Demand-Side Management of Domestic Electric Water Heaters Using Approximate Dynamic Programming

Khalid Al-jabery, *Student Member, IEEE*, Zhezhao Xu, Wenjian Yu, *Senior Member, IEEE*, Donald C. Wunsch, II, *Fellow, IEEE*, Jinjun Xiong, *Member, IEEE*, and Yiyu Shi, *Senior Member, IEEE*

***Abstract*—In this paper, two techniques based on *Q*-learning and action dependent heuristic dynamic programming (ADHDP) are demonstrated for the demand-side management of domestic electric water heaters (DEWHs). The problem is modeled as a dynamic programming problem, with the state space defined by the temperature of output water, the instantaneous hot water consumption rate, and the estimated grid load. According to simulation, *Q*-learning and ADHDP reduce the cost of energy consumed by DEWHs by approximately 26% and 21%, respectively. The simulation results also indicate that these techniques will minimize the energy consumed during load peak periods. As a result, the customers saved about $466 and $367 annually by using *Q*-learning and ADHDP techniques to control their DEWHs (100 gallons tank size) operation, which is better than the cost reduction that resulted from using the state-of-the-art ($246) control technique under the same simulation parameters. To the best of the authors' knowledge, this is the first work that uses the approximate dynamic programming techniques to solve the DEWH's load management problem.**

***Index Terms*—Approximate dynamic programming (ADP), load management, machine learning, Markov processes, power demand, smart grids, unsupervised learning.**

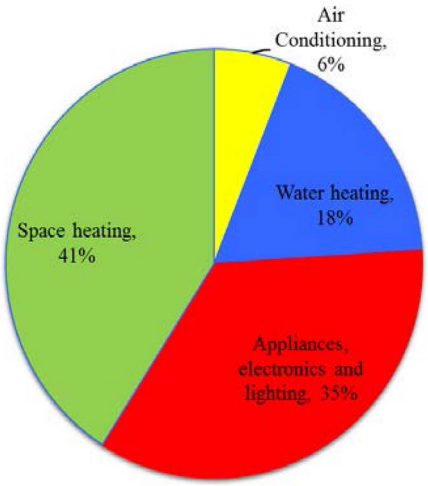


Fig. 1. House hold energy use distribution in the U.S. (Aug. 2013) [1].

I. Introduction

THE IMPORTANCE of domestic electric water heaters (DEWHs) can be seen from its effect on the overall grid load and energy consumption. For example, in the U.S. the average energy consumed by DEWHs is 18% as shown in Fig. 1. This share did not change during the last decade according to the U.S. Energy Information Administration [1]. The average annual cost of energy consumed by each DEWH is about $500 according to the office of energy and renewable energy [2]. Previous and current researches show that the total energy consumed in a city is highly dependent on the amount of power that the DEWHs consume [3]–[7]. For example, in the city of Quebec, Canada, the peaks in the grid load depends on the water heaters load, as illustrated in Fig. 2. The data plotted in Fig. 2, are from the field studies performed in two different cities [3], [4]. There is a clear relationship between the peaks in the grid load demand and those in the energy consumed for heating water in London and Quebec City. As a result, governmental agencies, policy makers, and of course energy companies, focus on domestic energy consumption with high priority for the water heaters. On April 30, 2015, president Obama signed into law S535, The Energy Efficiency Improvement Act, which established a new product category for large-capacity (75 + gallons) electric resistance "grid-enabled" water heaters for residential demand-response applications. Before that in the 15th of April, the Department of Energy provided new rule that requires all large capacity (55+ gallons) DEWH

Manuscript received February 14, 2016; revised May 23, 2016; accepted July 8, 2016. Date of publication August 8, 2016; date of current version April 19, 2017. This work was supported in part by the National Science Foundation, in part by the Missouri University of Science and Technology Intelligent Systems Center, in part by the Mary K. Finley Endowment, in part by the University of Missouri Research Board, and in part by the NSFC under Grant 61422402. This paper was recommended by Associate Editor X. Li. *(Corresponding authors: Donald C. Wunsch, II; Yiyu Shi.)*

K. Al-jabery, D. C. Wunsch, II, and Y. Shi are with the Electrical and Computer Engineering Department, Missouri University of Science and Technology, Rolla, MO 65409 USA (e-mail: kkatnb@mst.edu; dwunsch@mst.edu; yshi@mst.edu).

Z. Xu and W. Yu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhezhaoxu@gmail.com; yu-wj@tsinghua.edu.cn).

J. Xiong is with IBM Thomas J. Watson Research Center, Yorktown, NY 10598 USA (e-mail: jinjun@us.ibm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2016.2598563

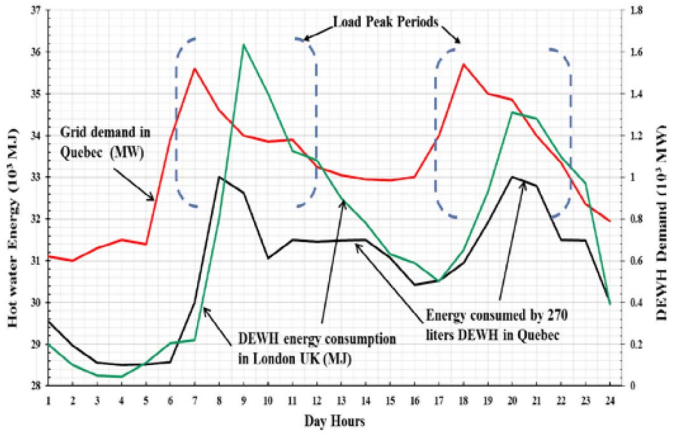


Fig. 2. Similarity in load peak periods between grid load demand and energy consumed by DEWH in the city of Quebec, Canada, and London, U.K. [2].

would have to be integrated electric heat pump water heaters. The current trend of research according to the peak load management agency is to design efficient grid enabled water heater, which is exactly what we have produced in this paper [8]. The attention to water heaters and load management is not only in the U.S. According to the reports of the Ministry of Power in India, DEWHs consumes nearly 23% of the electricity in the domestic sector [9].

The DEWHs are often selected for demand side management projects because both their load profiles and their average daily load profiles almost follow the same pattern as shown in Fig. 2. Furthermore, DEWH loads are easier to control than other domestic appliances, because of their energy storage ability. Although many researches have been conducted on DEWHs, most of them failed to be widely applied to the DEWH industry for various reasons. There are different demand side management strategies for controlling DEWH loads. In 1998, Nehrir *et al.* [6] introduced a fuzzy logic controller that can shift the DEWH load outside the peak demand period. Some of the suggested approaches significantly affected the temperature of the DEWH's output water, resulting in customer dissatisfaction, plus the complex modeling process it needs [7]. In 2007, Atwa *et al.* [10] used Elman neural network to control the power consumed by water heaters. Some researchers, used detailed analytical methods for modeling the DEWH and provided control strategies based on dividing the load into groups and control them through the thermostats [10]–[17]. In 2011, Moreau [3] described control strategies aimed at distributing and shifting DEWH's operation to within one or two hours outside the peak periods. The new coming technology in water heaters industry is the use of electric heat pump and gas condensing technology for water heaters with tanks that are larger than 55 gallons [18]. However, there are three concerns raised on this new technology, first it needs new installations, second it tends to be used in water heaters with large tanks only. The third concern and the most important one is low temperature operation when the heat pump water heater operates in electric resistance mode, it does not save energy or money compared to a conventional unit [19]. Even if the new heat pump water heaters dominated the market, the approaches demonstrated in this paper still can be used to improve the performance because they are designed to adapt with the user activities, grid load and the temperature of the output water as discussed next.

In this paper, we used action dependent heuristic dynamic programing (ADHDP) [20] and Q-learning approaches to solve the DEWH management problem, which is a multiobjectives optimization problem. The objectives of optimization are: minimizing the total cost of the energy consumed, reducing the load demand during peak periods, and achieving customer's satisfaction. The Q-learning algorithm and the ADHDP approach are both approximate dynamic programming (ADP) techniques [20], [21]. It should be pointed out that this paper is the first study using ADP techniques in the DEWH's load management.

The novelty of this paper lies in the way that the system is modeled. Three factors were used to define and control a DEWH: 1) the temperature of the water delivered to the customers; 2) instantaneous hot water consumption; and 3) estimated grid load demand (i.e., instantaneous energy price). In the Q-learning based approach, the three factors are considered linguistic variables. They are categorized as either "high," "medium," or "low." The problem was modeled as a semi-Markov decision process (SMDP) with two possible actions in each state: 1) "ON" and 2) "OFF." Specific fuzzy rules are used to determine the system's current state. Each DEWH is considered to be an artificial agent that was trained to adapt to the diversity within a user's consumption profile and grid load demand. The agent learns how to adapt, in the Q-learning approach, after finding the final Q-factors that specified the operating's policy during real time operation. In the ADHDP approach, the adaptation process consists of two phases: 1) critic training and 2) action training. The system learns to estimate the correct cost value during training the critic network, and then uses that cost in training the action network. The system tries to minimize the cost by adapting the weights of the action network [20].

These techniques can be applied to any DEWH, regardless of its capacity, heating elements, or operating environment. Furthermore, according to the simulation results, the approaches are able to reduce the energy consumed by DEWH more than the existing state-of-the-art methods [3]. The experiments show that the Q-learning and ADHDP controllers have reduced the cost of the energy consumed by the DEWHs by approximately 26% and 21%, respectively when using large (100 gallon tank) DEWH. As a result, both techniques will save about $466 and $367 per year, respectively, for the customers who use them to control their DEWH's operation. In comparison, only $246 would be saved annually by using the state-of-the-art control strategy [3], as illustrated in Table III.

II. System Modeling and the Approximate Dynamic Programming

ADP techniques have been used effectively in solving optimization problems that consist of sequences of control actions whose efficiency remains unknown until the end of sequence. For the demand side management problem of DEWH,

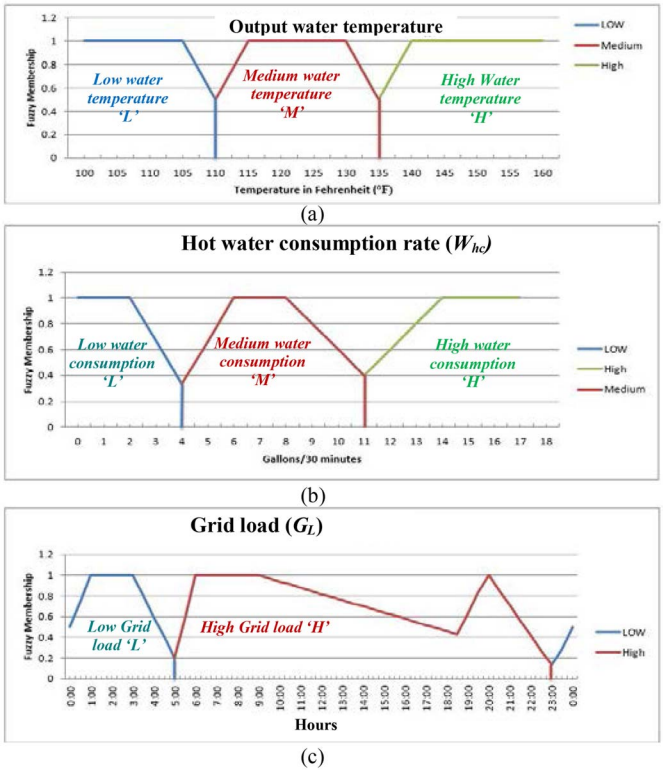


Fig. 3. Illustrations of the fuzzy membership functions. (a) $f_{z,1}(\cdot)$ for T_h. (b) $f_{z,2}(\cdot)$ for $W_{\rm hc}$. (c) $f_{z,3}(\cdot)$ for G_L, where G_L is a function of time in hour.

two ADP techniques: 1) ADHDP and 2) Q-learning (which is a special case of ADHDP) [20], were considered. The system modeling, and the training and controlling processes of both techniques are explained in the following sections.

A. System Model

The DEWH model is defined by three variables: *the output water temperature* (T_h), *hot water consumption rate* ($W_{\rm hc}$), and *the grid load* (G_L). $W_{\rm hc}$ is generated randomly (see Section III-C), T_h is calculated using (2) and (3) based on the selected action, and G_L depends on the city grid load profile. Therefore, the three variables are not correlated. However, the values of these variables were used differently for the two approaches presented in this paper.

In the Q-learning-based approach, the variables were converted into linguistic values using fuzzy membership functions, as illustrated in Fig. 3. The variables (T_h and $W_{\rm hc}$) have three possible values (low "L," medium "M," or high "H") while G_L has only two possible values low or high. [This assumption was based on the time-of-use (ToU) pricing profile that is used in this paper where there is no medium grid load or in other word medium power cost also it is meaningless practically to describe grid load as medium.]

Therefore, a discrete state system can be defined with $3 \times 3 \times 2 = 18$ different states. The demand side management problem is to decide whether to turn the DEWH ON or OFF at each event time, which refers to the time when user consumes any quantity of water from the DEWH's tank. In practice, the numeric value of each variable should be fuzzified and mapped to its corresponding linguistic value. The fuzzy membership functions $f_{z,i}(\cdot)$, $(i = 1, 2, 3)$ are defined for the three variables, and illustrated in Fig. 3. For variables T_h and $W_{\rm hc}$, values of L, M, and H correspond to 0, 1, and 2, respectively. For variable G_L, the values of L and H correspond to 0 and 1, respectively. The water temperature thresholds were specified based on the fact that legionella bacteria begin to die at temperatures above 120 °F [22].

TABLE I
SYSTEM STATES ENCODING (L: LOW = 0, M: MEDIUM = 1, AND H: HIGH = 2)

S_i	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}	S_{17}	S_{18}
T_h	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
W_{hc}	L	L	L	M	M	M	H	H	H	L	L	L	M	M	M	H	H	H
G_L	L	L	L	L	L	L	L	L	L	H	H	H	H	H	H	H	H	H

Suppose v_1, v_2, and v_3 are the numeric values of the three variables, respectively, and $S(v_1,\ v_2,\ v_3)$ is the corresponding state's index number. Then

$$S(v_1, v_2, v_3\ ) = \sum_{i=1}^{3} \left[3^{i-1} f_{z,i}(v_i)\right] + 1. \tag{1}$$

The system states are encoded as listed in Table I.

Equation (1) is used to determine the system's current state during the training phase. It is used during the simulation as well. The linguistic values are vital for calculating the immediate reward during training (see Section III). Actions are selected randomly with equal probability during the training phase in order to provide stochastic value iterations and update the Q-factors accordingly, as discussed in Section III-B.

Variable T_h's numeric value at time $(t + 1)$, denoted by $v_1(t+1)$ can be estimated based on the action decision made at time (t). From the law of energy conservation [23], [24], we derive

$$v_1(t+1)|_{a(t)=1} = \frac{9P\tau}{5K_jV} \cdot \frac{m_1 C_{\rm hw} T_h(t) + [C_{cw}T_c - C_{hw}T_h(t)] \cdot m_2(t)}{m_1 C_{\rm hw} + (C_{\rm cw} - C_{\rm hw}) \cdot m_2(t)} + 32 \tag{2}$$

$$v_1(t+1)|_{a(t)=2} = \frac{9}{5} \cdot \frac{m_1 C_{\rm hw} T_h(t) + [C_{cw}T_c - C_{\rm hw}T_h(t)] \cdot m_2(t)}{m_1 C_{\rm hw} + (C_{\rm cw} - C_{\rm hw}) \cdot m_2(t)} + 32 \tag{3}$$

where $v_1(t)$ is the current temperature of the water, $a(t)$ is the current action, 1 means ON, and 2 means OFF. m_1 is the total mass of water in the DEWH tank, $m_2(t)$ is the mass of water consumed at the time (t), and $m_2(t) = v_2(t) \times 3.785$ [25]. $C_{\rm hw}$ and $C_{\rm cw}$ represent the heat capacity of hot water and cold water, respectively [23]. T_c is the temperature of the water supplied to the DEWH, typically $10 \sim 13$ °C [26]. P is the power rating of the heating element (4500, 2800, or 36 000 Watts per hour in this paper). $K_j = 2.42$ W*h/gal*°F, which is the recovery rate calculation constant [27], [28], and V is the total volume of the DEWH tank. τ is the sampling period (30 min in this paper). $v_1(t)$ and $v_1(t+1)$ are in unit of °F. Heat dissipation and heat exchange between the DEWH

metal surface and air is negligible (*less than* 0.25 °C/h) [19]. These two equations are to estimate the water temperature using energy saving formula. According to the behavior and ranges of the calculated temperature at each time step in compare with field studies [3]–[5], the model was acceptable. However, accurate performance to these models have not presented in this paper. Due to the involvement of several random functions in profiles generation and the absence of real data.

An event driven simulator was designed to generate users' profiles. The simulator mimics the data that were collected in [3] and [4]. The distribution fitting toolbox in MATLAB was used in this paper to determine the random variable distribution. The designed event driven simulator generates the time of the events (which specifies the current grid load G_L) and the quantity of hot water used in each event W_{hc}. The linguistic value of the grid load factor (G_L) is determined based on the event time since previous studies have shown that there are specific periods during the day when the load demand becomes high [3]–[7], [13]. However, a ToU pricing profile is used to calculate the real cost of the consumed power [29], as illustrated in Section III-C. The instantaneous output water temperature, T_h, is calculated using (2) and (3) when the selected action at time (t) is OFF and ON, respectively. One of the advantages of the approach is that it avoids the complicated thermodynamic and heat transfer operations, which occur inside the DEWH's tank, by using (2) and (3) to estimate the value of T_h at time $(t + 1)$. Furthermore, this paper does not require any complex calculations such as that described in previous studies [7] to solve the optimization problem.

In the ADHDP approach, the system is modeled as a continuous state space system. The system's state is also defined by the same variables (T_h, W_{hc}, and G_L) used in the Q-learning approach, but there is no fuzzification/defuzzification process. The state variables are the inputs of the critic and the action neural networks in the ADHDP controller. Their normalized numeric values are used to train the neural networks. This will be explained in more detail in Section III-A.

B. Approximate Dynamic Programming

Approximate or adaptive dynamic programming (ADP), also known as the reinforcement learning, simulation-based dynamic programming, stochastic programming, and neuro-dynamic programming, refers to a group of algorithms designed to solve the problem of Markov and SMDPs given by [30]

$$J^*(i) = \max_{a\in A(i)} \left[\sum_{j=1}^{|S|} p(i,a,j)\big[r(i,a,j) + \lambda J^*(j)\big] \right] \tag{4}$$

where $J^*(i)$ is the ith element of the vector value function associated with the optimal policy. $A(i)$ is the set of all actions allowed in state i, $p(i, a, j)$ represents the transition probability of going from states i to j under the influence of action a. $r(i, a, j)$ is an immediate reward earned when action a is selected in state i and the system transfers to state j as a result. S represents the set of states in the Markov chain, and λ is the discounting factor.

Algorithm 1 Q-Learning

1 ***Set up the training parameter: imax, and initialize Q-factors = 0 and t = 0.***
2 ***Randomly select initial state and action (S_0,a_0).***
3 ***Repeat until (number of iterations > imax).***
- ***Apply action a(t) on DEWH model and read the current and the estimated values of vi(t) and vi(t+1), respectively.***
- ***Determine $S_{(t+1)}$ using*** (1).
- ***Calculate immediate Reward $r(S_t,a_t,S_{t+1})$***
- ***Update Total Reward R_t=R_t+ $r(S_t,a_t,S_{t+1})$.***
- ***Update: $t = t + 1$; $S_{(t-1)}$=$S_{(t)}$;$S_{(t)}$=$S_{(t+1)}$.***
- [a] ***Update learning rate: $\alpha = \alpha^{t+1}$.***
- ***Update Q-factors using*** (6).

4 ***Construct the final policy from (S, a) pairs with higher Q-factors using the following formula on each state (i): $P(i) = arg^{max}_{b\in A(i)} Q(i, b)$; P(i) is the policy at state (i) (i.e. action that lead to maximum reward on the long run).***
5 ***Record $\widehat{P}$(optimum policy for all states) and stop.***

[a] There are multiple ways to update the learning rate (α) [21]. In this work, we update α using: $\alpha^{t+1} = \frac{c_1}{(c_2+t)}$, where the positive constants c_1 and c_2 fulfils $c_1 < c_2$. (e.g. we set $c1 = 200$ *and* $c2 = 220$. *More discussion on learning rate selection can be found in [21].*

C. Q-Learning Algorithm

Watkins published the Q-learning algorithm in 1989. He defined this method as "a form of model free reinforcement learning and it can be viewed as a method of asynchronous dynamic programming" [31]. The Q-learning algorithm associates a scalar value, the Q-factor, with each state action pair. It solves (4) by updating the Q-factors associated with an optimal policy instead of approximating the cost function of a particular policy. Furthermore, it uses policy iteration (PI), as described in (5), to avoid the evaluation of multiple policies. The PI serves as the Q-factor version of the Bellman equation [21], [30]

$$Q(i,a) = \sum_{j=1}^{|S|} p(i,a,j)\left[r(i,a,j) + \lambda \max_{b\in A(j)} Q(j,b) \right] \tag{5}$$

where Q(i, a) and Q(j, b) are the Q-factors associated with state-action pairs (i, a) and (j, b), respectively.

Equation (5) still requires the transition probabilities. Therefore, the Robbins–Monro algorithm [32] was used to estimate the optimal Q-factors. The optimal Q-factors' estimation was achieved by expressing every Q-factor as an average of a random variable. Equation (6) represents the Q-factor version of the value iteration, which is the Q-learning algorithm for a discounted MDP. The derivation of (6) from (5) can be found in [30]

$$Q^{n+1}(i,a) = \left(1-\alpha^{n+1}\right)Q^n(i,a) + \alpha^{n+1}\left[r(i,a,j) + \lambda \max_{b\in A(j)} Q(j,b) \right] \tag{6}$$

where $\boldsymbol{\alpha}^{n+1}$ represents the adaptive learning rate and attenuating with time. $\boldsymbol{Q}^{n+1}$ is the updated Q-factor and $\boldsymbol{n}$ is the time step.

In this paper, the Q-learning version of the value iteration was used to solve the predescribed SMDP problem, which can be viewed as a discounted reward for the reinforcement

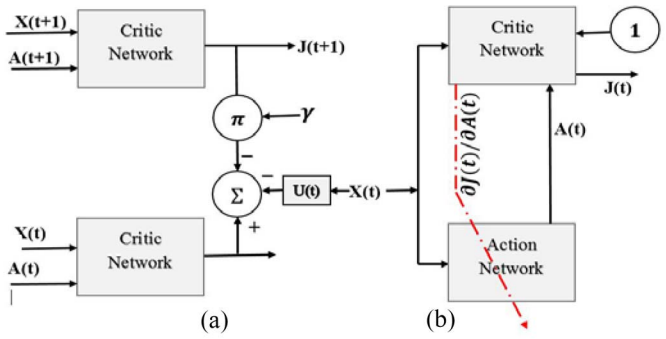


Fig. 4. (a) Critic adaptation in ADHDP/HDP. This is the same critic network in two consecutive moments in time. The critic's output $J(t+1)$ is necessary in order to give us the training signal $\gamma J(t+1) + U(t)$, which is the target value for $J(t)$. (b) Action adaptation. X is a vector of observables, and A is a control vector. We use the constant $\partial J/\partial J = 1$ as the error signal in order to train the action network to minimize J. This figure is adapted from [20].

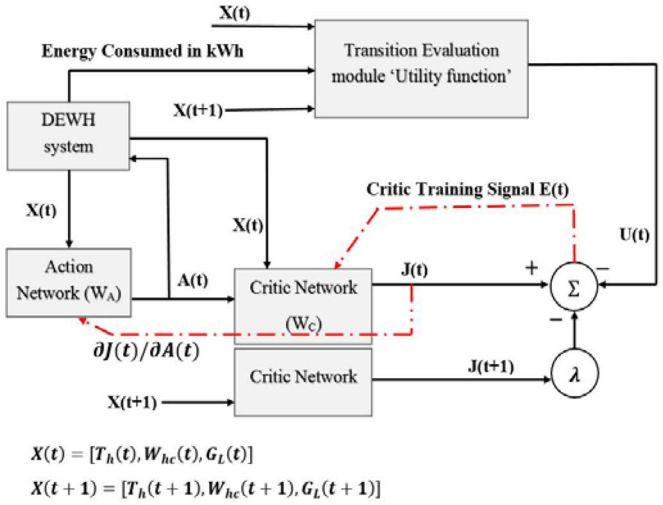


Fig. 5. Implemented ADHDP controller's structure. *X:* state, *J:* cost, *A:* action, any variable with (t) means current $(t+1)$ means next. $U(t)$: immediate transition cost "utility" and W_C and W_A: critic and actor networks weight matrices, respectively.

learning based on the stochastic value iteration. However, the Q-learning version used in this paper uses a specific cost function to generate the immediate cost/reward for each system state transition, as illustrated in Section III-A. This cost function eliminates the need for the transition reward matrix (TRM), which is usually used in Q-learning algorithms. The same cost function was used to evaluate the system's transition in the ADHDP approach as well (see Section II-D).

D. Action Dependent HDP

The family of adaptive critic design controllers has been presented by Werbos [33]. HDP, and its action dependent ADHDP forms, have a critic network that estimates the cost-to-go function J^* in (4) which calculates the Bellman equation of dynamic programming [20]. The standard structure of HDP and ADHDP is illustrated in Fig. 4. ADHDP is a generalization of Q-learning for the continuous domain system. In ADHDP, the critic is trained to provide an accurate estimation for the cost-to-go function J and to minimize the following error $E(t)$:

$$E(t) = J[X(t)] - \gamma J[X(t+1)] - U(t) \tag{7}$$

where $X(t)$ is the vector of observations/variables that define the system's current state.

The ADHDP controller adaptation process consists of two phases: 1) critic and 2) action networks training. These two processes are implemented continuously in sequence but not in parallel.

The critic network is designed to minimize a back propagated error signal (7), and the gradient of J with respect to the weights of the critic Wc is given by

$$\Delta W_c = -\eta_c[E(t)] \times \frac{\partial J}{\partial W_c} \tag{8}$$

where η_c is a positive learning rate ($0 < \eta_c = 1$). The action network is connected as shown in Fig. 4(b) in order to minimize J in the next time step and optimize the total cost over the entire domain. In the action network's adaptation phase, the gradient of J with respect to A (i.e., $\partial J/\partial A$) is back propagated, as illustrated in Fig. 4(b) and the following equation:

$$\Delta W_A = -\eta_a \times \frac{\partial J}{\partial W_A} \tag{9}$$

where $\partial J/\partial W_A = (\partial J/\partial A) \times (\partial A/\partial W_A)$ is the gradient of the cost-to-go function $\boldsymbol{J}$ with respect to the weights of the action network W_A. η_a is the action network learning rate [η_a does not have to be equal to η_c in (8)].

In HDP the immediate cost or the utility function $U(t)$ is approximated as well using neural networks, while in ADHDP, $U(t)$ is calculated using a model and the action network is connected directly to the critic [20]. The approaches described in this paper were designed to overcome the limitations presented by previous solutions. These improvements focused on the following.

1) Reducing the need for permanent communications and synchronizations between DEWHs and the smart grid infrastructure [7], [12].
2) Either shifting or eliminating the peaks with in the grid load [3].
3) Reducing power consumption during peak periods and as a result minimizing the cost of the power consumed without sacrificing customer satisfaction.

III. Training and Implementation

This section contains the discussion on the processes required to transform a normal DEWH into a smart appliance. The discussion clarifies and compares the technical implementation of the presented approaches: the ADHDP and the Q-learning.

A. ADHDP Implementation

The ADHDP controller is illustrated in Fig. 5. It consists of the following components.

1) The DEWH system module: which was described in (2) and (3) is the same module used during Q-learning

Algorithm 2 Utility Function

Calculate_cost(power in kwh, state: X(t)= [T_h, W_{hc}, G_L], t)

1 *if G_L is high*
  *J=energy in kwh*a; % penalty P1 {peak load}*
  if T_h< threshold
    $J = J + W_{hc}$; % increase cost {penalty}
  else: $J = J - b * W_{hc}$; % decrease cost {reward}
2 *else*
  if time (t) is between 3 and 5:30 am
   J = energy in kwh/a; % P2<P1 low load
  Else
    J= energy in kwh/b; % P1>P3>P2
  if T_h< threshold
    $J = J + W_{hc} * a$; % increase cost {penalty}
  else: $J = J - W_{hc}$; % decrease cost {reward}
3 *If Q_learning % see Section III.A*
  $J = -J$;
4 *Return J;*

and the final simulation (to be discussed further in Sections III-B and IV).

2) The critic network used in this paper consists of: one input layer with four neurons (for each state variable and the action network output), one hidden layer with 30 neurons each of which has a hyperbolic tangent sigmoid activation function, and one neuron in the output layer with a linear activation function. The output of the critic is $J(t)$, if the inputs were $X(t)$ and $A(t)$ or $J(t+1)$ if the inputs were $X(t+1)$ and $A(t+1)$.
3) The action network is almost identical to the critic network, but it only has three neurons in the input layer, and the sigmoid activation function is used in the hidden and also the output layer. The action network generates the control action (On or Off) during normal operation and simulation.

The transition evaluation module (or utility function) was designed to replace the TRM, as illustrated in Section II-C. This utility function (*in some literature called a cost function*) calculates the immediate transition reward/cost for each system's state transition. The same function was used in Q-learning as well. The utility function provides a more efficient evaluation than the TRM. The utility function calculates the immediate transition cost based on the energy consumed during the transition and all the other control variables ($T_h(t)$, $W_{\mathrm{hc}}(t)$, $G_L(t)$), as illustrated in Algorithm 2. The pseudo code of the cost function illustrated in Algorithm 2 is of major importance because it highly reduced the complexity of the training process for both ADHDP and Q-learning compared to previous work [34]. The utility function rewarded the agent for each gallon of output water supplied with $T_h > 120$ °F and penalized failure to do so. It also considers each consumed kWh as a penalty, but that penalty depends also on whether it was consumed during peak or normal load. If G_L is low, the penalty will be mitigated through dividing the kWh by (a) and vice versa, as illustrated in Algorithm 2. The guidance factors (a and b) provide control over the multiobjective optimization process (i.e., they encourage the agent to turn the heating element on during low load periods and to reduce the effect of the different scale between energy and output water units). Experimental results showed that: choosing a and b such that ($a \geq 2b, \forall b \geq 1$), provides better performance for the ADHDP controller. However, these factors have no effect at all on the Q-learning performance, as illustrated later in Table III.

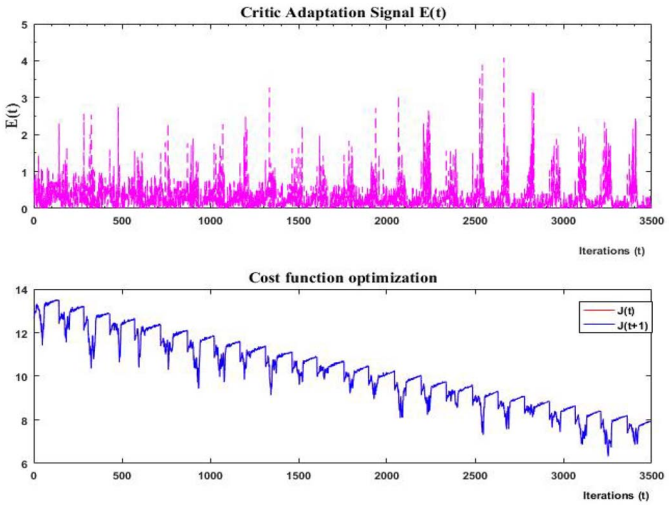


Fig. 6. ADHDP critic network adaptation. Top: instantaneous error signals during critic adaptation. Bottom: total cost reduction during critic adaptation.

The adaptation of the presented ADHDP controller was implemented in two phases.

1) *Offline Training:* In the offline training, the critic network was trained first using the data generated from the Q-learning algorithm. The critic training stops when the back propagated error signal from (7) becomes less than a prespecified small value, or when the training lasts for the maximum number of iterations. The critic training of the presented ADHDP is illustrated in Fig. 6. Furthermore, the action network is also trained during the offline phase using the same data used in critic adaptation. The size of the data sets depends on for how many simulation days Q-learning was trained, and each day contained about 150 samples. It was noted during experiences that repeating the offline training after the online training enhances the ADHDP controller's performance. The selection of the guidance factors has major effect on the ADHDP performance too, as illustrated in Sections V and VI.
2) *Online Training:* The online training is executed during the simulation phase. The action network keeps adapting during the simulation to minimize the system's cost to go (i.e., J). Simulation here is the same as real time operation, because of the use of the event driven simulator that explained in Section IV. The adaptation of the action network is illustrated in Section II-D.

The presented ADHDP controller showed better performance in cost reduction as the number of iterations increased. The critic adaptation is illustrated in Fig. 6 and was recorded during training for 100 simulation days with a total data set size of about 15 000 samples.

B. Q-Learning Implementation

The second optimization approach implemented in this paper is the Q-learning algorithm. The actions are selected

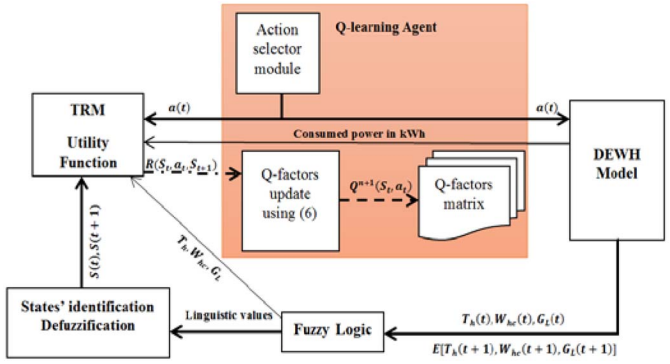


Fig. 7. Q-learning's schematic. $\mathbf{Q^{n+1}}$: new value of Q-factor, $\mathbf{R}(\boldsymbol{S_t}, \boldsymbol{a_t}, \boldsymbol{S_{t+1}})$: immediate reward due to system's transition from current state $\boldsymbol{S_t}$ to $\boldsymbol{S_{t+1}}$ using the current action a_t, $\boldsymbol{E}[\ldots]$ estimating the values of the enclosed factors, and Q-factors' matrix is an 18×2 matrix.

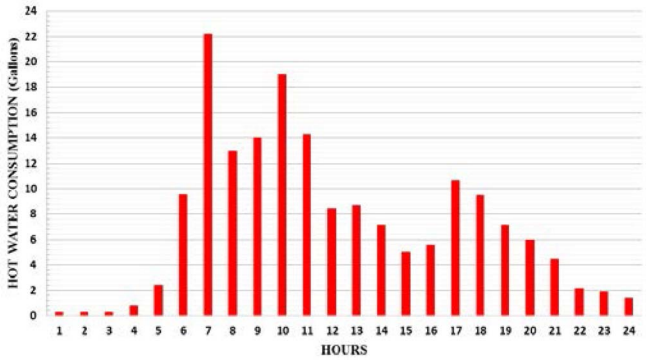


Fig. 8. Sample user profile generated from the event driven simulator. Where the profile generated using embedded code (see Section III-C).

randomly with equal probability at each time step to achieve better exploration of the solution space during the training phase. The algorithm then receives the control variables' estimated values at $(t+1)$ from the DEWH model and evaluates the performed action based on the utility function illustrated in Algorithm 2, which is the same utility function used in ADHDP. In step 3 from Algorithm 2, the calculated cost is negated. This is because Q-learning selects the optimum policy based on the Q-factor with the maximum value, unlike ADHDP which seeks to minimize the cost. The Q-factor associated with (S_t, a_t) was last updated using (6). This algorithm repeats the same procedures and continues until the maximum number of iterations are performed. The Q-factors have been stored in an (18×2) scalar matrix. The optimum policy is then derived, as illustrated in Algorithm 1 and Fig. 7.

Each iteration here represents a one-day simulation that contains about 150 time steps based on the event driven simulator used. The best value for the discount factor ë was derived heuristically as 0.9. Note that the same notation used in Fig. 5, to indicate the fact the same discount factor was used in ADHDP as well. The training phase for both of the presented approaches (ADHDP and Q-learning) was conducted using a DEWH's module with the following parameters.

1) The heating element power = 36, 4.5, 4.5, or 2.8 kWh.
2) Tank size = 120,100, 70, or 40 gallons, respectively.
3) Discount factor $\lambda = 0.9$.

The state's variables linguistic values were derived, as illustrated in Fig. 3, for Q-learning. The same specifications were used with in the simulation for all the other simulated approaches as well.

C. Data Processing

This section includes discussion on the process of generating the control variables' numerical values (T_h, $W_{\rm hc}$, and $G_{\boldsymbol{L}}$) and illustrates the *event driven simulator*.

The *event driven simulator* presented in this paper provided comparable results with those obtained from previous field studies [3], [7]. The simulator is designed to mimic human activities in consuming hot water. This simulation was required to provide a reliable assessment for the presented DEWH's control approaches. The simulator generates (*per simulation day*) unique and random profiles for each DEWH used in the simulation as shown in Fig. 8. The simulator assumes four occupants in each house, for simplicity we avoided considering the ages, gender, and other social factors for the occupants that may affect their hot water consumption rate. The generated profile shown in Fig. 8, is comparable to the profiles obtained from the field studies in the British Department for Environment, Food and Rural Affairs' report [4].

The simulation consists of two phases: 1) profile generation which was used in training and in performance evaluation or comparison and 2) comparator simulator, in which an evaluation process implemented among the presented approaches and the state-of-the-art approaches [3]. The profile's generator also provides the time of using the hot water and how much hot water was used. In the evaluation phase, different models for the DEWHs were used, uncontrolled "reference," scenarios 0, 1, and 2, Q-learning, and ADHDP. Each group of DEWH have the same number of DEWHs units, the same tank size (120, 100, 70, or 40 gallons) and heating element (36, 4.5, or 2.8 kWh). The simulator assumes that, we are creating six parallel universes or copies from every house dwelling and give each copy different brand (i.e., version) of DEWH. The performance of each group is evaluated during the simulation based how much energy each group can save with respect to the uncontrolled scenario (all scenarios operated for the same period of time and using the same user profile).

This criteria is used to guarantee a fair comparison among the different approaches. Otherwise, it is difficult to present an accurate comparison among the different control strategies. The user profiles which includes events' time indices and the consumed hot water quantities were generated using special combination of Poisson random variables. The choice of these variables is based on using the distribution fitting tool box from MATLAB on the previous studies' data. The profile generator function was adjusted empirically till it provides user profiles similar to the actual user profiles obtained by previous studies [3]–[5]. Artificial profiles needed due to the limitations of real data profiles. The numerical variables used for calculating the variables are explained as follows.

1) *Water Temperature:* In this paper, instead of diving deep inside the thermodynamic operations of the DEWH, we utilized the law of energy preservation to provide

a reasonable estimation for the temperature of the output water at the hot water faucet, as in (2) and (3). Since calculating the output water's exact instantaneous temperature is almost impossible without using an expensive embedded system to calculate the temperature of the DEWH's output water [23]–[25].

2) *Hot Water Consumption Rate:* Estimating or predicting any human activity is extremely difficult. This paper relied on statistics from field surveys, which have been performed in London, U.K., and Quebec City, Canada [3], [4]. However, to generate the required data, an embedded event driven simulator was designed, as illustrated in the previous section.
3) *Grid Load "Energy Cost":* The estimated instantaneous grid load can be obtained from the local utility companies, and they are time dependent, as illustrated in Fig. 3(c). As mentioned earlier the grid load characteristics used in this paper are based on data obtained from Quebec City and London [3], [4]. However, the numeric values of this factor are the time indices of the operation. The load peak periods occurred approximately between 5:30 and 10:00 A.M. and between 4:30 and 10:00 P.M. Furthermore, the final comparison was conducted using a ToU profile [29].
4) *Energy Consumed by the DEWH's Heating Element:* The amount of the consumed energy is calculated using the module described in (2) and (3). The values of the immediate energy consumption in kWh were used to calculate the value of the utility function at each system transition, as illustrated in Figs. 5 and 7.

The control variables' numerical values are normalized before being used as inputs for the action and the critic networks in the ADHDP controller. The same profiles generated during the *Q*-learning process were used in the ADHDP approach as well.

IV. Simulation and Evaluation

The simulation process was designed to provide the same operating conditions for all the simulated scenarios as discussed in the previous section. Five different approaches were simulated under the same operating conditions. These operating conditions are as follows.

1) The DEWH specifications as listed in Section III-B.
2) All DEWHs must supply water in a temperature higher than 120 °F.
3) A soft threshold specified to be 125 °F. This soft threshold was used to prevent the output water's temperature to fall below 120 °F for all the compared approaches [22] (*to maintain customers' satisfaction*). The DEWH heating element should be turned ON whenever the water temperature fell below the soft threshold. However, the simulator recorded even the quantities of water outputted to users below thresholds to provide more accurate evaluation to each of the control strategies, as illustrated in Tables III and IV.

The evaluation comparison is performed among five groups of DEWHs plus the uncontrolled operation as a reference.

TABLE II
Optimum Policies Selected by *Q*-Learning
(1 = On and 0 = Off)

S_t[a]	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}	S_{17}	S_{18}
a(t)[b]	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	1	0	0
a(t)[c]	1	1	1	1	1	0	1	1	0	1	0	0	1	0	0	1	0	0

[a] States (1-18)
[b] policy for 30 iterations
[c] policy for 100 iterations

Each group has the same number of identical DEWHs. The results in this paper were derived from simulating the operation of 100 DEWHs in each group. Any number of DEWHs can be used and from experiences no effect on the comparison. Since the same user profiles is being used for the different groups. In other words, the same group of DEWHs were simulated using five different control scenarios and the uncontrolled scenario. The final assessment was presented based on the percentage cost reduction of the consumed energy cost using a ToU pricing profile. The ToU profile gives three different prices for the kWh during the day: 5.62, 10.29, and 23.26 ¢ [29]. These prices were applied to the load profile of the city of Quebec that used in this paper and the state-of-the-art work that is compared with. As a result, the pricing profile that was used in our comparison simulator is as follows for the energy unit price.

1) 23.26 ¢ for each kWh consumed between 5:30 and 10:30 A.M. (load peak 1).
2) 10.29 ¢ for each kWh consumed between 4 and 9 P.M. (load peak 2).
3) 5.62 ¢ for each kWh consumed other times of the day (low grid load period).

The different scenarios presented in the state-of-the-art work in the field (i.e., scenarios 0, 1, and 2 in the list) [3], the uncontrolled scenario (i.e., scenario 4), and our scenarios (scenarios 3 and 6) are all discussed below.

1) *Scenario 0:* The demand pick-up at the end of the load shifting period is not controlled.
2) *Scenario 1:* The pick-up is controlled according to a prioritized random function that was spread over a range of 1 h after the peak period ended. In this scenario, the agent turn the heating element off during the peak period.
3) *Scenario 2:* The pick-up is controlled according to a prioritized random function that is spread over a range of 2 h after the peak period ended. The success of the simulator can be verified by looking at the energy consumption curves in Figs. 10–13.
4) *Scenario 3 (Q-Learning):* The entire operation of every DEWH in the group is controlled according to the policy selected after the presented *Q*-learning algorithm (also known as the trained group in the comparison charts), which is used to train the agent.
5) *Scenario 4 (Ref. or Uncontrolled Scenario):* The DEWHs that are simulated under this scenario are operating under no artificial control. The heating element is turned ON whenever the water temperature became less than the specified soft threshold and OFF if it exceeds

TABLE III
SIMULATION RESULTS FROM DIFFERENT EXPERIMENTS USING THE SAME USER PROFILES IN TRAINING AND SIMULATION

Approaches	Scenario-0	Scenario-1	Scenario-2	Q-learning		ADHDP		Ref.
				$a = b = 1$	$a = 2 * b$	$a = b = 1$	$a = 2 * b$	
Experiment .1: Tank size:70 gallons, 100 simulation days, heating element 4500Wh								
Energy Cost in US $ for all DEWHs	43765	45373	45325	37820	37996	54683	42917	48578
Percentage of cost's reduction	9.9%	6.6%	6.7%	22.4%	21.78%	-12.2%	11.65%	-
Estimated Customer's Annual saving	$173	$115	$117	$393	$381	$-213	$204	-
# times users receives water below '*th*'	815	1998	1916	0	0	2042	954	132
Output water below '*th*' in gallons	2190	5035	5006	0	0	9114	2674	331
Experiment .2: Tank size: 70 gallons, 30 simulation days, heating element 4500Wh								
Energy Cost in US $ for all DEWHs	13296	13766	13775	11438	11571	16203	12522	14848
Percentage of cost's reduction	10.45%	7.29%	7.23%	22.25%	22%	-10%	15.67%	-
Estimated Customer's Annual saving	$186	$130	$129	$393	$393	$-179	$279	-
# times users receives water below '*th*'	382	608	570	0	0	20	261	38
Output water below '*th*' in gallons	1031	1521	1453	0	0	56	692	124
Experiment .3: Tank size: 40 gallons, 100 simulation days, heating element 2800Wh								
Energy Cost in US $ for all DEWHs	40400	40922	40930	40014	40016	45661	40485	42843
Percentage of cost's reduction	5.7%	4.48%	4.46%	6.6%	6.6%	-6.6%	5.5%	-
Estimated Customer's Annual saving	$87.97	$69.16	$68.86	$101	$102	$-102	$85	-
# times users receives water below '*th*'	17707	19763	19880	0	4	5388	13421	19528
Output water below '*th*' in gallons	35423	39515	39747	0	4.49	10830	26902	39072
Experiment .4: Tank size: 100 gallons, 30 simulation days, heating element 4500Wh								
Energy Cost in US $ for all DEWHs	12574	13188	13048	10819	10856	16900	11566	14628
Percentage of cost's reduction	14%	9.84%	10.8%	26.4%	25.78%	-14.98%	20.93%	-
Estimated Customer's Annual saving	$246	$173	$190	$466	$453	$-264	$367	-
# times users receives water below '*th*'	33	288	315	3	0	1416	58	0
Output water below '*th*' in gallons	22.2	672.1	693.5	3.17	0	594.5	182	0
Experiment .5: Tank size: 120 gallons, 100 simulation days, heating element 36000Wh (Commercial product)								
Energy Cost in US $ for all DEWHs	48807	46592	46441	36647	36585	48391	44915	47582
Percentage of cost's reduction	-2.54%	2.08%	2.4%	22.98%	22.77%	-1.7%	5.6%	-
Estimated Customer's Annual saving	$-44	$35.7	$41	$394	$388	$-29.45	$96	-
# times users receives water below '*th*'	0	0	10	0	0	0	0	0
Output water below '*th*' in gallons	0	0	5.3	0	0	0	0	0

140 °F. This scenario (also known as the uncontrolled group in the comparison charts) is used as a reference to calculate the performance of all other scenarios.

6) *Scenario 5 (ADHDP):* All the DEWHs simulated with in this group are trained, as illustrated in Section III-A, using the adaptive critic technique ADHDP.

Scenarios 3 and 5 are the techniques implemented in this paper. Both scenarios (Q-learning and ADHDP) perform well and even better than the state-of-the-art strategies (scenarios 0, 1, and 2) in the existing work [3].

In scenario 0, the agent simply deactivated the heating element during peak periods unless T_h fell below the soft threshold, which is in this paper was 125 °F. The controller turned the heating element ON all the time it was outside the specified peak periods unless their (T_h) exceeded the maximum allowed temperature (140 °F). New peaks appeared when the heating elements for all DEWHs were reactivated simultaneously. In scenarios 1 and 2, the agent randomly reactivated the water heaters at the end of the load shifting period, giving priority to those that were having the lowest water temperature to be turned ON first. The time required for the water heaters to reactivate at the end of the shifting load was based on a random function. It was also based on the water's temperature at the end of the load shifting period.

Scenarios 3 and 5 represented the control approaches that were presented in this paper. In scenario 3, the operation of the DEWH's heating element was entirely controlled by the suboptimal policy that was achieved during the Q-learning's training phase. The same simulator that generated the control variables during training was used to calculate them during comparison as well.

Furthermore, in all scenarios, the DEWH controller overrode its control scenario on two occasions: when T_h either decreased below or exceeded the prespecified soft-threshold (125 °F) or maximum (140 °F) thresholds, respectively. The soft threshold was used in the comparator simulator in order to guarantee the same degree of customer's satisfaction for all scenarios (when all scenarios maintained their water's temperature above the hard threshold of 120 °F. It also provided clear performance measurement for the different scenarios, based on the consumed power cost only. The cost of the consumed power was calculated using a ToU pricing profile [29], as illustrated in Section III-C.

V. RESULTS

The event driven simulator, the described system's modeling, the Q-learning process, ADHDP controller back propagation training, and the simulator used for evaluating the performance of all approaches were all designed using MATLAB. Many simulations were conducted in this paper using different training parameters. As a result, the best learning schemes for Q-learning in terms of distance between Q-factors and policy stability were obtained using a discount factor of $\lambda = 0.9$. Table II illustrates the optimum policies

TABLE IV
SIMULATION RESULTS FROM DIFFERENT EXPERIMENTS USING USER PROFILE IN FIG. 8 FOR TRAINING AND THE PROFILE IN FIG. 9 IN EVALUATION

Experiment .6: Tank size:70 gallons, 100 simulation days, heating element 4500 Wh						
Approaches	Scenario-0	Scenario-1	Scenario-2	Q-learning	ADHDP	Ref.
Energy Cost in US $ for all DEWHs	56011	57796	57785	53500	57114	62841
Percentage of cost's reduction	10.87%	8.03%	8.05%	14.86%	9.12%	-
Estimated Customer's Annual saving	$249.29	$184.14	$184.54	$340.95	$209.04	-
# times users receives water below '*th*'	49709	60584	60199	0	57242	18645
Output water below '*th*' in gallons	150140	182500	181250	0	172760	56596
Experiment .7: Tank size: 70 gallons, 30 simulation days, heating element 4500 Wh						
Energy Cost in US $ for all DEWHs	16666	17227	17226	16032	16864	18873
Percentage of cost's reduction	11.52%	8.55%	8.553%	14.9%	10.47%	-
Estimated Customer's Annual saving	$268.52	$200.26	$200.39	$345.66	$244.43	-
# times users receives water below '*th*'	12708	14694	14530	0	14548	3599
Output water below '*th*' in gallons	38931	44903	44446	0	44464	11156

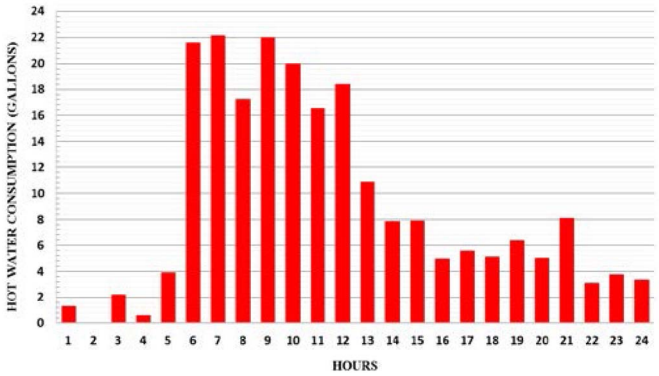


Fig. 9. Extreme user profile used in experiments 6 and 7. In this profile higher rate of hot water consumption is assumed for users.

selected by Q-learning for 30 and 100 iterations for DEWH with tank size 70 gallons.

The Q-learning agent showed the best performance in all experiences. The Q-learning approach implemented here is more advanced and comprehensive than that presented in previous work [34]. This paper uses realistic time events as generated by the event driven simulator. (*In the previous work [34], the controller made a decision every 30 min.*) The ADHDP approach is also conducted in all experiments and it outperformed the state-of-the-art approaches [3] when setting the appropriate values for the guidance parameters (a and b, see Section III-A). The ADHDP approach is based on the continuous state space version of the problem, not a discrete state space like Q-learning [20], [21], [30], [31], [35]. Several experiments were implemented, as illustrated in Tables III and IV. Table III contains results for the experiments that were implemented using the same profiles during training and evaluation, with user profiles illustrated in Fig. 8. Table III also includes the results for different values for the guidance parameters (a and b) and clearly shows their effect on the ADHDP performance. Table IV contains results for experiments that uses extreme profiles during the evaluation phase, as illustrated in Fig. 9. All the experiments in Table III were implemented twice using two different combinations of the guidance factors (a and b). The results are recorded twice for Q-learning and ADHDP, since these are the only approaches that may get affected by the influence of the guidance factors on the cost function. The remaining results were recorded when using ($a = 2 * b$). There were slight fluctuations in the values due to the random profile generation.

In experiment 1, a 100 typical DEWH with tank size of 70 gallons and 4.5 kWh heating element were simulated for 100 iterations (i.e., simulation days). Experiment 2 repeats experiment 1 but using 30 iterations only. Experiment 3 evaluates the performance of all approaches for smaller tank size DEWH. DEWH of 40 gallons tank was used in this experiment. Experiments 4 and 5 show and compare the performance of all the different scenarios using DEWHs with larger tanks (i.e., 100 and 120 gallons). In experiment 5, a commercial DEWH model with 36 kWh heating element was simulated. The experiments listed in Table IV (experiments 6 and 7) were performed using extreme user profiles during the evaluation process, in order to measure the robustness of the presented approaches. The results obtained from experiments 1–5, showed outstanding performance for the Q-learning approach regardless of the guidance factors (a and b). The ADHDP approach outperformed the state-of-the-art techniques when $a = 2 * b; b = 2$. But it performed poorly when setting $a = b = 1$. It was observed during some additional experiments that ADHDP has shown better performance in cost reduction when setting $a \gg b$ (e.g., $a = 8 * b; b = 1$).

The comparison simulation was implemented, as illustrated in Section IV. The simulation parameters were the same for the different scenarios, and each scenario had the same number of DEWHs. Furthermore, the ADP approaches were tested using more extreme user profiles than those they were trained with, yet the Q-learning method was able to provide better cost reduction rates than those presented in the state-of-the-art techniques, as illustrated in Table IV and Fig. 13.

Tables III and IV compare the performance of the ADP approaches with the state-of-the-art control strategies [3] in terms of energy cost reduction and customers satisfaction. The amounts in the first row of each experiment were calculated in the code by accumulating the instantaneous costs of energy consumed in all the 100 dwellings. The cost is based on the ToU profile described earlier [29]. The percentage cost reduction rates (i.e., numeric values in the second row) were calculated using the following equation:

$$\%\ \text{Cost Reduction} = \frac{\text{Approach's Total saving}}{\text{Ref. energy cost}} \times \%100 \tag{10}$$

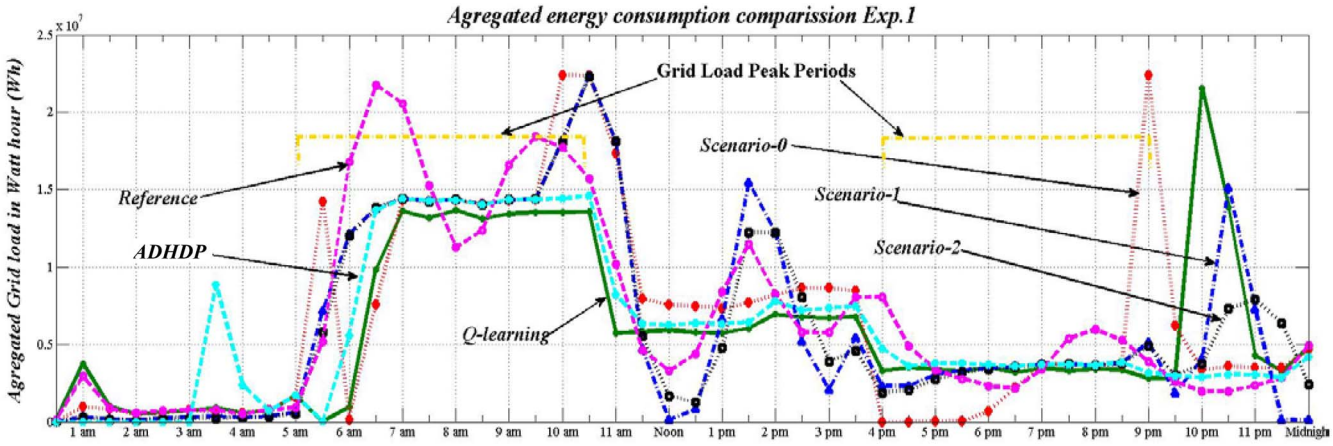


Fig. 10. Aggregated energy consumed by all approaches during experiment 1. DEWH's tank size 70 gallons, 100 simulation days.

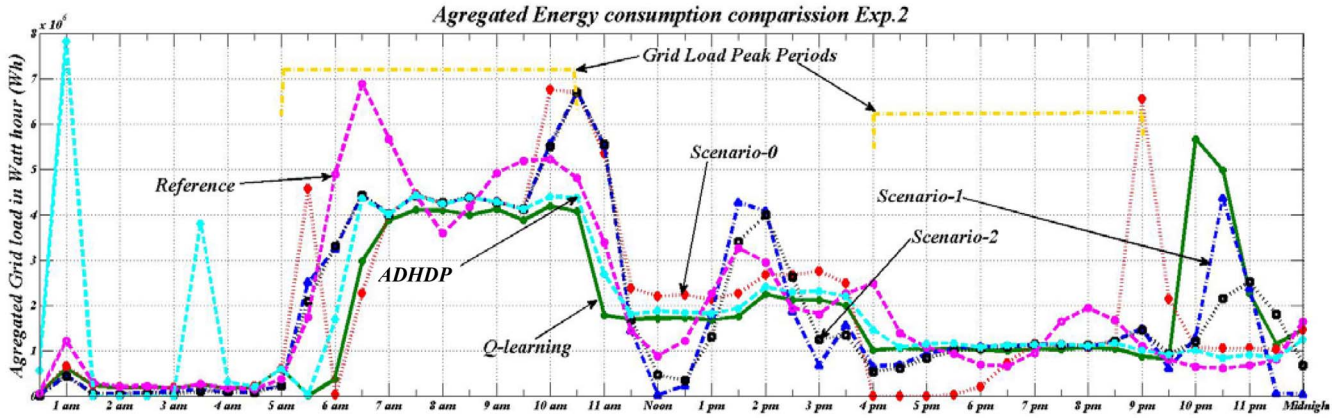


Fig. 11. Aggregated energy consumed by all approaches during experiment 2. DEWH's tank size 70 gallons, 30 simulation days.

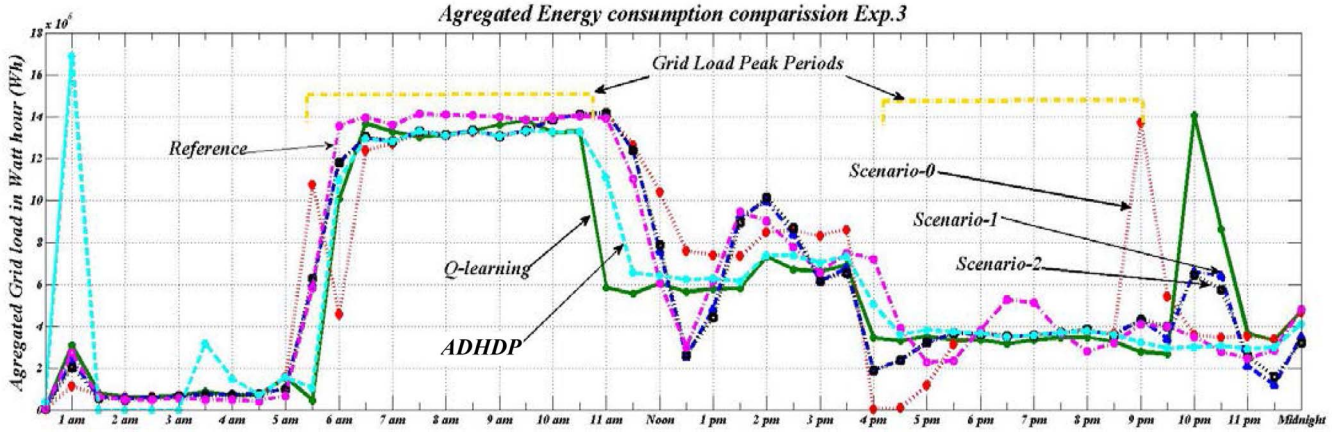


Fig. 12. Aggregated energy consumed by all approaches during experiment 3. DEWH's tank size 40 gallons, 100 simulation days.

where approach's total saving = (Cost of energy consumed using the uncontrolled approach "Ref." − the cost of energy consumed using the approach). The uncontrolled scenario is used as an index for comparison or to evaluate between all the control strategies. The estimated annual saving was calculated from multiplying the per-day saving by 365, as illustrated in (11). The customer's satisfaction evaluation was based on the quantities of the output water with temperature less than "th" (i.e., 120 °F) and the number of times that happened in the entire 100 dwellings

$$\text{EAS} = \frac{\text{Approach's Total saving}}{\#\ \text{DEWHs}} \times \frac{365}{\#\ \text{simulation days}}. \tag{11}$$

The results illustrated in Figs. 10–13 and Tables III and IV indicate that, in most cases, using Q-learning to control the

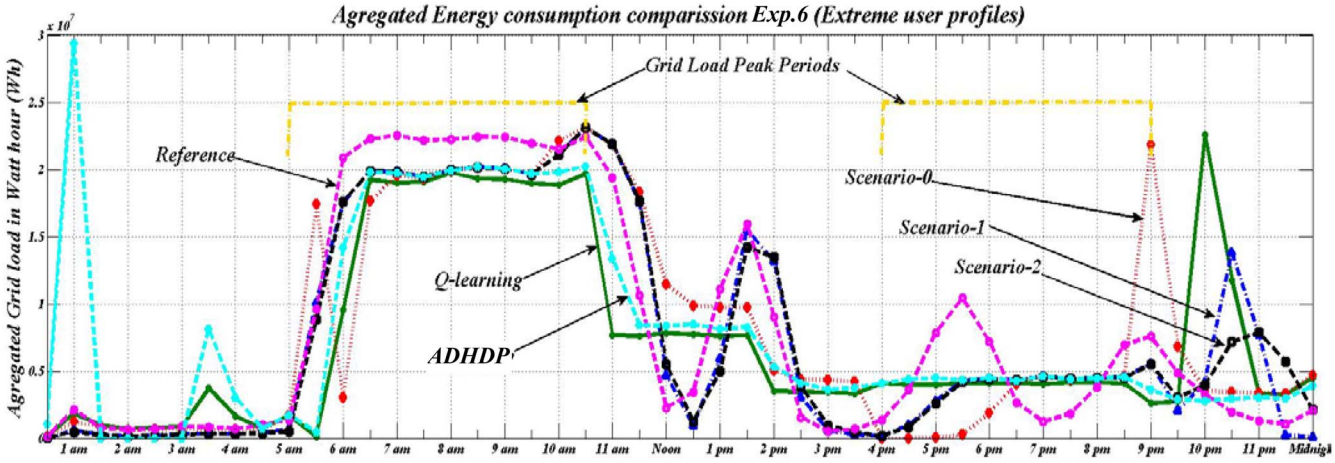


Fig. 13. Aggregated energy consumed by all approaches during experiment 6. Users' profiles in Fig. 9 were used in comparison.

100 DEWHs reduced the cost of the consumed power by 22% which is twice the cost reduction resulted from the state-of-the-art technique "scenario 0." The ADHDP approach also outperforms the state-of-the-art techniques, through reducing the total cost by 16% and 12% when trained for 100 and 30 days, respectively.

In experiment 3, Q-learning reduced the total cost by 6.6% which is still higher than the reduction from other techniques (e.g., 5.7% for scenario 0). ADHDP in this experiment reduced the cost by only 5.5% only which is less than scenario 0, but higher than scenarios 1 and 2. However, the best performance recorded was in experiment 4 with tank size $=$ 100 gallons and heating element of 4.5 kWh. The percentage cost reduction rates were $\sim$ 26% and 21% for the Q-learning and ADHDP, respectively. The annual savings were approximately \$453 and \$367 for Q-learning and ADHDP, respectively.

In experiments 6 and 7 (Table IV), a higher user profile was used in the evaluation than that used during training. The Q-learning also outperformed all other scenarios by producing about 15% cost reduction. According to the simulation results, the Q-learning controller maintained the temperature of the output water above the prespecified threshold (120 °F), except when the 40 gallons tank size was used, it provided a small amount of water slightly below the threshold, th. The ADHDP approach outperformed the state-of-the-art scenarios if the training was performed on the same data used during evaluation and a large tank size (70 gallons) was used.

VI. Conclusion

In conclusion, the Q-learning approach can at least save a family of four persons between \$102, \$393, and \$453 annually if they are using a DEWH with 40, 70, and 100 gallons, respectively. Even for the commercial product the ADP approaches provide excellent annual saving of (\$394) for DEWH with larger heating element (36 kWh) and tank size of 120 gallons. Furthermore, the simulation showed that the Q-learning controller maintained the water temperature above 120 °F, indicating an opportunity to enhance the system further by designing a flexible threshold. The simulation results shown in Tables III and IV clearly illustrate that Q-learning has the best performance in terms of cost reduction, customer's satisfaction, and even in terms of load peak elimination or shifting, as illustrated in Figs. 10–13.

Another opportunity for further enhancement would be using real user profiles to control DEWH in real time. ADP may also be used to improve the most recent heat pump water heaters. The presented techniques do not depend on the technology used in the DEWH, but depend only on the grid load demand (i.e., instantaneous energy cost), the temperature of the output water, and the user profile. The authors therefore believe that various ADP approaches are worth further investigation. Q-learning outperformed ADHDP and previous state-of-the-art methods in these experiments. The authors speculate that ADHDP will still prove useful in scenarios that play to its strengths in continuous state spaces and dynamic environments such as adaptive thresholds. The authors are also aware that some of the references cited in this paper demonstrated better performance with ADHDP than with Q-learning. However, for these experiments, Q-learning outperformed it and all other methods, probably due to the reduced state space and the limited complexity of the implemented state space model. This is encouraging; for Q-learning is a simple and robust, easily deployable ADP approach. The results presented here strongly suggest it should not be difficult to use simple machine learning techniques to achieve substantial cost savings and environmental benefits.

References

[1] Eia.gov. (2015). *Residential Energy Consumption Survey (RECS)—Analysis & Projections—U.S. Energy Information Administration (EIA)*. Accessed on Oct. 24, 2015. [Online]. Available: http://www.eia.gov/consumption/residen

[2] Energy.gov. (2015). *Office of Energy Efficiency & Renewable Energy | Department of Energy*. Accessed on Oct. 25, 2015. [Online]. Available: http://energy.gov/eere/office-energy-efficiency-renewable-energy

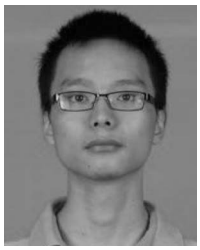
[3] A. Moreau, "Control strategy for domestic water heaters during peak periods and its impact on the demand for electricity," *Energy Procedia*, vol. 12, pp. 1074–1082, Dec. 2011.

[4] "Measurement of domestic hot water consumption in dwellings," Dept. Environ. Food Rural Affairs, London, U.K., Tech. Rep., 2008. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/48188/3147-measure-domestic-hot-waterconsump.pdf

[5] S. Lu and M. Kintner-Meyer, "Scoping study for the demand response DFT II project in Morgantown," Pac. Northwest Nat. Lab., U.S. Dept. Energy, Washington, DC, USA, Tech. Rep. PNNL-17474, 2008.
[6] M. H. Nehrir, B. J. LaMeres, and V. Gerez, "A customer-interactive electric water heater demand-side management strategy using fuzzy logic," in *Proc. IEEE Power Eng. Soc. Winter Meeting*, vol. 1. New York, NY, USA, Jan. 1999, pp. 433–436.
[7] A. Sepulveda *et al.*, "A novel demand side management program using water heaters and particle swarm optimization," in *Proc. IEEE Elect. Power Energy Conf. (EPEC)*, Halifax, NS, Canada, Aug. 2010, pp. 1–5.
[8] Peakload.org. (2015). *2014 Annual Report to Members—Peak Load Management Alliance*. Accessed on Oct. 25, 2015. [Online]. Available: http://www.peakload.org/?page=2014Report
[9] Powermin.nic.in. (2016). *Annual Reports Year-Wise Indian Ministry of Power*. Accessed on Feb. 2, 2016. [Online]. Available: http://powermin.nic.in/annual-reports-year-wise
[10] Y. M. Atwa, E. F. El-Saadany, and M. M. Salama, "DSM approach for water heater control strategy utilizing elman neural network," in *Proc. IEEE Canada Elect. Power Conf. (EPC)*, Montreal, QC, Canada, Oct. 2007, pp. 382–386.
[11] N. Saker, M. Petit, J. C. Vannier, and J. L. Coullon, "Demand side management of electrical water heaters and evaluation of the cold load pick-up characteristics," in *Proc. 17th Power Syst. Comput. Conf.*, Trondheim, Norway, 2011, pp. 1–8.
[12] S. Lefebvre and C. Desbiens, "Residential load modeling for predicting distribution transformer load behavior, feeder load and cold load pickup," *Int. J. Elect. Power Energy Syst.*, vol. 24, no. 4, pp. 285–293, May 2002.
[13] C. Diduch *et al.*, "Aggregated domestic electric water heater control—Building on smart grid infrastructure," in *Proc. 7th Int. Power Electron. Motion Control Conf. (IPEMC)*, vol. 1. Harbin, China, Jun. 2012, pp. 128–135.
[14] B. Ramanathan and V. Vittal, "A framework for evaluation of advanced direct load control with minimum disruption," *IEEE Trans. Power Syst.*, vol. 23, no. 4, pp. 1681–1688, Nov. 2008.
[15] S. Tiptipakorn and W.-J. Lee, "A residential consumer-centered load control strategy in real-time electricity pricing environment," in *Proc. IEEE 39th North Amer. Power Symp. (NAPS)*, Las Cruces, NM, USA, Sep./Oct. 2007, pp. 505–510.
[16] N. Lu and S. Katipamula, "Control strategies of thermostatically controlled appliances in a competitive electricity market," in *Proc. IEEE Power Eng. Soc. Gen. Meeting*, San Francisco, CA, USA, Jun. 2005, pp. 202–207.
[17] B. Rautenbach and I. E. Lane, "The multi-objective controller: A novel approach to domestic hot water load control," *IEEE Trans. Power Syst.*, vol. 11, no. 4, pp. 1832–1837, Nov. 1996.
[18] Aceee.org. (2015). *Water Heaters Get an Efficiency Makeover Courtesy of the Department of Energy | ACEEE*. Accessed on Oct. 30, 2015. [Online]. Available: http://aceee.org/blog/2015/02/water-heaters-get-efficiency-makeover
[19] Appliance-standards.org. (2015). *The Good News, and the Not-So-Good News, on the New DOE Water Heater Test Procedure | ASAP Appliance Standard Awareness Project*. Accessed on Oct. 30, 2015. [Online]. Available: http://www.appliance-standards.org/blog/good-news-and-not-so-good-news-new-doe-water-heater-test-procedure
[20] D. V. Prokhorov and D. C. Wunsch, "Adaptive critic designs," *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 997–1007, Sep. 1997.
[21] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1. Belmont, MA, USA: Athena Sci., 1995.
[22] (2015). *Who.int*. Accessed on Nov. 13, 2015. [Online]. Available: http://www.who.int/water_sanitation_health/e
[23] R. E. Sonntag, C. Borgnakke, and G. J. Van Wylen, *Fundamentals of Thermodynamics*. New York, NY, USA: Wiley, 1998, pp. 356–357.
[24] J. J. Klemes, R. Smith, and J.-K. Kim, Eds., *Handbook of Water and Energy Management in Food Processing*. Burlington, MA, USA: Elsevier, 2008.
[25] F. Cardarelli, *Encyclopaedia of Scientific Units, Weights and Measures: Their SI Equivalences and Origins*. London, U.K.: Springer, 2003.
[26] A. M. Omer, "Energy, environment and sustainable development," *Renew. Sustain. Energy Rev.*, vol. 12, no. 9, pp. 2265–2300, 2008.
[27] A. S. Mujumdar, "A review of 'mathematical principles of heat transfer,'" *Drying Technol.*, vol. 24, no. 2, p. 245, 2006.
[28] Formulas and Facts. (n.d.). (Jan. 13, 2014). *Contractorsinstitute.com*. [Online]. Available: http://www.contractorsinstitute.com/downloads/Solar/Contractors'%20Domestic%20Hot%20Water%20Educational%2020PDF's/Hot%20Water%20Formulas%20and%20Facts.pdf
[29] S. Borenstein, "Time-varying retail electricity prices: Theory and practice," in *Electricity Deregulation: Choices and Challenges*. Chicago, IL, USA: Univ. Chicago Press, 2005, pp. 111–130.
[30] C. J. Watkins and P. Dayan, "*Q*-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[31] A. Gosavi, *Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning*, vol. 25. New York, NY, USA: Springer, 2003.
[32] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, 1951.
[33] P. J. Werbos, *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*, vol. 1. New York, NY, USA: Wiley, 1994.
[34] K. Al-jabery, D. C. Wunsch, J. Xiong, and Y. Shi, "A novel grid load management technique using electric water heaters and *Q*-learning," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Venice, Italy, 2014, pp. 776–781.
[35] G. K. Venayagamoorthy, R. G. Harley, and D. C. Wunsch, "Comparison of heuristic dynamic programming and dual heuristic programming adaptive critics for neurocontrol of a turbogenerator," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 764–773, May 2002.

Khalid Al-jabery (S'14) received the B.S. and M.S. degrees in computer engineering from Basrah University, Basrah, Iraq, in 2005 and 2009, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Missouri University of Science and Technology, Rolla, MO, USA.

His current research interests include simulation, machine learning, bio-medical data analysis, clustering, and adaptive critic design.



Zhezhao Xu received the B.S. degree in electronic science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2014. He is currently pursuing the M.S. degree in computer science with Tsinghua University, Beijing, China.

His current research interests include modeling and simulation of complex systems.

Wenjian Yu (S'01–M'04–SM'10) received the B.S. and Ph.D. degrees in computer science from Tsinghua University, Beijing, China, in 1999 and 2003, respectively.

He joined Tsinghua University, in 2003, where he is an Associate Professor with the Department of Computer Science and Technology. He was a Visiting Scholar with the Department of Computer Science and Engineering, University of California at San Diego, San Diego, CA, USA, twice during the period from 2005 to 2008. He has authored two books and over 130 papers in refereed journals and conferences. His current research interests include modeling and simulation of complex systems (integrated circuits and others), numerical algorithms, and their applications.

Dr. Yu was a recipient of the Distinguished Ph.D. Award from Tsinghua University in 2003, the Excellent Young Scholar Award from the National Science Foundation of China in 2014, and the Best Paper Award from the Design, Automation and Testing in Europe Conference in 2016.

Donald C. Wunsch, II (F'87) received the B.S. degree in applied mathematics from the University of New Mexico, Albuquerque, NM, USA, and Jesuit Core Honors Program, Seattle University, Seattle, WA, USA, the M.S. degree in applied mathematics and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, and the Executive M.B.A. degree from Washington University at St. Louis, St. Louis, MO, USA.

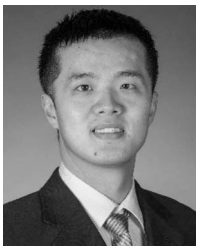
He is the Mary K. Finley Missouri Distinguished Professor with the Missouri University of Science and Technology (Missouri S&T), Rolla, MO, USA. He was with Texas Tech University, Lubbock, TX, USA, Boeing, Chicago, IL, USA, Rockwell International, Albuquerque, NM, USA, and International Laser Systems, Holyoke, MA, USA. He has supervised 18 Ph.D. students in computer engineering, electrical engineering, and computer science and attracted over $10 million in sponsored research. He has over 400 publications including nine books with over 11 000 citations. His current research interests include clustering/unsupervised learning, adaptive resonance and reinforcement learning architectures, hardware and applications, neurofuzzy regression, traveling salesman problem heuristics, robotic swarms, and bioinformatics.

Prof. Wunsch was a recipient of the NSF CAREER Award and the 2015 INNS Gabor Award. He was the INNS President, an INNS Fellow, and a Senior Fellow from 2007 to 2013. He served as the IJCNN General Chair, and on several boards, including the St. Patrick's School Board, the IEEE Neural Networks Council, the International Neural Networks Society, and the University of Missouri Bioinformatics Consortium. He has Chaired the Missouri S&T Information Technology and Computing Committee as well as the Student Design and Experiential Learning Center Board.

Jinjun Xiong (S'05–M'07) received the Ph.D. degree from the University of California at Los Angeles, Los Angeles, CA, USA, in 2006.

He is currently the Program Director and a Research Staff Member with IBM Thomas J. Watson Research Center, Yorktown, NY, USA, responsible for building world-class programs on cognitive computing systems research. He has published over 100 technical papers in refereed international conferences and journals. His current research interests include future computing systems, cognitive computing, big data analytics, smarter energy, and very large-scale integrated circuit designs.

Dr. Xiong was a recipient of numerous Best Paper Awards, the Best Paper Award nominations, and the Outstanding Ph.D. Award from the University of California at Los Angeles.



Yiyu Shi (SM'06) received the B.S. degree (Hons.) in electronic engineering from Tsinghua University, Beijing, China, in 2005, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Los Angeles, Los Angeles, CA, USA, in 2007 and 2009, respectively.

He is currently an Associate Professor with the Departments of Computer Science and Engineering and Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA. His current research interests include 3-D integrated circuits, hardware security, and renewable energy applications.

Dr. Shi was a recipient of many best paper nominations in top conferences, the IBM Invention Achievement Award in 2009, the Japan Society for the Promotion of Science Faculty Invitation Fellowship, the Humboldt Research Fellowship for Experienced Researchers, the IEEE St. Louis Section Outstanding Educator Award, the Academy of Science (St. Louis) Innovation Award, the Missouri University of Science and Technology Faculty Excellence Award, the National Science Foundation CAREER Award, the IEEE Region 5 Outstanding Individual Achievement Award, and the Air Force Summer Faculty Fellowship.