Evaluation of Reinforcement Learning for Optimal Control of Building Active and Passive Thermal Storage Inventory

Simeng Liu
Bes-Tech Inc.,
3910 South Interstate Highway 35,
Suite 225,
Austin, TX 78704
e-mail: sliu@bes-tech.net

Gregor P. Henze
Department of Architectural Engineering,
University of Nebraska-Lincoln,
Omaha, NE 68182
e-mail: ghenze@unl.edu

This paper describes an investigation of machine learning for supervisory control of active and passive thermal storage capacity in buildings. Previous studies show that the utilization of active or passive thermal storage, or both, can yield significant peak cooling load reduction and associated electrical demand and operational cost savings. In this study, a model-free learning control is investigated for the operation of electrically driven chilled water systems in heavy-mass commercial buildings. The reinforcement learning controller learns to operate the building and cooling plant based on the reinforcement feedback (monetary cost of each action, in this study) it receives for past control actions. The learning agent interacts with its environment by commanding the global zone temperature setpoints and thermal energy storage charging/discharging rate. The controller extracts information about the environment based solely on the reinforcement signal; the controller does not contain a predictive or system model. Over time and by exploring the environment, the reinforcement learning controller establishes a statistical summary of plant operation, which is continuously updated as operation continues. The present analysis shows that learning control is a feasible methodology to find a near-optimal control strategy for exploiting the active and passive building thermal storage capacity, and also shows that the learning performance is affected by the dimensionality of the action and state space, the learning rate and several other factors. It is found that it takes a long time to learn control strategies for tasks associated with large state and action spaces. [DOI: 10.1115/1.2710491]

Introduction

The advantage of shifting building cooling load by using active and passive thermal storage capacity has been known for decades. Properly utilizing thermal storage may offer substantial cost savings and potentially increases plant efficiency. By definition, *active* building thermal capacity refers to the thermal energy storage (TES) system, which is either a chilled water or ice based system; the *passive* building thermal storage capacity refers to the building envelope, internal construction, and furniture, which affect the building cooling load. The motivation for the current study stems from a research project to investigate predictive optimal control of active and passive building thermal storage inventory. In that project, a model-based predictive optimal control approach was developed in order to evaluate the merits of controlling active and passive thermal storage optimally in a continuous closed-loop fashion. Both numerical analysis and experimentation demonstrated the optimal controller can operate the building and cooling plant efficiently and significant cost saving may be achieved. However, modeling complexity and inaccuracy of this model-based approach were revealed as well, as the model-based predictive optimal control approach is strongly affected by the quality of the model. In reality, there are always deviations between the model and the actual building. The efforts to develop and maintain the model can be very demanding and time consuming, possibly unaffordable. Meanwhile, numerous applications of reinforcement learning control to engineering problems provide a new direction to tackle this control problem in a potentially more efficient manner.

To this end, a model-free learning control is investigated for the operation of electrically driven chilled water systems in heavy-mass commercial buildings. In this paper, we first give a brief introduction to reinforcement learning, followed by the methodology to implement the learning controller. Finally, selected results are presented to evaluate the performance of the learning approach compared with the model-based optimization approach and other conventional control strategies.

Review of Past Work

Previous studies on building thermal mass utilization demonstrate the potential of reduction of peak cooling load and associated electrical demand, and the results show that cost savings vary widely among published case studies [1–4]. In a simulation study presented by Braun [5], cost savings for a design day varied from zero to 35% depending on system type and utility rate. Anderson and Brandemuehl [6] demonstrated energy and cost savings potential by precooling the building structure, calling attention to the importance of the mass of furnishing which significantly affects the precooling strategy. Braun et al. [7] developed a tool to evaluate different precooling strategies by comparing the heating, ventilation, and air conditioning (HVAC) utility costs in each application. Simulation studies were carried out for selected locations, climates, and utility rate structures. A comparison showed cost savings varying from 40% at best to zero or even excess costs for some less favorable cases. In a review article on load control using building thermal mass, Braun [8] concluded that the savings potential is very sensitive to the utility rates, building and plant characteristics, and weather conditions and occupancy schedule. The greatest cost savings were realized for the case of heavy

Contributed by the Solar Energy Engineering Division of ASME for publication in the JOURNAL OF SOLAR ENERGY ENGINEERING. Manuscript received May 22, 2005; final manuscript received October 31, 2006. Review conducted by Moncef Krarti. Paper presented at the 2005 International Solar Energy Conference (ISEC2005), Orlando, FL, USA, August 6–12, 2005.

construction, good part-load characteristics and low ambient temperature which enabled free cooling during night ventilation.

Optimal control of TES has been investigated by several researchers. To evaluate the theoretical potential of ice storage systems in reducing operating cost, a detailed analysis was performed by Henze et al. [9,10] using a simulation environment based on dynamic programming. Within this environment, three conventional control strategies were compared to optimal control to determine how well conventional controls harness the system's cost saving potential. The effect of uncertainty in external variables, such as weather variables and cooling loads, on the performance of optimal control was studied in companion papers by Henze et al. [11] and Henze and Krarti [12].

The combined use of both active and passive storage media under optimal control has been investigated by Kintner-Meyer and Emery for a 24 h deterministic simulation study which revealed that significant operating cost savings (18%) and electrical demand reduction can be achieved [13]. Optimal building control proved most effective in dry climates with large diurnal temperature swings, in the presence of utility rates strongly encouraging load shifting, and when cool storage systems allow more effective load shifting than building precooling alone.

The project "Predictive optimal control of active and passive building thermal storage inventory" sponsored by the U.S. Department of Energy attempts to combine these merits together, and mathematically analyze the optimization simultaneously. A simulation study was carried out to investigate the combined usage of both active and passive building thermal storage inventory by Henze et al. [14]. The analysis showed that when an optimal controller for combined utilization is given perfect weather forecasts and when the building model used in the model-based predictive control perfectly matches the actual building, the utility cost savings are significantly greater than for either passive or active storage alone (although less than the sum of savings from passive and active storage alone), and the on-peak electrical demand for cooling can be drastically reduced. Further research by Henze et al. [15] also demonstrates that prediction uncertainty in the short-term weather forecasts can affect the controller's cost saving performance. Liu and Henze [16] investigated the impact of five categories of building modeling mismatch on the performance of model-based predictive optimal control of combined thermal storage using perfect prediction. The results showed that a simplification or mismatch of the building geometry and zoning only marginally affect the optimization strategy. However, the mismatch of internal heat gain, building construction and energy system efficiency can lead to the significant deviation in optimization. Henze et al. [17] demonstrates model-based predictive optimal control of active and passive building thermal storage inventory in a test facility in real-time using time-of-use differentiated electricity prices without demand charges. The experiment essentially confirms the previous findings in the numerical analysis of optimal control of building thermal storage. However, the savings associated with passive building thermal storage inventory proved to be small because the test facility is not an ideal candidate for the investigated control technology. Moreover, the facility's central plant revealed the idiosyncratic behavior that the chiller operation in the ice-making mode was more energy efficient than in the chilled-water mode.

While the analysis of model-based optimization of active and passive building thermal storage demonstrated significant cost saving potential, the accuracy of the weather prediction and building model affect the optimization significantly. In the model-based optimization scenario, it would be necessary to implement an online model calibration procedure to maintain model fidelity within acceptable limits. However, this would further increase the complexity of the approach and its computational load. Thus, is there any other way? The question leads us to seek a new approach to solve the problem, and learning control may be appropriate. Although techniques of machine learning have been applied in many disciplines, the concept of learning control is still rather new in the domain of HVAC.

Kretchmar et al. [18] employed reinforcement learning assisted by artificial neural networks to learn to improve multiple-input multiple-output (MIMO) control performance of a heating system within a stable environment guaranteed by robust control. Henze and Dodier [19] investigated learning control of a grid-independent photovoltaic system consisting of a collector, storage, and a load. Q learning, a model-free reinforcement learning algorithm, was applied to optimize control performance of the system. Simulation analysis compared the performance between a conventional control strategy giving priority to the photovoltaic array (PV-priority) and the cost optimal control strategy. Better performance was found by applying the reinforcement learning to optimize the operation of the system. Henze and Schoenmann [20] investigated the application of reinforcement learning control to optimization of thermal energy storage system. Though reinforcement learning control proved sensitive to the selection of state variables, level of discretization, and learning rate, it effectively learned a difficult task of controlling thermal energy storage and displayed good performance. The cost savings compare favorably with conventional cool storage control strategies, but do not reach the level of predictive optimal control. These studies encouraged the authors to further pursue reinforcement learning to explore the optimal control of active and passive building thermal storage inventory.

Introduction to Reinforcement Learning

In our study, the problem of optimal control of active and passive building thermal storage inventory is formulated as sequential decision making problem, in which an agent or decision maker is repeatedly facing a problem to find a proper control strategy or policy given a certain state in order to achieve a long-term goal. The objective function J for this problem is

$$J = J(u_1^*, \ldots, u_l^*) = \min\left(\sum_{k=0}^{l} r_k P_k \Delta t\right) \tag{1}$$

Equation (1) describes that a controller in our problem takes a sequence of actions $u_1, \ldots, u_l$ over a selected time horizon of l time steps to minimize the electrical energy cost without demand charges for the building energy systems, where r_k is the price of electricity at time k; Δt is the time interval; and P_k is the total building electricity consumption, which could be either the sum of cooling and noncooling electrical loads, or the cooling related load only. Equation (1) does not account for demand charges but could be expanded to do so. The present study shows results for time-of-use utility rate structures without demand charges. The actions variables are defined as building global zone air temperature setpoint T_{sp}, which is the control variable for passive thermal storage, and the control command for the active thermal storage system, which is either a charging (+) or discharging (−) rate. The time step of the controller may be 1 h or defined less flexibly by introducing a *building mode*, which summarizes the consecutive hours that shares the same feature of building load and utility rate structure. For example, there may be three building modes, with the first one starting at midnight and continuing until the onset of the on-peak period, the second one covering the on-peak period, and the third one starting at the end of the on-peak period and continuing until midnight. The advantage of using the building mode is to decrease the dimension of the state space, which makes the task easier to learn.

It is not an exaggeration to say that the application of reinforcement learning algorithms to sequential decision making problems accounts for most of the current interest in reinforcement learning by machine learning researchers. Figure 1 sketches the general layout of a typical reinforcement learning problem.

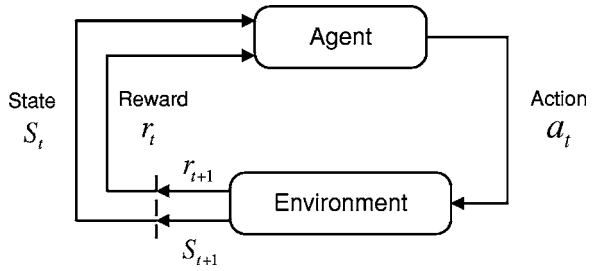

Fig. 1 Schematic of the reinforcement learning problem

As depicted in Fig. 1, at any moment t, the agent first senses the current condition of the environment, which is represented as state S_t, then selects an action a_t. The action cause the environment to change to a new state S_{t+1}, and after the state transition, the agent receives a reward r_t. Reinforcement learning problems describe a situation in which an agent interacts with the external environment to achieve a long-term goal, which is a measure of cumulated rewards over a finite or infinite sequence of decisions. Most reinforcement learning control problem adopt the framework of a Markov decision process (MDP), which exhibits the Markov property. A process is Markovian if the next state of the environment depends and only depends on the current state and current action to take. This property does not mean that the historical states are not important, but that all the historical information can be retained by the current state. For the Markovian case, a transition probability function then is introduced, defined as

$$p^a_{ss'} = \Pr(s_{t+1} = s' | s_t = s, a_t = a) \tag{2}$$

Similarly, the next reward is also defined as a function of current state, current action, and next state

$$R^a_{ss'} = E(r_{t+1} | s_t = s, a_t = a, s_{t+1} = s') \tag{3}$$

The policy is a mapping between the state space and the action space. In MDP, we usually define this mapping as a probability function $\pi(s,a)$ of taking action a when state is s. Given a policy and certain state, the value function is defined as

$$V^\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\} \tag{4}$$

where a discount factor γ discounts future rewards. Similarly, we define the value of taking action a in state s according to a policy as

$$Q^\pi(s,a) = E_\pi\{R_t | s_t = s, a_t = a\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\} \tag{5}$$

Notice that there is no terminal cost considered. The goal of reinforcement learning is to find the optimal policy that maximizes the value function. The core feature of reinforcement learning control can be viewed as combination of searching and long-term memory: Over time, continuous exploration eventually builds up a statistical library that contains the memory of previous experience. In any given state, the agent can choose the best action according to the library. Meanwhile, the library is updated with new experience available. This feature makes the reinforcement learning particularly attractive when dealing with the environment that may vary over time.

An important issue of reinforcement learning is known as *behavioral variety*, which describes the trade off between *exploration* and *exploitation*. Exploration refers the activity of evaluating the value of available actions, and exploitation means to utilize the current knowledge of action values to maximize the return. In a continuous problem, in a certain state, should we try a new action or use that action which is currently associated with the highest reward? The goal of optimal control is to maximize the return of the process over time. Choosing the action with the highest return is called the greedy action, but what if we do not know the available actions very well yet? The action with a lower immediate reward may possibly bring a higher return in the long run. Some action may be less frequently chosen than others in a certain state, this makes the estimation of action-value function inaccurate. There are many techniques to balance the tradeoff between exploration and exploitation. One of simplest approaches is called ε-*greedy* method, in which, instead of being greedy all the time, the agent takes nongreedy action once in a while, for example with the probability of ε. Another category of methods is called *softmax action selection* methods, among which the Gibbs or Boltzmann distribution is one of the most popular methods. This method defines the rule of choosing an action with probability

$$P(s,a) = \frac{e^{Q(s,a)/\tau}}{\sum_{b=1}^{n} e^{Q(s,a)/\tau}} \tag{6}$$

where τ is a positive parameter called temperature. High temperatures cause all actions to be (nearly) equiprobable; low temperatures cause a greater difference in selection probability for actions that differ in their value estimates. As $\tau \to 0$, softmax selection becomes the greedy action selection.

The Q-Learning Algorithm

The Q-learning algorithm introduced by Watkins [21] is considered one of the most important breakthroughs in reinforcement learning. The simplicity of the Q-learning algorithm makes it one of most popular reinforcement learning algorithms, meanwhile, it is also one of most efficient model-free learning methods. The key concept is that the action-value function $Q(s,a)$ is used directly to approximate the optimal value $Q^*(s,a)$. The updating rule is defined as

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \tag{7}$$

where $0 \leqslant \alpha \leqslant 1$ is a learning rate, and $0 \leqslant \gamma \leqslant 1$ is the discount factor as introduced previously. The $Q(s,a)$ values are usually stored in a lookup table, which is the so-called Q table, and each entry represents an estimation of the Q value of a state-action pair. A simple interpretation of the Q-learning algorithm is that the estimation of the optimal action-value function $Q(s,a)$ is a combination of past memory and new experience. The learning procedure is mostly influenced by two learning parameters: the discount factor γ and learning rate α. The discount factor expresses the confidence in the estimates of distant future costs and learning rate α determines the speed of evolution of the Q table. Higher learning rates will update the Q values faster, but it may not mean that learning is accelerated because the learning controller may learns towards the "wrong" direction. The Q-learning algorithm necessarily converges to the optimal solution for a Markov decision problem under the following two conditions:

1. The learning rate α has to decrease over time in accordance with the usual conditions common in stochastic optimization: not too fast and not too slow; and
2. In the limit, each action is tried in each state infinitely often.

This is a result of great importance since it shows that the Q policy will yield optimal sequences of actions. However, in practice, some of the requirements of the convergence theorem may not be met. In particular, the system may not be exactly a controlled MDP; it is also very hard to know when the algorithm has learned "long enough."

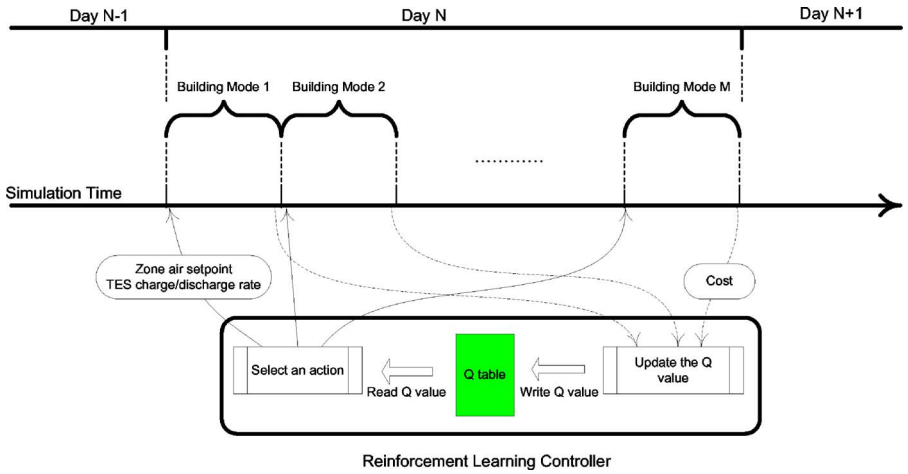

Fig. 2 Schematic of the learning controller

Development of Simulation Environment

A model of the dynamic thermal response and energy consumption of the HVAC system for a commercial building was developed in the technical computing environment Matlab/Simulink. The building thermal dynamics and energy consumption are modeled by four major groups of modules. The first group consists of the external and internal heat gain models including weather data, solar radiation, and building internal heat gain. The second group comprises the building envelope modules. Using state space modeling, the transient heat transfer through each construction element of the building is calculated using a second-order lumped capacitance model with three resistances and two capacitances per construction element. The third group contains all of the HVAC components modules required, which includes a VAV terminal box with reheat coil, an air handling unit including an economizer, a cooling coil, and a circulation fan, and finally a simple plant module including an electrical chiller, a cooling tower, and a chilled water pump. All modules are based on the models in ASHRAE Secondary Toolkit [22]. The modules for heat gain, building envelope, and HVAC components are linked to the fourth group, which is a thermal and humidity balance function that updates the thermal condition in each time step. There is a utility block included to calculate the electrical energy cost of the building energy systems, which is subject to the time-of-use rate structure.

The Matlab/Simulink model is calibrated against an EnergyPlus [23] model, which is developed using identical building information. For the purpose of evaluating the learning controller (Fig. 2), the EnergyPlus model imitates the actual building performance. Figure 3 compares the cooling load profiles of the EnergyPlus model and Simulink model both before and after calibration.

The developed building model is about 3000 m^2, and lighting power densities of typical office buildings are assumed. The utility rate structure and occupancy schedule are depicted in Fig. 4. It can be seen that the onset of the utility on-peak rate and full occupancy are synchronized in order to simplify the problem.

One of the main objectives of the simulation study is to evaluate the feasibility of the reinforcement learning algorithm for op-

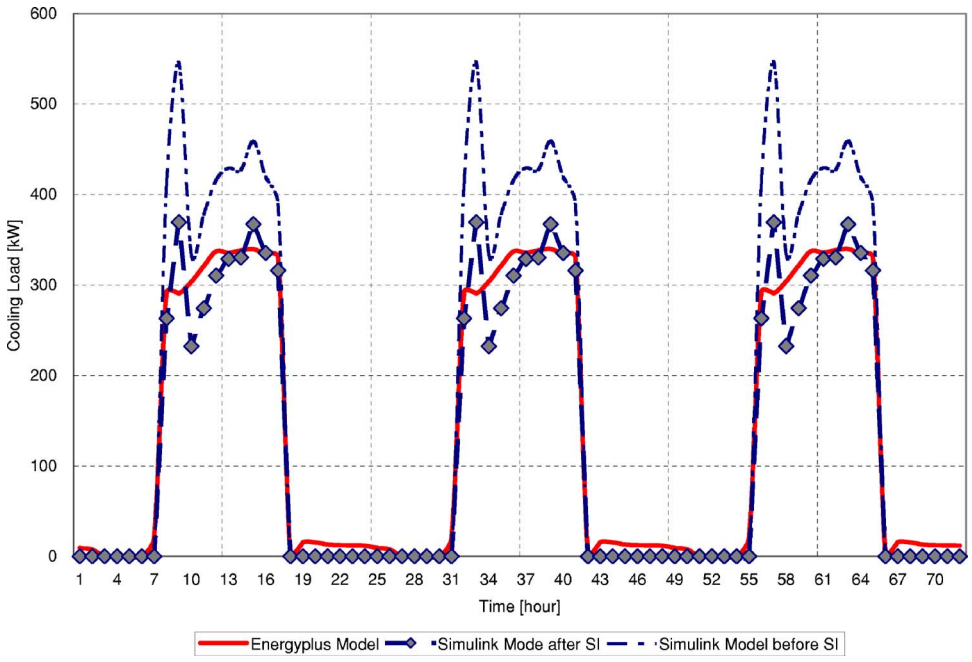

Fig. 3 Cooling load profiles of EP model and Simulink model before and after calibration

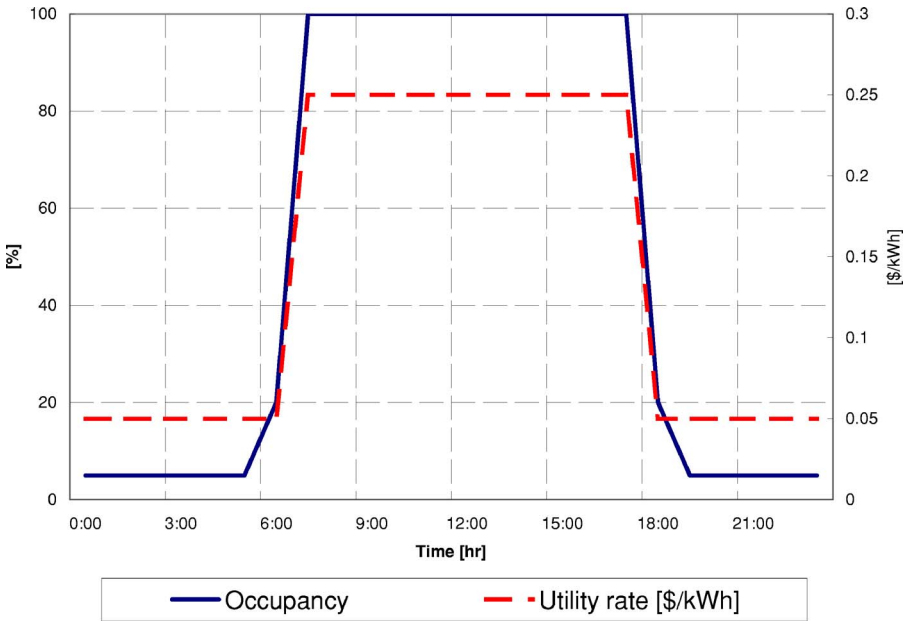

Fig. 4 Schedule of utility rate and internal heat gain

timal control of thermal storage. The uncertainty of the environment and configuration of state and action space are the determining factors that affect the performance of the learning controller. In the initial step of simulation analysis, we try to minimize the uncertainty of the environment. One of our concerns is the influence of weather. Weather conditions are not considered as state space variables due to following considerations:

1. Modern commercial buildings, which is the objective application of this study, are mostly internal heat gain dominated buildings. Improved construction materials provide good insulation of the building envelope, and the heat transfer from the ambient environment only slightly contributes to the total cooling load. The internal heat gains such as lighting, electric equipment, and occupants comprise the majority of the cooling load; and
2. The reinforcement learning control algorithms are designed to work in an unknown environment. Uncertainty brought about by the variation of ambient conditions is not supposed to greatly influence the optimal policy found by the learning controller.

Typical meteorological weather data (TMY2) for Omaha, NE is selected as the ambient weather condition. A repeating sequence block is implemented in the simulation model, which simulates a single-day weather profile including dry bulb temperature, relative humidity and wind speed. The function of this block is to repeat the weather vectors through the whole simulation or training period. By doing so, the assumption is made that the same weather condition is maintained for all the training days.

The reinforcement learning controller is implemented as depicted in Fig. 2. Each building mode is considered as a stage in the sequential decision making problem, which can be flexibly defined as 1 h or multiple hours that have the same characteristics. A Simulink block for the Q-learning controller is implemented into the simulation environment and the calling sequence of the Q-learning controller is presented in Fig. 2. The Q value of each action-state pair is stored in the Q table, which contains the current knowledge of the learning controller about the environment. At the beginning of each building mode period, the controller observes the current state of the environment, then chooses an action from the action space (the set of pairs of values of global zone air temperature setpoint and TES charging/discharging rate). The selected action is executed by the simulation model, and at the end of the current building mode period (beginning of next mode) the cost of the chosen action is computed, which is fed back along with the information on the new state into the learning controller. The cost and new state are used to update the Q table according to the Q-learning algorithm. In our simulation study, we assume that the capacity of the HVAC system is always capable of reaching the control setpoint, but this may not be true in an actual application. In the case where the HVAC system does not have enough capacity, the selected action will not lead to the expected state transition. However, even in this case, as long as the state is fully observable and the actual state can be ascertained, the Q table is still updated properly, and the optimal policy is still idenified eventually.

Construction of the state and action space is the most critical task in the development of the learning controller. Even though past research efforts improved the efficiency of reinforcement learning algorithms to solve sequential decision making problems, they still suffer from the curse of dimensionality. The performance of the learning controller can easily be poor when dealing with a complex environment and high-dimensional state and action space. Table 1 lists the candidate state and action variables in our case.

The choice of action space is comparatively straightforward. The action variables are global zone air temperature setpoint T_{sp}

Table 1 Possible state and action space configurations

	Variable Name	Dimension
State space	Building mode	3, 6, 9, …, 24
	State of charge	10, 20, … (depends on the resolution)
	Zone air temperature	10, 20, … (depends on the resolution)
	Ambient temperature	10, 20, … (depends on the resolution)
Action space	Global zone air temperature setpoint	10, 20, … (depends on the resolution)
	TES Charging/discharging rate	10, 20, … (depends on the resolution)

Table 2 Constraints for the action variables

		On peak	Off peak
Global zone air	Upper bound	24	30
Temperature setpoint	Lower bound	20	15
TES (charging/discharging rate)	Upper bound	$u_{\max}$	
	Lower bound	$u_{\min}$	

and TES charging/discharging rate u, which are discretized within the allowable range. Table 2 lists the constraints for each action variable.

As Table 2 shows, the constraints on the zone air temperature setpoint are determined by the requirement of thermal comfort. The limits on the TES charging/discharging rate are time dependent, as expressed in the following equations

$$u_{\min,k} = \max\{-Q_k, u_{x_{k+1}=x_{\min}}\}$$

$$u_{\max,k} = \min\{\mathrm{CCAP}_k - Q_k, u_{x_{k+1}=x_{\max}}\} \quad (8)$$

where $u_{x_{k+1}=x_{\min}}$ denotes the control action u that leads to a state-of-charge x' when operating the cooling system over the next time step $k \rightarrow k+1$ at control u. Thus, no actions can be taken that would lead to states of charge outside the limits, i.e., full and empty storage tank, respectively. Furthermore, the chiller capacity CCAP_k minus the current load Q_k defines the maximum charging rate $u_{\max,k}$, i.e., how much can be charged after meeting the load.

Analysis

The following sections present results for an application of the reinforcement learning controller to active and passive thermal storage capacity. The discussion starts with the single-task scenarios, in which either the active or passive thermal storage capacity is considered. In these cases, the corresponding state and action space is comparatively small, which makes the learning task easier. The present study confirms the feasibility of the reinforcement learning approach, and analyzes the effect of different state and action space configurations and learning parameters on the learning procedure.

Reinforcement Learning Controller for Passive Thermal Storage Only

Model-based predictive optimal control was shown to successfully find the optimal setpoint profile that can substantially reduce the peak cooling load and operating cost under time-of-use utility rate structures. The objective of this section is to demonstrate that the optimal control policy can also be found by reinforcement learning control, based only on the interaction between the controller and the environment. Liu and Henze [24] summarize the results from an early study of the application of reinforcement learning to thermal mass control. A parametric analysis had been carried out subsequently to analyze the effect of different learning parameters and state–action space configurations on the performance of the learning controller. It was found that for all of the cases the controller approaches the optimal policy, yet different learning parameter settings significantly affect the speed of learning. The softmax method shows an advantage over the ε-greedy method because it relates the policy of action selection with the Q value for the each action–state pair. A dynamic learning rate proved to facilitate the convergence of the learning control as expected. It was also found that the dimensions of the state–action space affect the learning controller with respect to both speed and performance. Figure 5 compares the optimal zone air setpoint profiles of three learning control cases with the ones for model-based optimal control, which is considered the "true" optimal solution given no model mismatch and perfect predictions. In Fig. 5, the numeral preceding "BM," e.g., 3BM, indicates the number of building modes per weekday. As shown in Fig. 5, although the reinforcement learning cases do not reach the "true" optimum, they do recognize the benefit of precooling the building during the off-peak period and find a near-optimal policy. The cost savings are presented in Table 3.

The cost savings shows that the reinforcement learning case is close to the model-based optimization case. However, the learning controller usually takes a long time to find the optimal policy and the training time increases with the dimension of the Q table. In the reinforcement learning case with nine building modes, even after over 6000 training days, the controller is only able to find a near-optimal policy, and the cost savings is reduced by about 5% compared with the value found for three building modes. The computation efficiency of reinforcement learning is only competitive with the model-based approach in the low-dimensional cases, and deteriorates quickly when the dimension of state space and action space grows. In summary, the simulation results confirm that reinforcement learning is a feasible methodology to find the optimal policy to control the passive building thermal storage inventory; however, the performance of the learning controller is substantially affected by the selection of learning parameters of the configuration of state and action space.

It is surprising to see that reinforcement learning with dynamic learning rate, softmax method, and three building mode achieves an even lower daily cost than the model-based optimization. Com-

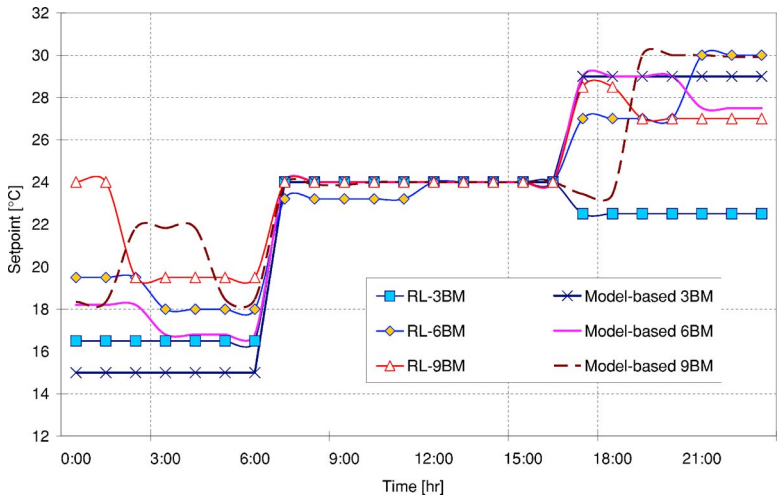

Fig. 5 Optimal zone air temperature setpoint profile

Table 3 Reinforcement learning controller for passive thermal storage only, compared to model-based approach

		Reinforcement learning			Model based cases		
Case	Base case	3BM	6BM	9BM	3BM	6BM	9BM
Cost ($)	216	181.2	189.8	193	184.1	184.5	189.1
Saving (%)	—	16.1	12.1	10.6	14.7	14.5	12.4
			Computation efficiency				
No. of training days		<500	>3000	>6000	—	—	—
No. of Iterations		—	—	—	485	487	1555

paring the optimal setpoint, it can be found that both cases found almost identical setpoints for building Modes 1 and 2, but differ in the third building mode. The model based decides to let the setpoint float, but the learning controller chooses to cool down the building. The difference can be interpreted as follows: the model-based approach uses a single day's cooling load profile, and there is no cooling load after the third building mode. So the optimal action is to let the plant shut down and temperature float. On the other hand, in the training of learning controller, the building is operated continuously. Since the controller realizes there is a cooling load after the third building mode, it will find out which action is best eventually.

Reinforcement Learning Controller for Active Thermal Storage Only

Henze and Schoenmann [20] investigated the application of reinforcement learning control to the optimization of a thermal energy storage system. A reinforcement learning controller was designed to learn to charge and discharge a commercially available thermal energy storage system in a commercial building to exploit load-shifting utility rate incentive. The reinforcement learning control was compared with three conventional control strategies including chiller priority, storage priority, as well as model-based predictive optimal control. As expected, model-based predictive optimal control provides the highest savings in all three cases. Reinforcement learning control performance is only moderately inferior compared to the performance of the storage-priority control. This conclusion is also confirmed in the present investigation. Table 4 compares the daily operating cost and the savings for various cases.

The model-based optimization provides the best cost savings. TES-priority control provides the highest cooling load shifting, however, it consumes too much energy during the off-peak period partially because the coefficient-of-performance (COP) of the ice-making chiller is lower than its nominal value in the (subfreezing) charging mode. The result confirms the previous study in that reinforcement learning control offers slightly inferior cost saving than TES-priority control. The training period is limited to 6000 days; it can be expected that the learning controller can do better if more training is applied.

Multitask Scenario

The study of single-task scenarios confirms the feasibility of reinforcement learning algorithm in optimal control of either active or passive building thermal storage inventory. A past parametric analysis of different configuration of state and action space and learning parameters revealed the effect of each parameter on the performance of the learning controller and provided valuable experience for the following analysis. In this section, the reinforcement learning controller is designed to learn to control the active and passive thermal storage simultaneously. The results show that the curse of dimensionality dramatically slows down the learning procedure of the controller in standard Q learning. The term standard Q-learning approach refers to the ordinary methodology applied in the previous study of single-task scenarios. The Q table starts with zeros for all the entries, and the learning rate may vary with learning but is applied uniformly to all Q-table coordinates, which is also called synchronous Q learning. Table 5 lists the configuration of the state and action space.

Besides adjusting the learning parameters, efforts have been made to improve the performance of the learning controller with three additional techniques. The first one is to use the randomization of the initial Q table. As mentioned earlier, the simulation previously initializes the Q table with zeros for all the entries. By doing so, there is no preference for the controller to select the action at the early period of the learning procedure. This represents the case where the learning controller has no prior knowledge at the beginning. Alternatively, there is the option to initialize the learning controller, by randomly assigning values to each state–action pair. It is interesting to find that better learning results are achieved by using the randomization of the initial Q table. The reason behind this effect might be interpreted as that the randomness introduces certain "subjective" knowledge, which makes it easier for the learning controller to "jump around" in either ε-greedy or softmax methods. By doing so, the state–action space is explored faster and the learning process is accelerated.

The second option is known as asynchronous Q learning. Dar and Mansour [25] state that there are two variations of the Q-learning algorithms by varying the definition of the learning rate. The first one is called *synchronous Q learning*, in which the learning rate varies uniformly for all the entries in the Q table

$$Q(s,a) = Q(s,a) + \alpha(t)[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (9)$$

Equation (9) is the same as Eq. (7) except for varying the learning rate as a function of time. *Asynchronous Q learning* is defined as

Table 4 Comparison of cost and savings for active thermal storage optimal control

Case	Base	Chiller priority	TES priority	Reinforcement learning	Model based
Cost ($)	216	208.2	187.2	189.4	181.5
Saving (%)	—	3.6	13.3	12.3	16.2

Table 5 State and action space configuration for multitask scenario

	Variable name	Dimension
State space	Building mode	3, 6
	State-of-charge	10
Action space	Global zone air Temperature setpoint	10
	TES Charging/discharging rate	20

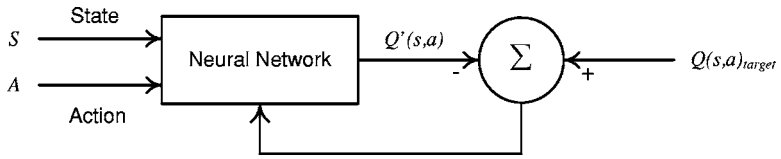

Fig. 6 Updating *Q* value with an artificial neural network

$$Q(s,a) = Q(s,a) + \alpha(t,s,a)[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \tag{10}$$

In contrast to Eqs. (7) and (9), Eq. (10) associates the learning rate not only with time, but also with the state–action pair. There is always a situation in a learning process where some entries of the Q table are visited only a few times; meanwhile others have been sampled "enough." Applying the same learning rate to these less sampled entries would be unfair because they have not been fully explored. Asynchronous Q learning provides an efficient way to avoid this discrimination. Instead of uniformly changing the learning rate, the learning rate is determined by the number of times a specific entry has been visited. Higher learning rates are applied when the state action is new, and the more the state action is sampled, the more the learning rate is decreased. It was found that the asynchronous Q-learning approach converges faster than synchronous Q learning in most simulation cases. Compared with the previous simulation cases, only one-third or even less previous training time is required to reach the near-optimal solution, although the cost did not decrease in most cases.

The third option applies artificial neural networks (ANNs) to replace the Q table. One of the most powerful application of ANN is function approximation. By embedding a vast number of simple neurons in an interconnected parallel processing system, ANN provides computational power for very sophisticated information processing. For this reason, ANN offers an efficient solution to overcome the curse of dimensionality problem that bedevils attempts to model nonlinear functions with large numbers of variables, such as the value function in reinforcement learning problems. The procedure of updating the Q value $Q(s,a)$ of a specific entry is replaced with training the neural network to approximate the target value, which is depicted in Fig. 6.

Simulation studies carried out in the context of this paper, confirm that ANNs are a feasible solution to replacing the Q table, but it did not reveal any advantages over standard Q learning. The simulation of reinforcement learning using ANN consumed much more computation time because the training process of the neural network is more involved than simply replacing old Q-table values. Nevertheless, application of ANN leads the learning controller to approach a lower daily cost faster in terms of training days. Figure 7 compares the moving average daily cost profiles of standard Q learning with the improved learning case.

It can be seen that the techniques discussed above did improve the performance of the learning controller with respect to learning speed and performance. The randomly initialized Q table finds the best optimal result overall, and asynchronous Q learning and application of ANN improve learning speed. Even though asynchronous Q learning did not show the lowest daily cost among the simulation cases investigated, we believe that improved learning parameters would not only make it the fastest learning but also the best performing learning controller.

Figures 8 and 9 show an example of how the controller learns in the present study. Figure 8 presents the profiles of zone temperature setpoint T_{sp} in each building mode. It can be seen that during building Modes 3 and 4 the setpoints are gradually raised in order to reduce cooling requirements during on-peak periods. Two building modes find the merit of precooling, which are building Modes 2 and 6. It is not surprising to see that the controller tries to precool the building in building Mode 6 because the simulation runs continuously and the precooling effect can consequently lower future cooling load. Yet, in building Mode 1 the controller only finds suboptimal solutions as the building is not precooled. Figure 9 illustrates the learning profiles of the TES charge/discharge rate u in each building mode. The controller finds the advantage of load shifting very quickly in building Modes 3 and 4, which coincide with the on-peak period. It also gradually learns to charge during building Modes 1, 2, and 6, respectively. However, building Mode 5 presents a discharge action that does not make sense since the controller decides to dis-

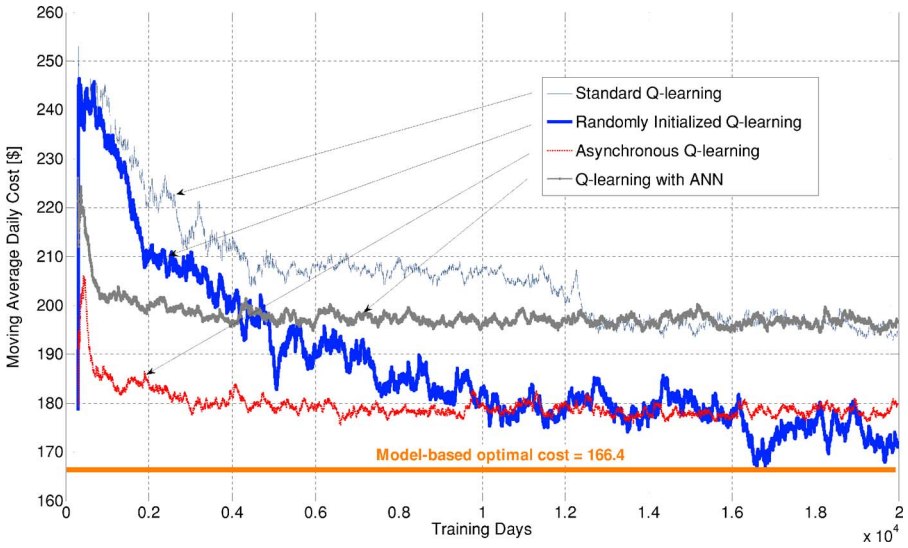

Fig. 7 Comparison of all reinforcement learning with multitasks scenario

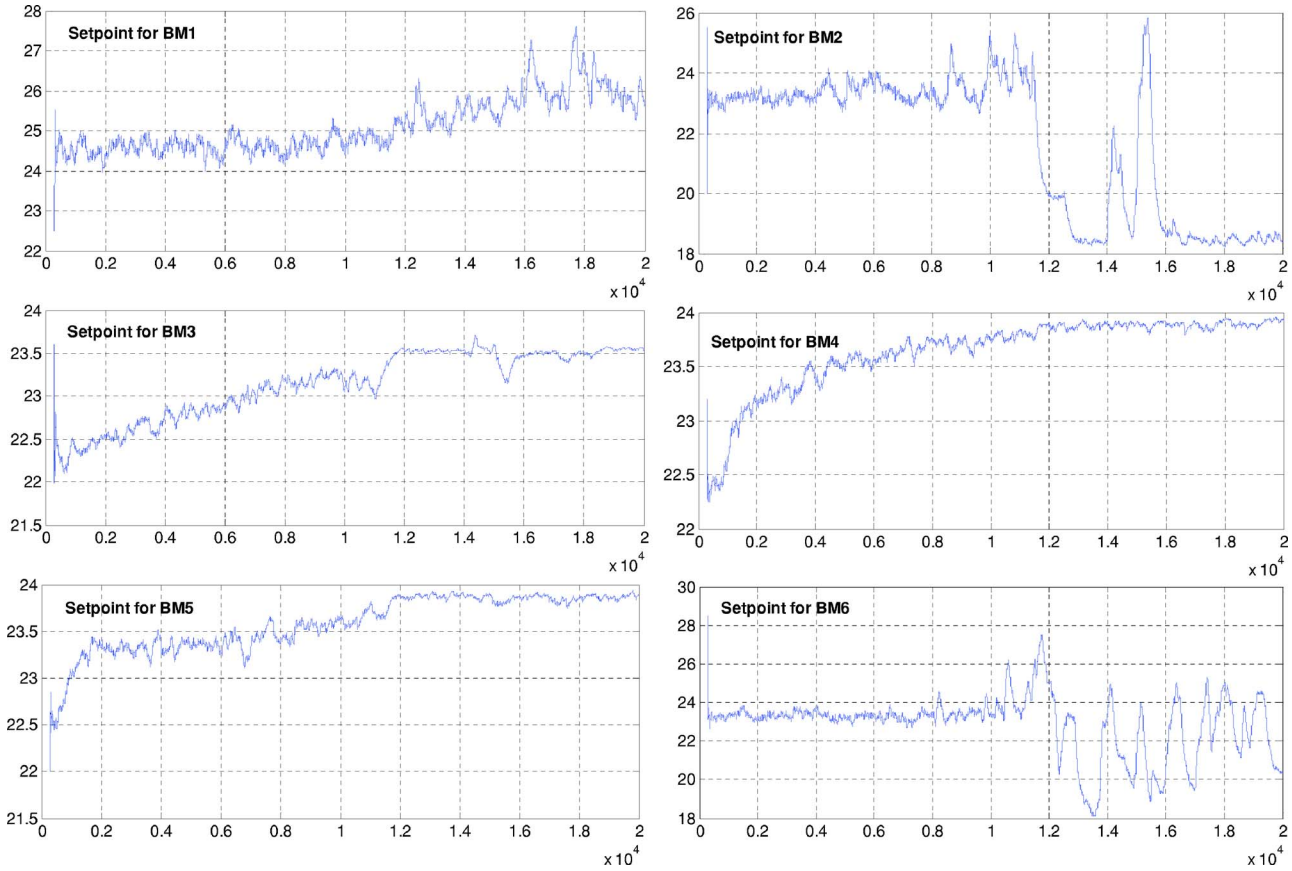

Fig. 8 Learning of zone temperature setpoints T_{sp}

charge during the off-peak period from 6 p.m. to 9 p.m., i.e., it depletes TES inventory when energy is inexpensive.

Similar to the single-task scenarios, the performance of the learning controller is evaluated by comparing the cost savings with other control strategies. The training computation required is very great in this case; the minimum training of the reinforcement learning controller is more than 6000 training days even in the fastest case, which is asynchronous Q learning, and the result is still only near optimal. The reinforcement learning algorithm performs acceptably well compared with model-based optimization in this highly dimensional optimization problem. Table 6 lists the cost savings of the investigated cases.

Table 6 lists all control strategies utilizing both active and passive thermal storage inventory. Cases 2 and 3 stand for the conventional control strategies for the active TES system, but the optimized zone air temperature setpoint is applied, which is obtained from the model-based optimization approach. It is interesting to see that Case 3 provides higher saving than model-based optimization because of the following reason. The model-based optimization approach is introduced as a benchmark to evaluate the learning control performance, which uses the operating cost of a single day simulation. As mentioned above, the model-based optimal controller did not fully utilize the active TES inventory because single day simulation cannot see the coming cooling load from the next day. That is why the TES is dormant in the second off-peak building mode of the day. To overcome this effect, the model-based optimization is also carried out in a continuous fashion (Case 5). Case 5 is slightly better than Case 3 because of the initial condition of state of charge of the TES. The learning controller is trained continuously, and the inventory of TES leftover will be carried forward to the next day, which could be utilized to meet the cooling load.

6 Conclusion

An extensive investigation has been carried out to analyze the application of model-free learning control on the optimization of building active and passive thermal storage inventory. The simulation study covers different learning scenarios including single task, where either active or passive thermal storage media are considered, and multitask that simultaneously control both active and passive storage, as well as analyzes the effect of different learning parameters and the configuration of state and action space. In general, the results of the simulation study confirm that reinforcement learning control is a feasible methodology to derive the optimal control policy for this specific problem. The learning controller successfully learns to precool the building when the building is unoccupied, and charges or discharges the TES according to the utility rate structure. The operating savings of the plant do not reach the value of model-based predictive optimal control, but are still substantial compared with conventional control strategies. It is worth mentioning that there are some simulation cases that did reach the optimum found by the model based, especially when the dimension of the learning problem is low. Theoretically the reinforcement learning algorithm can reach the true optimum given properly selected learning parameters and enough learning time. However, it seems to be easier for the learning controller to find general optimal action patterns rather than to extract a refined control policy.

The savings were only achieved given sufficiently long learning periods for the reinforcement learning controller. The amount of training is not realistic if the controller is directly implemented into a real application. This constitutes the major drawback of the reinforcement learning control approach. It is also found that the selection of state and action variables strongly affects the learning

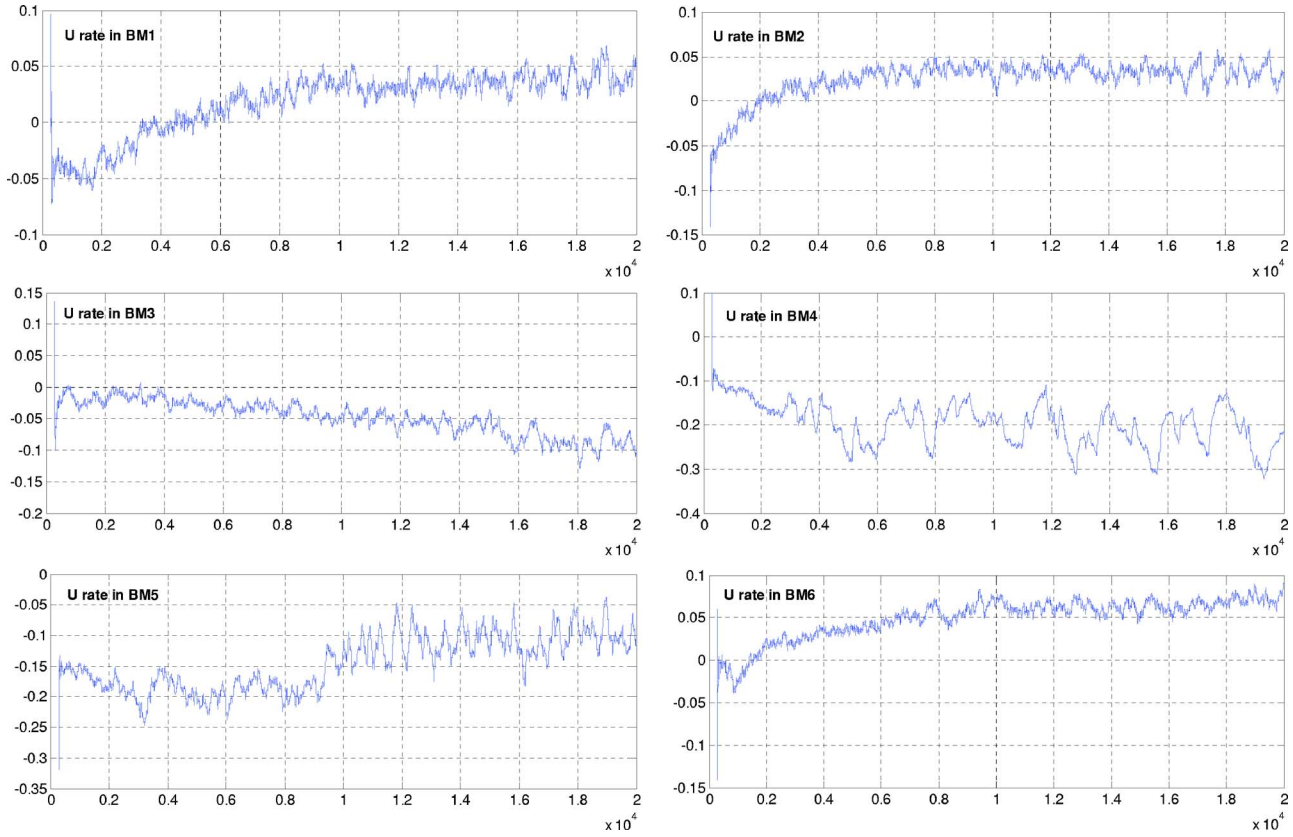

Fig. 9 Learning of TES charge/discharge rate *u*

procedure of the controller. The curse of dimensionality is evident in the application of reinforcement learning: both single-task and multitask scenarios showed that the performance of the controller goes down when the dimension of the problem increases.

Selection of proper learning parameters is another critical consideration for the learning controller. The present study shows that the discount factor should be low when more uncertainty is involved in the estimation of the value function. As supported by past work, the learning rate must be decreased over the training process, but the definition of the dynamics of the learning rate can be defined freely. Also, the softmax exploration algorithm was shown to outperform the ε-greedy method.

Several techniques have been identified to be effective in improving the performance of the learning controller, namely, randomization of the Q table, synchronized Q learning, and the application of artificial neural networks. However, even after greatly shortening the learning period, the time and effort to make the controller learn is still very long. This indicates that a learning controller with no prior domain knowledge is not a practical solution and that in some form contextual information needs to be introduced to expedite learning of the fundamental features of the problem, while reinforcement learning can accommodate fine tuning of the domain knowledge.

Table 6 Comparison of cost and savings for active thermal storage optimal control

No.	Case	Cost ($)	Savings (%)
1	Base case	216.2	—
2	Chiller priority with optimized T_{sp}	179.1	17.2
3	TES priority with optimized T_{sp}	161.1	25.4
4	Model-based opt. $SOC_{ini}=0$ single day	166.4	23.0
5	Model-based opt. $SOC_{ini}=0$ continuous	157.2	27.4
6	Reinforcement learning control	168.1	22.2

Acknowledgment

The authors gratefully acknowledge the financial support of this work through U.S. Department of Energy NASEO Cooperative Agreement No. DE-FC-36-03GO13026.

Nomenclature

a = action
s = state
P = transition probability function
R = reward function
α = learning rate
γ = discount factor
V = value function for given state
Q = value function for given state and action
π = policy function

References

[1] Conniff, J., 1991, "Strategies for Reducing Peak Air-Conditioning Loads by Using Heat Storage in the Building Structure," ASHRAE Trans., **97**(1), pp. 704–709.
[2] Morris, F. B., Braun, J., and Treado, S., 1994, "Experimental and Simulated Performance of Optimal Control of Building Thermal Storage," ASHRAE Trans., **100**(1), pp. 402–414.
[3] Rabl, A., and Norford, L., 1991, "Peak Load Reduction by Preconditioning Buildings at Night," Int. J. Energy Res., **15**, pp. 781–798.
[4] Keeney, K., and Braun, J., 1996, "A Simplified Method for Determining Optimal Cooling Control Strategies for Thermal Storage in Building Mass," HVAC&R Res., **2**(1), pp. 59–78.
[5] Braun, J., 1990, "Reducing Energy Costs and Peak Electrical Demand through Optimal Control of Building Thermal Mass," ASHRAE Trans., **96**(2), pp. 876–888.

[6] Andresen, I., and Brandemuehl, M., 1998, "Heat Storage in Building Thermal Mass: A Parametric Study," ASHRAE Trans., **98**(1), pp. 910–918.
[7] Braun, J., Montgomery, K., and Chaturvedi, N., 2001, "Evaluating the Performance of Building Thermal Mass Control Strategies," HVAC&R Res., **7**, pp. 403–428.
[8] Braun, J., 2003, "Load Control Using Building Thermal Mass," J. Sol. Energy Eng., **125**(3), pp. 292–301.
[9] Henze, G., Krarti, M., and Brandemuehl, M., 1997, "A Simulation Environment for the Analysis of Ice Storage Controls," HVAC&R Res., **3**(2), pp. 128–148.
[10] Henze, G., Krarti, M., and Brandemuehl, M., 2002, "Guidelines for Improved Performance of Ice Storage Systems," Energy Build., **35**(2), pp. 111–127.
[11] Henze, G., Dodier, R., and Krarti, M., 1997, "Development of a Predictive Optimal Controller for Thermal Energy Storage Systems," HVAC&R Res., **3**(3), pp. 233–264.
[12] Henze, G., and Krarti, M., 1999, "The Impact of Forecasting Uncertainty on the Performance of a Predictive Optimal Controller for Thermal Energy Storage Systems," ASHRAE Trans., **105**(2), pp. 553–561.
[13] Kintner-Meyer, M., and Emery, A., 1995, "Optimal Control of an HVAC System Using Cold Storage and Building Thermal Capacitance," Energy Build., **23**(3), pp. 19–31.
[14] Henze, G., Felsmann, C., and Knabe, G., 2004, "Evaluation of Optimal Control for Active and Passive Building Thermal Storage," HVAC&R Res., **9**(3), pp. 259–275.
[15] Henze, G., Kalz, D., Felsmann, C., and Knabe, G., 2003, "Impact of Forecasting Accuracy on Predictive Optimal Control of Active and Passive Building Thermal Storage Inventory," HVAC&R Res., **9**(3), pp. 259–275.
[16] Liu, S., and Henze, G. P., 2004, "Impact of Modeling Accuracy on Predictive Optimal Control of Active and Passive Building Thermal Storage Inventory," ASHRAE Trans., Technical Paper No. 4683, **110**(1), pp. 151–163.
[17] Henze, G. P., Kalz, D., Liu, S., and Felsmann, C., 2005, "Experimental Analysis of Model-Based Predictive Optimal Control for Active and Passive Building Thermal Storage Inventory," HVAC&R Res., **11**(2), pp. 189–214, American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Atlanta, GA.
[18] Kretchmar, R. M., Young, P. M., Anderson, C. W., Hittle, D. C., Anderson, M. L., Delnero, C. C., and Tu, J., 2001, "Robust Reinforcement Learning Control With Static and Dynamic Stability," Int. J. Robust Nonlinear Control, **11**, pp. 1469–1500.
[19] Henze, G., and Dodier, R., 2003, "Adaptive Optimal Control of a Grid-Independent Photovoltaic System," J. Sol. Energy Eng., **125**(1), pp. 34–42.
[20] Henze, G., and Schoenmann, J., 2003, "Evaluation of Reinforcement Learning Control for Thermal Energy Storage Systems," HVAC&R Res., **9**(3), pp. 259–275.
[21] Watkins, C., and Dayan, P., 1992, "Q-Learning," Mach. Learn., **8**, pp. 279–292.
[22] Brandemuehl, M. J., 1993, *HVAC 2 TOOLKIT*, 1st ed., American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA.
[23] U.S. Department of Energy, 2005, "Energyplus 1.2.1," http://www.eere.energy.gov/buildings/energyplus/
[24] Liu, S., and Henze, G. P., 2004, "Investigation of Reinforcement Learning for Building Thermal Mass Control," *Proceedings of SimBuild 2004*, Boulder, CO, August 4–6, International Building Performance Simulation Association.
[25] Dar, E., and Mansour, Y., 2003, "Learning Rates for q-Learning," J. Mach. Learn. Res., **5**(1), pp. 1–25.