Model Predictive Prior Reinforcement Learning for a Heat Pump Thermostat

Kuo Shiuan Peng
Electrical and Computer Engineering
University of Arizona
Tucson, Arizona, USA
kspeng@email.arizona.edu

Clayton T. Morrison
School of Information
University of Arizona
Tucson, Arizona, USA
claytonm@email.arizona.com

ABSTRACT

We combine results from model predictive control, reinforcement learning, and set-back temperature control to develop an algorithm for adaptive control of a heat-pump thermostat. The algorithm borrows from model predictive control the concept of optimizing a controller based on a model of environment dynamics, but then updates the model using online reinforcement learning. An adaptive set-back heuristic further improves energy savings while maintaining target temperature goals. We evaluate the framework in simulation, demonstrating its advantages over standard model predictive control and reinforcement learning alone.

CCS Concepts

•**Theory of computation** → **Theory and algorithms for application domains;** *Machine learning theory;* Reinforcement learning; Sequential decision making; •**Computing methodologies** → **Artificial intelligence ;** *Control methods ;* Computational control theory;

Keywords

heat pump thermostat; model predictive control; model-free reinforcement learning; set-back strategy

1. INTRODUCTION

Residential and commercial buildings around the world consume about 20–40% of global energy [1, 2]. This is especially true of heating, ventilation and air conditioning (HVAC) systems, which consume over half of this energy. Heat-pump thermostats have been actively studied for decades, with the goal of improving building energy consumption efficiency. However, with changes in technology, energy and comfort management in smart energy buildings remains an open problem and active research area [3, 4].

In this paper, we develop and evaluate a method for synthesizing an efficient control strategy for a heat-pump thermostat. In our application domain, heat-pump control regulates building temperature to maintain a thermal comfort condition while maximizing power consumption efficiency. In our approach, we combine model predictive control (MPC) and reinforcement learning (RL) methods for control strategy synthesis. MPC uses an explicitly formulated model of the process to solve open-loop deterministic optimal control problems [5, 6]. MPC generally achieves state of art performance and is a popular choice for thermal control regimes. However, the approach relies on the availability of an accurate and stable model of building thermal dynamics, and it can be costly to develop and maintain model quality [6]. RL, on the other hand, provides a method for synthesizing near-optimal control strategies in a model-free setting, based on direct interaction with the control task environment. This approach does not rely on a pre-existing environment dynamics model and in principle can adapt to changes in the environment dynamics. However, training may be relatively slow, requiring a significant amount of costly experience interacting with the environment in order to achieve competitive performance. We develop a hybrid method that combines the strengths of MPC and RL while minimizing either's shortcomings.

The paper is structured as follows. Section 2 presents a review of prior work in model predictive control, reinforcement learning, and set-back control as applied to heat-pump thermostats. In section 3 we present our approach, which combines a model predictive prior with set-point temperature control. Section 4 presents an evaluation of our method compared to MPC and RL alone in simulation, and section 5 concludes with discussion and future work.

2. RELATED WORK

Model predictive control has become the dominant popular approach to heat-pump thermostat control [5, 6, 7]. At each decision point, the controller selects an action by solving a fixed-horizon optimization problem. This process depends on an accurately calibrated model of the task environment. Several different types of HVAC system control schemes have been developed, including conventional controllers, hard controllers, soft controllers, and hybrid controllers [8]. These methods optimize lower energy consumption and better transient response to changes in indoor air temperature. However, MPC controllers are only as good as their models and require accurate knowledge of the operating environment conditions, which realistically may change over time, for example due to changes in local building neighborhood, changes in global weather patterns, or changes in building occupancy patterns.

Feedback Computing '16 July 19, 2016, Wurzburg, Germany

ACM ISBN 978-1-4503-2138-9.

DOI: 10.1145/1235

An adaptive controller synthesis paradigm is preferred when the task environment dynamics are not assumed to be completely known ahead of time or might change. Q-learning [9] has been proposed as one approach to minimizing the electricity cost of thermal energy storage systems [10]. Reinforcement reward functions that incorporate both thermal target and energy consumption objectives have been studied [11]. Methods have also been explored for reducing the amount of online training required through the use of training in simulation that may only approximate the true system dynamics, but helps prime the policy so that required online training is reduced [12, 13, 14]. How to design the simulated training environment for optimally efficient Q-learning remains and open research problem.

In the problem of designing the control agent of a heat-pump thermostat, there are a series decisions that must be made that require knowledge of the system dynamics. In a review comparing RL with MPC [7], the conclusion was that the proper way to address this kind of problem was to combine model-based technology such as MPC and learning-based techniques such as RL.

In addition to combining RL and MPC approaches to controller synthesis, we also incorporate the use of a "set-back" strategy for heat-pump control [15]. The conventional control paradigm for a heat pump keeps its temperature set-point constant during the day. The set-back strategy relaxes the set-point (controlled target) temperature during convenient times of the day, for example when the occupants are not in the building, to reduce power consumption. This method reduces the overall power consumption, but can itself become a key source of energy usage inefficiency as the controller engages to change the temperature from the set-back state to the set-point state.

In this paper, we make two contributions. First, our approach seeks a compromise between MPC and RL by using the concept of a prior environment dynamics, as used by MPC, to determine the preliminary policy, and then refine and update the policy by RL during online training. Second, we employ our MPC+RL method in the context of the set-back strategy to produce an adaptive set-point control method to reduce overall power consumption. We demonstrate that the proposed method gains the advantages of MPC, RL, and the set-back strategy to provide a high performance heat-pump controller synthesis method.

3. MODEL PREDICTIVE PRIOR REINFORCEMENT LEARNING

3.1 Problem Statement

In this work, we make the simplifying assumption that energy consumption, E, is equivalent to heat-energy production, Q. These quantities are related to building temperature control by the following equation:

$$E \approx Q = UA\Delta T = UA(C - B). \tag{1}$$

Here, U represents thermal capacity, the degree to which materials conduct or resist heat.[1] A is the area of the surface that the heat is flowing through. ΔT is the temperature difference between the target controlled temperature C (as set by the heat-pump controller) and the building temperature B. In a specific environment, U and A are assumed constant. The temperature difference ΔT is the main factor determining energy consumption E. In order to simplify the problem, we assume that the efficiency of converting heat-energy production to energy consumption is constant in this study. We also ignore the ramping-up or ramping-down time and the corresponding energy consumption to activate to the target controlled temperature in the the heat-pump thermostat.

[1]Thermal capacity is strictly positive and represents how much energy the material needs to rise 1 degree celsius. The baseline value $U = 1$ represents water.

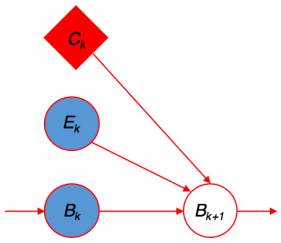


Figure 1: Graphical model representing building temperature control problem.

Heat-pump control can be naturally formulated as a sequential decision making problem. At each state $s \in S$, the decision-making agent selects an action $a \in A$ that maximizes the reward r_k. The reward r_k is considered as the negative value of the linear combination of temperature error and the energy consumption:

$$r_k = -(1 - w_e) \times (\Delta T_k) - w_e \times (e_k) \ , \tag{2}$$

where ΔT_k is the temperature difference, which refers to the temperature error, between the desired target set-point Temperature T_{sp} and the building temperature B_k at time k, e_k is the energy consumption at time k, and w_e $(0 \leq w_e \leq 1)$ is a weight trading the contribution of ΔT_k.

Figure 1 shows a graphical model representing how the next state building temperature, B_{k+1}, is a function of current building temperature, B_k, the external environment temperature, E_k, and the temperature control signal from the heat-pump, C_k, at time k. The system dynamics is expressed in the following equation:

$$\begin{aligned} B_{k+1} &= B_k + (E_k - B_k)\frac{\Delta t}{SCAP_e} + (C_k - B_k)\frac{\Delta t}{SCAP_i} \\ &= B_k(1 - \frac{\Delta t}{SCAP_e} - \frac{\Delta t}{SCAPi}) \\ &\quad + E_k\frac{\Delta t}{SCAP_e} + C_k\frac{\Delta t}{SCAP_i}, \end{aligned} \tag{3}$$

where Δt is the time interval, and $SCAP_e$ and $SCAP_i$ are the external and internal system thermal capacity.

The goal of the controller agent is to then identify a policy, $\pi(s, a)$ that determines what action, $a = C_k$, to select in each state, $s = (B_k, E_k)$, in order to minimize energy consumption while also minimizing divergence from the set-point target temperature, as described below.

3.2 Methodology

In this paper we refer to two general (and not mutually-exclusive) approaches to formulating and solving decision-making problems: *planning* and *learning*. In planning, which includes the framework of Model Predictive Control [5, 6, 8], it is assumed that a complete model of the task environment is available and the planner induces a policy for choosing the

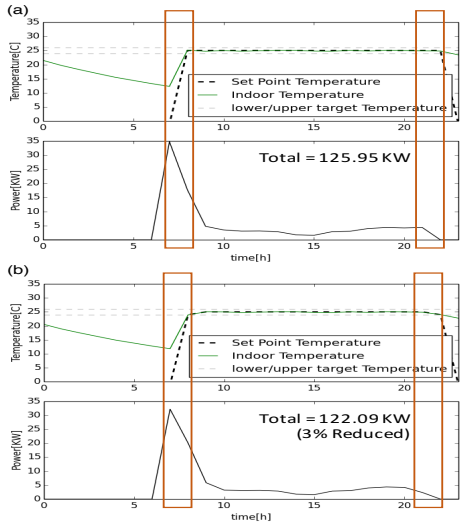


Figure 2: Set-back control scenario with standard and adaptive set-points; (a) standard set-point target; (b) adaptive set-point target.

action in each state that achieves optimal performance in terms of total long term reward. The learning approach, on the other hand, does not assume the environment is known ahead of time. Instead, the learning agent has to interact directly with the environment to gather data about the effects of actions on the world and their reward value, and while doing so searches for an optimal policy for action that maximizes long term reward. This is the classic setting of reinforcement learning [10, 11, 12, 13, 9].

The learning framework provides a general approach solving sequential decision-making problems without relying on a pre-existing model of the task environment, but incurs the cost of online interaction with the environment, which in some circumstances may be prohibitively high. We can potentially get the advantages of both approaches through a hybrid approach in which we use a suitable simulated environment for offline training. In this case, the initial policy learned in simulation is used to "bootstrap" the learner to reasonable performance that is then transferred to and fine-tuned in real-world interaction, often reducing the amount of costly real-world experience required to achieve high performance [12, 14].

In this work we seek to optimize two potentially conflicting goals, achieving target temperature while also minimizing power consumption. As has been demonstrated previously, a good way to reduce overall power consumption is to adopt a "set-back" strategy in which the set-point controller is only engaged during periods when the temperature-controlled area is being used [15]. However, because this method allows the building temperature to drift uncontrolled during the set-back period, it is possible for the desired target set-point temperature be very far from the set-back state when it comes time to reengage temperature control. A controller that only optimized for minimizing divergence from the set-point temperature risks expending an enormous amount of energy to immediately achieve the set-point goal.

Figure 2 demonstrates a scenario with set-back and set-point control regions. The orange outlined regions highlight the time interval of the transition from set-back to set-point. The green line represents indoor temperature,

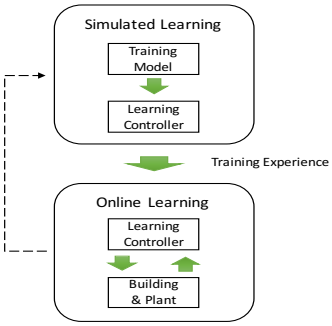


Figure 3: Schematic of combined simulation and real-world reinforcement learning.

the black dashed-line represents the target set-point temperature, which ramps up in the transition from set-back to set-point phases, and the horizontal gray-dashed lines represent the window of acceptable target temperatures. As we detail below, Figure 2(a) shows the results of standard set-back control, while Figure 2(b) shows that by "smoothing" the set-point state, power consumption may be reduced.

In the following sections, we explain our overall method, first describing our simulated reinforcement learning method for inducing an MPC prior policy, followed by how we formulate the adaptive set-point temperature goal.

3.2.1 *Simulated Reinforcement Learning: Hybrid Control Scheme with Model Predictive prior*

In order to minimize the amount of expensive real-world experience required by the relatively slow Q-learning process, we provide a simulated reinforcement learning environment to learn an initial, if noisy, policy [12, 14]. Figure 3 shows a schematic of the overall learning process.

In the Simulated Learning phase, the learning controller is trained by a simulator to learn a preliminary model. The simulator is less accurate than the real world, but by adapting the policy to the simulation approximation, we can reduce the online training time needed to achieve comparable performance. In the Online Learning phase, the learning controller is embedded in and interacting with the actual environment and continues to improve the performance of the policy. By interacting directly with the environment in this phase, the policy can adapt to specific features of the environment dynamics not represented in the simulation.

In our work, the simulation dynamics are represented in Eq. 4. The policy $\pi_0(s,a)$ is modeled as

$$\pi_0(s,a) = \frac{SCAP_i}{\Delta t}\left(T_{k+1} - B_k(1 - \frac{\Delta t}{SCAP_e} - \frac{\Delta t}{SCAP_i}) - E_k\frac{\Delta t}{SCAP_e}\right). \quad (4)$$

The model predictive prior $\pi_0(s,a)$ replaces the whole training phase to provide the initial policy model as the preliminary policy model.

3.2.2 *Adaptive set-point Temperature*

The set-back strategy [15] activates the heat pump only during specified intervals. During these intervals, the set-point is the target temperature. As the bottom plot of Figure 2(a) shows, the predominant power usage occurs during the transition from set-back to set-point state (the interval in the orange rectangle). Here we have two competing goals:

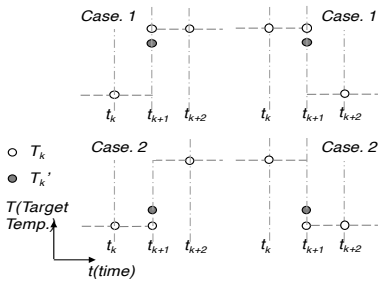


Figure 4: Cases for calculating target set-point temperature.

maintain as close to thermal comfort factor (set-point) as possible, but also minimize power use. We propose to accomplish both goals by smoothing the transition between the set-back and set-point phases. Figure 4 presents a schematic representation of two conditions, depending on whether the current temperature is above or below the set-point temperature. The target set-point temperature T'_k is determined by the previous (T_{k-1}), current (T_k), and next (T_{k+1}) set-points. The transition between states is reduced by the updated set-point temperature T'_k. We call this the adaptive set-point temperature (ASPT) method. The next step is to determine by how much to adjust the target set-point, T_{adj}. Here we use a heuristic that bases the adjustment on the observation that a comfortable range of temperature variation is within $\pm 1^{\circ}$C [16]. Thus, $T_{adj} = 1$.

Based on this, the updated current set-point temperature T'_k is determined as follows:

$$T'_{k+1} = \begin{cases} T_k - T_{adj}, & Case1 \\ T_k, & \text{no change} \\ T_k + T_{adj}, & Case2 \end{cases} \quad (5)$$

where

$$\begin{aligned} Case1: \quad & [(T_k < T_{k+1}) \& (T_{k+1} = T_{k+2})] \text{ or} \\ & [(T_k = T_{k+1}) \& (T_{k+1} > T_{k+2})] \\ Case2: \quad & [(T_k > T_{k+1}) \& (T_{k+1} = T_{k+2})] \text{ or} \\ & [(T_k = T_{k+1}) \& (T_{k+1} < T_{k+2})]. \end{aligned}$$

Using this model, the dominant power usage is reduced under considering the thermal comfort factor. Fig. 2(b) shows the result of applying this method to the same scenario as in Fig. 2(a) gaining a 3% savings in energy efficiency.

3.2.3 Algorithm

The complete model predictive prior reinforcement learning (MPPRL) with the adaptive set-point temperature (ASPT) algorithm is given in Algorithm 1. The algorithm adopts the model predictive prior to initialize the preliminary policy model π_0 and then runs the learning agent in the real world to do the online training. In the learning process, the set-point temperature is dynamically determined by the action a of the ASPT method to achieve energy efficient control during transitions. The evaluation function is the action value function $Q_t(s_t, a)$, which is determined by the reward function $R(s_t, a, s_{t+1})$ and state value function $V(s_t + 1)$. The γ is the learning rate. At each time, the optimal action a_{max} of the max quality value $Q_t(s_t, a)$ is selected to maximize the state value $V(s_t + 1)$. The policy model π is continuously updated by the optimal action a_{max} to adapt to environment changes.

Algorithm 1

```
1: procedure MPPRL
2:     Input: π0(s, a)
3:     for each t in active_time do
4:         calculate Action_set based on ASPT
5:         for each a in Action_set do
6:             Qt(st, a) = R(st, a, st+1) + γV(st + 1)
7:         end for
8:         V(st+1) = max_{a_max} Qt(st, a)
9:         π(st, a_max)
10:    end for
11: end procedure
```

4. SIMULATION RESULTS

Here we compare the performance of the proposed reinforcement learning agent in the MPPRL algorithm to a Model Predictive Control (MPC) and standard reinforcement learning (std RL) agent [10].

4.1 Simulation Setup

In the experiment, the simulated building temperature dynamics is modeled by the thermal model of Equation 3. The parameters of external ($SCAP_e$) and internal ($SCAP_i$) system thermal capacity are set as 9.896 and 2.441, based on [15]. The set-point temperature of the building is 25°C during hours 8 to 22, when the inhabitants are in the building. The heat pump changes its power set-point temperature every hour with 60 discrete actions ($0 \sim 60^{\circ}$C).

The heat-pump thermostat is assumed to be equipped with sensors to measure the environment temperature; in our model, these measurements are based on the TMY3 Tucson International Airport dataset [17]. This simulation assumes two cases of the thermal model for the building: an ideal case and a practical case. In the ideal case, we assume the thermal model equation 3 is exactly the same as true thermal dynamic behavior of the building. In the practical case, errors are introduced between the equation 3 and the true behavior of the building; in our experiment, these errors were set to be 20% above in external ($SCAP_e$) and 20% below in internal system thermal capacity ($SCAP_i$).

In the proposed learning agent, the Q-learning learning rate γ (line 7 of Algorithm 1), and the weighting of the energy consumption w_e of Equation 2, were both set to 0.8. The standard learning agent uses the same learning rate but there is no energy consumption term in the reward function.

4.2 Evaluation Results

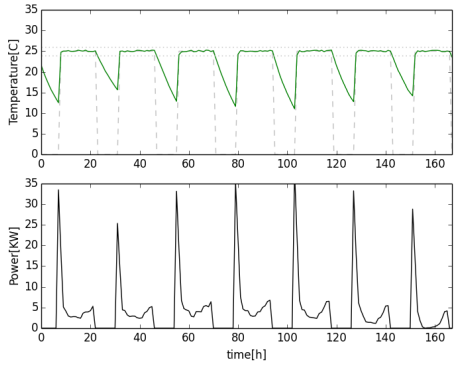
The experiments evaluated the temperature variation as the thermal comfort factor ($C.F.$) and the power consumption, as measured by energy E, of the three control strategies: MPC, std RL, and MPPRL. Smaller $C.F.$ means smaller temperature variation and implies that human inhabitants are more comfortable.

The quality of initial model of the MPPRL is an important factor of the performance. If the initial model is closer to the behavior of the real system, MPPRL will have the better performance. In order to have the same baseline in the evaluation between MPC and MPPRL, we set the predictive model of MPC as the same as the initial model of MPPRL.

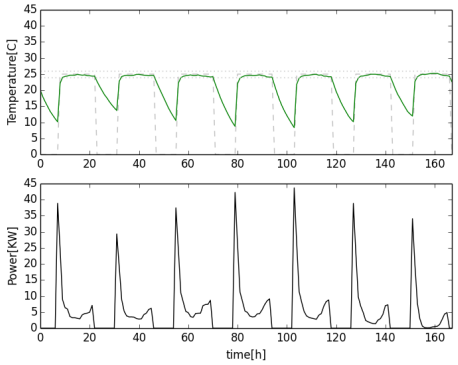
The evaluation considers the ideal and practical cases in

Metrics	Cases	Winter			Summer		
		MPC	Std RL	Proposed RL	MPC	Std RL	Proposed RL
C.F.[°C]	Ideal Case	0.1	0.12	0.27	0.09	0.16	0.31
	Practical Case	0.59	0.32	0.37	0.33	0.19	0.28
P [KW]	Ideal Case	774	769.25	725.5	323	314.3	287.5
	(comparing to MPC)	100.00%	99.39%	93.73%	100.00%	97.31%	89.01%
	Practical Case	962.8	906.2	889.3	389.6	383.7	355.6
	(comparing to MPC)	100.00%	94.12%	92.37%	100.00%	98.49%	91.27%

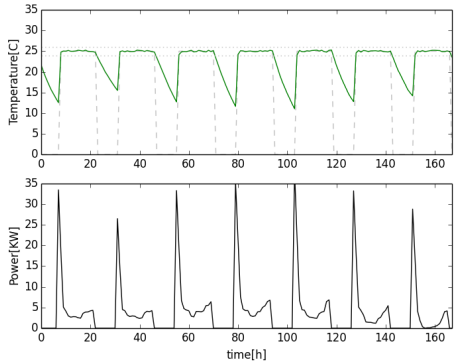
Tabel 1: Summary of Performance Evaluation of MPC, std RLC, and proposed RLC



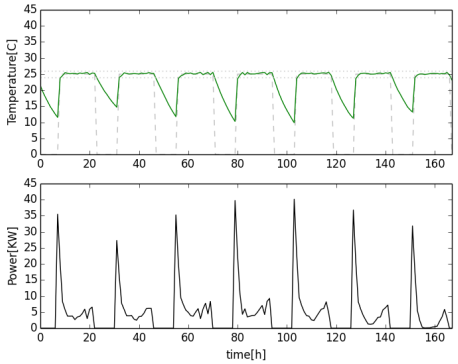
(a) MPC: C.F. = 0.1°C, P = 774.0 kW



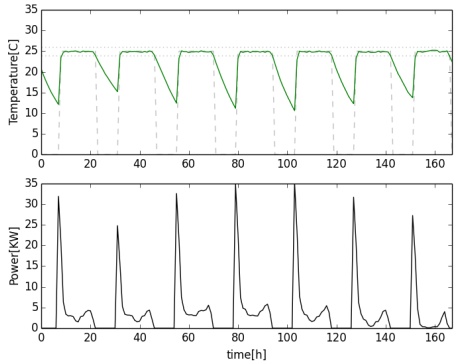
(d) MPC: C.F. = 0.59°C, P = 962.8 kW



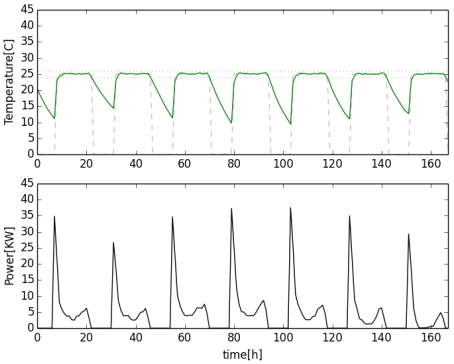
(b) Std RL: C.F. = 0.12°C, P = 769.3 kW



(e) Std RL: C.F. = 0.32°C, P = 906.2 kW



(c) Proposed RLC: C.F. = 0.27°C, P = 727.5 kW



(f) Proposed RLC: C.F. = 0.37°C, P = 889.3kW

Figure 5: The performance evaluation in a winter week under the ideal (left column, (a), (b) and (c)) and practical (right column, (d), (e), and (f)) thermal model cases.

winter and summer days separately. Table 1 shows the experimental results for the three controllers. Comparing to the MPC, MPPRL reduced power consumption by 6.3% during the winter and 11% during the summer in the ideal case. In the practical case, the saving rate is approximately 7.6% during the winter and 8.7% during the summer.

In the ideal case, the comfort factor (*C.F.*) of MPC is the optimal performance of 0.1°C in winter and 0.9°C in summer. Std RL provides a near optimal *C.F.* of 0.12°C for winter and 0.16°C for summer. MPPRL keeps the *C.F.* about 0.3°C. The profiles of the building temperature and the power consumption of a week in the winter are shown in Figures 5(a)-(c).

In the practical case, the learning agents are able to adapt to the noise better than MPC, which is based on the prior assumed model. The *C.F.* of MPC changes are larger and even worse than the learning agents. The power consumption of the MPPRL algorithm is lower than MPC by about 7.6% in the winter and 8.7% in the summer. The profiles of the building temperature and the power consumption of a week in the winter are shown in Figures 5(d)-(f).

The building temperature profile of the MPPRL is smoother than that of the MPC and std RL controllers in the region of the state change. This reduces the most significant part of the power consumption of each day.

5. CONCLUSION AND FUTURE WORK

We have proposed and evaluated a reinforcement learning controller for a heat-pump thermostat with a model predictive prior and incorporated an adaptive set-point temperature heuristic. The proposed method combines the strengths of two controller synthesis method, gaining the adaptivity of reinforcement learning while reducing online training cost through the use of a prior policy induced in offline simulation [7, 12, 14]. We demonstrated that the adaptive set-point temperature reduces the most significant part of power consumption, in the transition interval between set-back and set-point control. The proposed learning agent is an efficient and robust controller for a heat-pump thermostat of a building. We are currently working on extending the adaptive set-point heuristic to determine the degree of set-point adjustment as part of optimization, and are also extending the MPPRL control framework to distributed, multi-grid energy system optimization.

6. ACKNOWLEDGMENTS

This research was supported under a University of Arizona WEES Renewable Energy Networks Faculty Exploratory Grant (5825479). We thank our collaborators, Shane I. Smith, Christopher Lasch and Pierre Lucas.

7. REFERENCES

[1] U.S. Energy Information Administration (EIA). EIA online statistics. Technical report, http://www.iea.org/topics/electricity/, 21 May 2015.

[2] P. Bayer, D. Saner, S. Bolay, L. Rybach, and P. Blum. Greenhouse gas emission savings of ground source heat pump systems in europe: A review. *Reneable & Sustainable Energy Reviews*, 16:1256–1267, 2012.

[3] P. H. Shaikh, N. Bin Mohd Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim. A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renewable and Sustainable Energy Reviews*, 34:409–429, 2014.

[4] A. I. Dounis and C. Caraiscos. Advanced control systems engineering for energy and comfort management in a building environment–a review. *Renewable and Sustainable Energy Reviews*, 13(6):1246–1261, 2009.

[5] M. Morari and J. H. Lee. Model predictive control: Past, present and future. *Computers & Chemical Engineering*, 23(4):667–682, May 1999.

[6] J. Maciejowski. *Predictive Control with Constraints*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[7] D. Ernst, M. Glavic, F. Capitanescu, and L. Wehenkel. Reinforcement learning versus model predictive control: A comparison on a power system problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):517–529, 2009.

[8] A. Afram and F. Janabi-Sharifi. Theory and applications of hvac control systems–a review of model predictive control (mpc). *Building and Environment*, 72:343–355, 2014.

[9] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. Cambridge: The MIT Press, 1998.

[10] G. P. Henze and J. Schoenmann. Evaluation of reinforcement learning control for thermal energy storage systems. *HVAC&R Research*, 9(3):259–275, 2003.

[11] K. Dalamagkidis, D. Kolokotsa, K. Kalaitzakis, and G. S. Stavrakakis. Reinforcement learning for energy conservation and comfort in buildings. *Building and Environment*, 42(7):2686–2698, 2007.

[12] S. Liu and G. P. Henze. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 1. theoretical foundations. *Energy and Buildings*, 38(2):142–147, 2006.

[13] S. Liu and G. P. Henze. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory: Part 1. theoretical foundations. *Energy and Buildings*, 38(2):147–161, 2006.

[14] M. Cutler, T. J. Walsh, and J. P. How. Real-world reinforcement learning via multifidelity simulators. *IEEE Transactions on Robotics*, 31(3):655–671, 2015.

[15] F. Ruelens, S. Iacovella, B. J. Claessens, and R. Belmans. Learning agent for a heat-pump thermostat with a set-back strategy using model-free reinforcement learning. *Energies*, 8(8):8300–8318, August 2015.

[16] J. F. Nicol and M. A. Humphreys. Adaptive thermal comfort and sustainable thermal standards for buildings. *Energy and Buildings*, 34(6):563–572, 2002.

[17] National Renewable Energy Laboratory. National Solar Radiation Data Base. Tucson International AP [Excel file]. Technical report, http://rredc.nrel.gov/solar/old_data/nsrdb/1991-2005/tmy3/by_state_and_city.html, 13 April 2016.