful for suitable feature construction. For non-linear approximation approaches like ANN, deep ANN and deep convolution networks, proper design and training of the networks can contribute a great deal to RL [25]. Impressive applications make building energy control attractive. For instance, the image of temperature scale or humidity is a suitable case to serve as the input state to the convolution network.

Optimal policies of multi-agent reinforcement learning are usually defined in the context of game theory and the cooperative case is expected for maximising the joint reward. However, many challenges may intervene the cooperative agents and lead to sub-optimal policies. One crucial future task is to learn optimal policies for buildings in a dynamic and independent environment.

The promising RL method opens a new way of efficiently controlling HVAC systems, water heaters, lighting and window opening. As a part of smart buildings, energy-related systems have great potential for being updated by experience. More learning experiences can assist the agent in faster exploring the desired solution space and reducing the trial-and-error attempts.

Acknowledgment

The authors are thankful for the financial support from the Micro-data analysis research profile at Dalarna University, and the Swedish Energy Agency.

References

[1] Fanti MP, Mangini AM, Roccotelli M. A simulation and control model for building energy management. Cont Engi Prac 2018; 72:192-205.

[2] Pérez-Lombard L, Ortiz J, Pout C. A review on buildings energy consumption in formation. Energy and Buildings 2008; 40:394-98.

[3] Cai WG, Zhong YW, Ren H. China building energy consumption: Situation, challenges and corresponding measures. Energy Policy 2009; 37:2045-59.

[4] Xu P, Chan EH, Qian QK. Success factors of energy performance contracting (EPC) for sustainable building energy efficiency retrofit (BEER) of hotel buildings in China. Energy Policy 2011; 39(11):7389-98.

[5] Nejat P, Jomehzadeh F, Taheri MM, Gohari M, Majid MZA. A global review of energy consumption, CO_2 emissions and policy in the residential sector (with an overview of the top ten CO2 emitting countries). Renew Sustain Energy Rev 2015; 43:843–62.

[6] Wang N, Phelan PE, Harris C, Langevin J, Nelson B, Sawyer K. Past visions, current trends, and future context: A review of building energy, carbon, and sustainability. Renew Sustain Energy Rev 2018; 82(1):976-93.

[7] Wang W, Zmeureanu R, Rivard H. Applying multi-objective genetic algorithms in green building design optimization. Buil and Enri 2005; 40(11):1512-25.

[8] Shaikh PH, Nor NBM, Nallagownden P, Elamvazuthi I, Ibrahim T. A review on optimized control systems for building energy and comfort management of smart sustainable buildings. Renew Sustain Energy Rev 2014; 34 409–29.

[9] Zhao P, Suryanarayanan S, Simoes MG. An Energy Management System for Building Structures Using a Multi-Agent Decision-Making Control Methodology. IEEE Tran on Indu Appl 2013; 49(1): 322-30.

[10] Yu Z, Huang G, Haghighat F, Li H, Zhang G. Control strategies for integration of thermal energy storage into buildings: State-of-the-art review. Energy and Buildings 2015; 106:203–15.

[11] Dounis AI, Caraiscos C. Advanced control systems engineering for energy and comfort management in a building environment-A review. Renew Sustain Energy Rev 2009; 13:1246–61.

[12] Aste N, Manfren M, Marenzi G. Building Automation and Control Systems and performance optimization: A framework for analysis. Renew Sustain Energy Rev 2017; 75:313–30.

[13] Afram A, Janabi-Sharifi F. Theory and applications of HVAC control systems – A review of model predictive control (MPC). Building and Environment 2014; 72:343–55.

[14] Naidu DS, Rieger CG. Advanced control strategies for heating, ventilation, air-conditioning, and refrigeration systems-an overview: Part I: Hard control. HVAC&R Research 2011; 17(1):2–21.

[15] Naidu DS, Rieger CG. Advanced control strategies for HVAC&R systems-an overview: Part II: Soft and fusion control. HVAC&R Research 2011; 17(2):144–58.

[16] Kaelbling LP, Littman ML, More AW. Reinforcement Learning: A Survey. Jour of Arti Inte Rese 1996; 4:237-85.

[17] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015; 518:529-33.

[18] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Driessche GVD. Mastering the game of Go with deep neural networks and tree search. Nature 2016; 529:484-89.

[19] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A. Mastering the game of Go without human knowledge. Nature 2017; 550:354-59.

[20] Riedmiller M, Gabel T, Hafner R, Lange S. Reinforcement learning for robot soccer. Auton Robot 2009; 27:55-73.

[21] Claus C, Boutilier C. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. National Conference on Artificial Intelligence Proceedings 1998; 746-52.

[22] Gu S, Lillicrap T, Sutskever I, Levine S. Continuous Deep Q-Learning with Model-based Acceleration. Proceeding of Conference on Machine Learning 2016; JMLR: W&CP 48.

[23] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Yuval, et al. Continuous control with deep reinforcement learning. International Conference on Learning Representations 2016; arXiv:1509.02971v5 [cs.LG].

[24] Sutton RS, Barto AG. Reinforcement Learning: An introduction. 1st ed. MIT Press, Cambridge, MA; 1998.

[25] Sutton RS, Barto AG. Reinforcement Learning: An introduction. 2nd ed. MITPress, Cambridge, MA: E-Publishing; 2018.

[26] Bellman RE. A Markov decision process. Journal of Mathematics and Mechanics 1957; 6(5):679–84.

[27] Bellman RE. Dynamic Programming. Princeton University Press, Princeton; 1957.

[28] Silver D. UCL courses on RL, http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html; 2015 [accessed May 2018].

[29] Buşoniu L, Babuška R, Schutter BD, Ernst D. Reinforcement learning and dynamic programming using function approximators. CRC Press, Boca Raton, FL; 2010.

[30] Watkins CJCH. Learning from Delayed Rewards. Ph.D. thesis, University of Cambridge; 1989.

[31] Rummery GA, Niranjan M. On-line Q-learning using connectionist systems. Cambridge University, Cambrige; 1994.

[32] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning 1992; 8(3-4):229-56.

[33] Deisenroth MP, Rasmussen CE. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In Proceedings of the 28th International Conference on machine learning (ICML11) 2011; 465–72.

[34] Royapoor M, Antony A, Roskilly T, A review of building climate and plant controls, and a survey of industry perspectives. Energy and Buildings 2018; 158:453-65.

[35] Li X, Wen J. Review of building energy modeling for control and operation. Renew Sustain Energy Rev 2014; 37:517-37.

[36] Zhang J, Zhang K. A Particle Swarm Optimization Approach for Optimal Design of PID Controller for Temperature Control in HVAC. IEEE Int Conf on Meas Tech and Mech Auto (ICMTMA) 2011; 230-33.

[37] Bai J, Zhang X. A new adaptive PI controller and its application in HVAC systems. Energy Conv Mana 2007; 48:1043-54.

[38] Wang YG, Shi ZG, Cai WJ, PID autotuner and its application in HVAC systems. IEEE American Control Conf 2001; 2192-6.

[39] Bianchi FD, Mantz RJ, Christiansen CF. Gain scheduling control of variable-speed wind energy conversion systems using quasi-LPV models. Control Engineering Practice 2005; 13(2):247–55.

[40] Moradi H, Saffar-Avval M, Bakhtiari-Nejad F. Nonlinear multivariable control and performance analysis of an air-handling unit. Energy and Buildings 2012; 43:805–13.

[41] Isidori A. Nonlinear Control Systems. 3rd ed. London: Springer; 2000.

[42] Zaheer-Uddin, M. Optimal, sub-optimal and adaptive control methods for the design of temperature controllers for intelligent building. Building and Environment 1993; 28(3):311–22.

[43] Vinter R. Optimal Control. Springer, Birkhauser, Boston; 2010.

[44] Branu JE. Reducing energy costs and peak electrical demand through optimal control of building thermal storage. ASHRAE transactions 1990; 96(2):876–88.

[45] Henze GP, Dodier RH, Krarti M. Development of a Predictive Optimal Controller for Thermal Energy Storage Systems. HVAC&R Research 1997; 3(3):233–64.

[46] Dong B. Non-linear optimal controller design for building HVAC systems. Control Applications (CCA), IEEE International Conference 2010; 201-5.

[47] Garcia CE, Prett DM, Morari M, Model predictive control: theory and practice-a survey. Automatica 1989; 25(3):335-48.

[48] Afram A, Janabi-Sharifi F, Fung AS, Raahemifar K. Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system. Energy and Buildings 2017; 141:96-13.

[49] Ma J, Qin J, Salsbury T, Xu P. Demand reduction in building energy systems based on economic model predictive control. Chemi Eng Science 2012; 67(1):92-100.

[50] Cole WJ, Powell KM, Hale ET, Edgar TF. Reduced-order residential home modeling for model predictive control. Energy and Buildings 2014; 74:69-7.

[51] Muhammad A, Chen C, Chau C-K, Talal R. Automatic HVAC control with real-time occupancy recognition and simulation-guided model predictive control in low-cost embedded system. Energy and Buildings 2017; 154:141-56.

[52] Ascione F, Bianco N, Stasio CD, Mauro GM, Vanoli GP. A new comprehensive approach for cost-optimal building design integrated with the multi-objective model predictive control of HVAC systems. Sust. Cities and Soci.2017; 31:136-50.

[53] Zhou K, Doyle JC. Essentials of robust control. Upper Saddle River, NJ: Prentice Hall; 1998.

[54] Anderson M, Buehner M, Young P, Hottle D, Anderson C, Tu J, et al. MIMO Robust Control for HVAC Systems. IEEE Transactions on Control Systems Tech. 2008; 16(3):475-83.

[55] Underwood CP. Robust control of HVAC plant II: controller design. Proceeding of CIBSEA, Building Serv Eng Res Tech 2000; 21(1):63-71.

[56] Moradi H, Saffar-Avval M, Bakhtiari-Nejad F. Nonlinear multivariable control and performance analysis of an air-handling unit. Energy and Buildings 2011; 43:805-13.

[57] Thosar A, Patra A, Bhattacharyya S. Feedback linearization based control of a variable air volume air conditioning system for cooling applications. ISA Transactions 2008; 47(3):339-49.

[58] Åström KJ, Wittenmark B. Adaptive Control. 2nd ed. Dover Pub INC., Mineola NY; 2008.

[59] Huang WZ, Zaheeruddin M, Cho SH. Dynamic simulation of energy management control functions for HVAC systems in buildings. Energy Conv and Manag 2006; 47(7-8):926-43.

[60] Bai J, Li Y, Chen J. Real-time performance assessment and adaptive control for a water chiller unit in an HVAC system. Earth and Environmental Science 2018; 121:1-7.

[61] Liang J, Du R. Thermal comfort control based on neural network for HVAC application. Proc of IEEE Conf on Contr Appl 2005; 819-24.

[62] Javed A, Larijani H, Ahmadinia A, Gibson D. Smart Random Neural Network Controller for HVAC using Cloud Computing Technology. IEEE Trans in Indus Inform 2017; 13:351-60.

[63] Javed A, Larijani H, Ahmadinia A, Emmanuel R, Mannion M, Gibson D. Design and implementation of a cloud enabled random neural network-based decentralized smart controller with intelligent sensor nodes for HVAC. IEEE Internet of Things Journal 2017; 4:393-03.

[64] Henze GP, Hindman RE, Control of air-cooled chiller condenser fans using clustering neural networks. ASHRAE Tran 2002; 108:232-44.

[65] Shepherd AB, Batty WJ. Fuzzy control strategies to provide cost and energy efficient high quality indoor environments in buildings with high occupant densities. Build Serv Eng Res and Tech 2003; 24(1):35-45.

[66] Ahn J, Cho S, Chung DH. Analysis of energy and control efficiencies of fuzzy logic and artificial neural network technologies in the heating energy supply system responding to the changes of user demands. Applied energy 2017; 190:222-31.

[67] Işik E, Inalli M. Artificial neural networks and adaptive neuro-fuzzy inference systems approaches to forecast the meteorological data for HVAC: The case of cities for Turkey. Energy 2018; 154:7-16.

[68] Marvuglia A, Messineo A, Nicolosi G. Coupling a neural network temperature predictor and a fuzzy logic controller to perform thermal comfort regulation in an office building. Buil and Envi 2014; 72:287-99.

[69] Drees KH, Braun JE. Development and Evaluation of a Rule-Based Control Strategy for Ice Storage Systems. HVAC&R Research 1996; 2(4):312-36.

[70] Anderson CW, Hittle DC, Katz AD, Kretchmar RM. Synthesis of reinforcement learning, neural networks and PI control applied to a simulated heating coil. Artificial Intelligence in Engineering 1997; 11:421-29.

[71] Anderson CW, Hittle DC, Kretchmar RM, Young P. Robust reinforcement learning for heating, ventilation and air conditioning control of building. In Handbook of learning and approximate dynamic programming IEEE Press, NJ US 2004; 517-34.

[72] Henze GP, Dodier RH. Adaptive optimal control of a grid-independent photovoltaic system. International solar energy conference 2002; 139-48.

[73] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with Deep Reinforcement Learning. NIPS Deep Learning Workshop 2013; arXiv:1312.5602 [cs.LG].

[74] Dalamagkidis K, Kolokotsa D, Kalaitzakis K, Stavrakakis GS. Reinforcement learning for energy conservation and comfort in buildings 2007; 42:2686-98.

[75] Claessens BJ, Vanhoudt D, Desmedt J, Ruelens F, Model-free control of thermostatically controlled loads connected to a district heating network. Energy and Buildings 2018; 159:1-10.

[76] Gracia AD, Fernández C, Castell A, Mateu C, Cabeza LF. Control of a PCM ventilated facade using reinforcement learning techniques. Energy and Buildings 2015; 106:234-42.

[77] Gracia AD, Barzin R, Fernández C, Farid MM, Cabeza LF. Control strategies comparison of a ventilated facade with PCM –energy savings, cost reduction and CO2 mitigation. Energy and Buildings 2016; 130:821-8.

[78] Schmidt M, Moreno MV, Schülke A, Macek K, Mařík K, Pastor AG. Optimizing legacy building operation: The evolution into data-driven predictive cyber-physical systems. Energy and Buildings 2017; 18:257-79.

[79] Liu S, Henze GP. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory Part 1. Theoretical foundation. Energy and Buildings 2006; 38:142-7.

[80] Liu S, Henze GP. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory Part 2. Results and analysis. Energy and Buildings 2006; 38:148-61.

[81] Liu S, Henze GP. Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory. Journal of Sol Energy Eng 2007; 129(2):215-25.

[82] Cheng Z, Zhao Q, Wang F, Jiang Y, Xia L, Ding J. Satisfaction based Q-learning for integrated lighting and blind control. Energy and Buildings 2016; 127:43-55.

[83] Mocanu E, Nguyen PH, Kling WL, Gibescu M. Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building transfer learning. Energy and Buildings 2016; 116:646-55.

[84] Costanzo GT, Iacovella S, Ruelens F, Leurs T, Claessens BJ. Experimental analysis of data-driven control for a building heating system. Sust. Energy. Grids and Net. 2016; 6:81-90.

[85] Yang L, Nagy Z, Goffin P, Schlueter A. Reinforcement learning for optimal control of low exergy buildings. Applied Energy 2015; 156:577-86.

[86] Dalamagkidis K, Kolokotsa D. Reinforcement Learning for Building Environmental Control, Reinforcement Learning Cornelius Weber 2008 (Chapter 15), IntechOpen; 283-94.

[87] Henze GP, Schoenmann J. Evaluation of Reinforcement Learning Control for Thermal Energy Storage Systems. HVAC&R Research 2003; 9(3):259-75.

[88] Yu Z, Dexter A. Online tuning of a supervisory fuzzy controller for low-energy building system using reinforcement learning. Contr Engi Prac 2010; 18:532-39.

[89] Ruelens F, Claessens BJ, Quaiyum S, Schutter BD, Babuška R, Belmans R. Reinforcement Learning Applied to an Electric Water Heater: From Theory to Practice. IEEE Transactions on Smart Grid 2016.

[90] Chen Y, Norford LK, Samuelson HW, Malkawi A. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning. Energy and Buildings 2018; 168:195-05.

[91] Xu X, He H, Hu D. Efficient reinforcement learning using recursive least-squares methods. Journal of Artificial Research 2002; 16:259-92.

[92] Ernst D, Geurts P, Wehemkel L. Tree-Based Batch Mode Reinforcement Learning. Journal of Machine Learning Research 2005; 6:503-56.

[93] Sutton RS, McAllester DA, Singh SP, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems (NIPS) 1999; 12:1057-63.

[94] Grondman I, Busoniu L, Lopes GAD, Babuska R. A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients. IEEE Trans on Syst, Man and Cyber, Part C 2012; 42(6):1291-307.

[95] Stone P, Veloso M. Multiagent systems: a survey from machine learning perspective. Autonomous Robots 2000; 8(3):345-83.

[96] Sycara KP. Multiagent systems. AI Magazine 1998; 19:79-92.

[97] Ye D, Zhang M, Vasilakos AV. A Survey of Self-Organization Mechanisms in Multiagent Systems. IEEE Trans on Syst, Man and Cyber: Syst 2017; 47(3):441-61.

[98] Wooldridge M. An introduction to multiagent systems. 2nd ed. John Wiley and Sons Ltd, Chichester, UK; 2009.

[99] Yang R, Wang L. Multi-zone building energy management using intelligent control and optimization. Sust Cities Soci 2013; 6:16-21.

[100] Wang Z, Wang L. Occupancy pattern based intelligent control for improving energy efficiency in buildings. IEEE Inte Conf on Auto Scie and Engi 2012; 804-09.

[101] Wang Z, Yang R, Wang L. Multi-agent Control System with Intelligent Optimization for Smart and Energy-efficient Buildings. 36th annual conference on IEEE Indu Elec Soci 2010; 1144-49.

[102] Yang R, Wang L. Multi-objective optimization for decision-making of energy and comfort management in building automation and control. Sust Cities and Soci 2012; 2(1):1-7.

[103] Wang Z, Yang R, Wang L. Multi-agent intelligent controller design for smart and sustainable buildings. Annu IEEE Syst Conf 2010; 277-82.

[104] Klein L, Kwak JY, Kavulya G, Jazizadeh F, Becerik-Gerber B, Varakantham P, et al. Coordinating occupant behavior for building energy and comfort management using multi-agent systems. Auto in Cons 2012; 22:526-36.

[105] Littman ML. Markov games as a framework for multi-agent reinforcement learning. Conference on Machine Learning, New Brunswick 1994; 157-63.

[106] Bowling M, Veloso M. Multiagent learning using a variable learning rate. Arti Itne 2002; 136(2):215-5.

[107] Schwartz HM. Multi-agent machine learning. John Wiley and Sons Inc, New Jersey; 2014.

[108] Buşoniu L, Babuška R, Schutter BD. A comprehensive survey of multiagent reinforcement learning. IEEE Tran on Syst, Man and Cybe-Part C: Appl and Revi 2008; 38:156-72.

[109] Buşoniu L, Babuška R, Schutter BD. Multi-agent Reinforcement Learning: An Overview. Inno in Multi-Agent Syst and Appl, Springer 2010; 183-221.

[110] Ye D, Zhang M, Vasilakos A. A Survey of Self-Organization Mechanisms in Multiagent Systems. IEEE Tran on Syst, Man and Cybe: Syst 2017; 47(3):441-61.

[111] Hurtado Munoz LA, Mocanu E, Nguyen HP, Gibescu M, Kamphuis IG. Enabling cooperative behavior for building demand response based on extended joint action learning. IEEE Transactions on Industrial Informatics 2018; 14(1):127-36.

[112] Jain R, Chiu DM, Hawe W. A Quantitative Measure of Fairness and Discrimination For Resource Allocation in Shared Conputer Systems. Technical Report TR-301, DEC Research Report 1984.

[113] Bollinger LA, Evins R. Multi-agent reinforcement learning for optimizing technology deployment in distributed multi-energy systems. In: 23rd International Workshop of the European Group for Intelligent Computing in Engineering. Krakow 2016.

[114] Kazmi H, Suykens J, Driesen J. Valuing knowledge, information and agency in Multi-agent Reinforcement Learning: a case study in smart buildings. Proceeding of AAMAS 2018; arXiv:1803.03491v1 [cs.MA]

[115] Snoonian D. Smart buildings. IEEE Spectrum 2003; 40(8):18-23.

[116] Gosavi A, Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning, ser. Operations Research/Computer Science Interfaces. New York: Springer; 2003.