Residential Demand Response Using Reinforcement Learning

Daniel O'Neill*, Marco Levorato*, Andrea Goldsmith* and Urbashi Mitra†
* Dept. of Electrical Engineering, Stanford University, Stanford, CA 94305 USA
† Dept. of Electrical Engineering, University of Southern California, Los Angeles, USA.
e-mail: dconeill@stanford.edu, levorato@stanford.edu, andreag@stanford.edu, ubli@usc.edu.

***Abstract*— We present a novel energy management system for residential demand response. The algorithm, named CAES, reduces residential energy costs and smooths energy usage. CAES is an online learning application that implicitly estimates the impact of future energy prices and of consumer decisions on long term costs and schedules residential device usage. CAES models both energy prices and residential device usage as Markov, but does not assume knowledge of the structure or transition probabilities of these Markov chains. CAES learns continuously and adapts to individual consumer preferences and pricing modifications over time. In numerical simulations CAES reduced average end-user financial costs from 16% to 40% with respect to a price-unaware energy allocation.**

I. Introduction

Demand response (DR) systems [1]–[3] enables the dynamic adjustment of electrical demand in response to pricing signals. DR offers several benefits. By suitably adjusting energy prices, electrical load can be shifted from periods of high or peak demand to other periods, thereby reducing the ratio of peak to average loads. This, in turn, can improve operational efficiency, reduce operating costs, improve capital efficiency, and reduce harmful emissions and risk of outages. DR has been extensively investigated for larger energy users and has been implemented in many areas (*e.g.*, [4], [5]).

Residential DR [6]–[9] potentially offers similar benefits but also faces several challenges. Technical challenges include the deployment of an infrastructure supplying real time pricing information to consumers in a useful way, networking home devices, ensuring security, advanced metering [9], [10] and developing fully-automated Energy Management Systems (EMS) [9], [11] to manage home energy usage in a manner acceptable to consumers.

There are also important decision-making challenges. Consumers make many energy-related decisions each day. With real time variable pricing, consumers face an infinite sequence of decisions to either use a particular device now and consume energy at current (known) prices or to defer using the device until later at possibly unknown prices. Each decision implicitly requires the consumer to estimate what future energy prices may be and weigh this differential cost against the dis-utility of waiting. This decision is further complicated by consumers selecting devices in a correlated fashion (*e.g.* the range and then the dishwasher.) We hypothesize that few consumers will be willing to continuously make a sequence of decisions of this type, especially when many of these decisions will have limited short term financial impact on the consumer. Under this hypothesis, fully-automated EMS algorithms are a necessary part of residential DR.

We present a novel EMS algorithm that learns a residence's behavior and automatically makes optimal energy scheduling and allocation decisions. Our contribution is in modeling the problem and using on line learning methods to optimally and adaptively solve it. We call the algorithm Consumer Automated Energy Management System (CAES.) The consumer decides what devices need to be run and makes an *energy reservation* by selecting the device by, for example, pushing the start button. CAES then schedules when the devices are run and how much energy to allocate to the devices. CAES factors in the statistical impact of future prices and the correlation between devices selected by consumers. The CAES algorithm minimizes the infinite horizon average financial cost of consuming energy and the average dis-utility to the consumer of delaying operation of the selected device. The CAES algorithm models both consumer energy reservations and energy prices as Markov chains but with unknown transition probability distributions. To overcome this problem, CAES uses Q-learning [12], a type of temporal-difference learning, to allow the algorithm to learn the behaviors of consumers and to optimally make energy consumption decisions.

Prior work such as [10], [13], [14], statically defines a response to high energy prices using different cost functions. CAES is different in that it

- learns the relevant statistical behavior of residential users and residential energy pricing;
- automatically adapts to modifications of consumer behavior;
- schedules energy usage taking into account the long term financial cost and dis-utility of delay seen by customers.

These features, together with the possibility for incorporating additional constraints and cost functions, make CAES a powerful and flexible control tool for price and consumer-aware energy management in smart residences.

Numerical results show that CAES significantly reduces the financial cost of consuming energy with respect to price-unaware energy management. Moreover, CAES smooths energy usage over time and reduces the probability of demand peaks. Simulation results show that CAES enables a reduction of the average end-user financial cost ranging from 16% to 40% with respect to price-unaware energy allocation depending on the *value* given to the satisfaction of the consumer.

The remainder of this paper is organized as follows: Section II describes the system model, and Section III presents the system performance metric. Section IV poses the problem as

an infinite horizon discounted cost Markov decision process and describes its properties. Section V describes Q-learning and the CAES algorithm. Section VI describes numerical examples, and Section VII summarizes our conclusions.

II. Model

We consider a single residence receiving real time variable pricing. Time is discrete $t = 1, \ldots,$ and pricing signals are indicated by the *pricing sequence* $p(t) \in \mathbf{R}_+$. The residence has $m = 1, \ldots, M$ electrical devices, such as air conditioners, ovens, dishwashers, computers, electric vehicles or other devices. Each device requires a certain quantity γ_m of electrical energy to operate and complete its work. In some cases this quantity can be spread over several time slots, perhaps by running the device at a different speed or intensity. Residents select a device and indicate their desire to run it by, for example, depressing its start button. The *consumer reservation vector* $z(t) \in \mathbf{R}_+^M$ captures this consumer action. If at time t, the consumer selects device m, then $z_m(t) = \gamma_m$. Otherwise the value of $z_m(t)$ is zero. However, this action does not necessarily start the device, but rather makes a reservation with CAES to run the device at an optimal time to minimize measures of long term costs, described more fully below.

The sequence of consumer reservation vectors, $z(t)$, captures the visible actions of the set of consumers in a residence. Frequently, there will be statistical relationships among consumer selections about what devices to run that can be exploited to reduce overall costs (*e.g.* using the cook top suggests the dishwasher will soon be run.) Thus, we model $z(t)$ as a Markov chain. When these behaviors are periodic, such as diurnally or weekly cycles, a block Markov model can be used. For clarity, we use a simple Markov model in this paper. The transition probabilities associated with this Markov chain can vary with different residences and are thus assumed unknown. The Markov model can be extended to force immediate operation of a device by depressing a button a second time.

Similarly, the energy pricing sequence, $p(t)$, is assumed to be Markov, also with an unknown probabilistic structure. Note that correlation between the processes $z(t)$ and $p(t)$ can be easily modeled by representing the overall system's state as the aggregate $(z(t), p(t))$. In this case, the algorithm will learn the statistics of the aggregate process.

As shown in Figure 1, the CAES algorithm takes as input the sequence of consumer reservation vectors $z(t)$ and the sequence of energy prices $p(t)$. The output is the vector of energies allocated to devices $u(t) \in \mathbf{R}_+^M$ and is called the *energy control policy*. The value of $u_m(t)$ is the energy allocated to device m at time t and is constrained to be less than the pending energy requirement of device m. The CAES algorithm finds the optimal energy control policy $u(t)$ to minimize the infinite horizon discounted "cost", which is discussed in Section III. CAES may delay the start of a device or allocate only a portion of the total energy necessary[1], postponing completion of the device's operation until a later

[1]For instance, light intensity, monitors' brightness, battery charging speed can be reduced in order to reduce the instantaneous demand for energy.

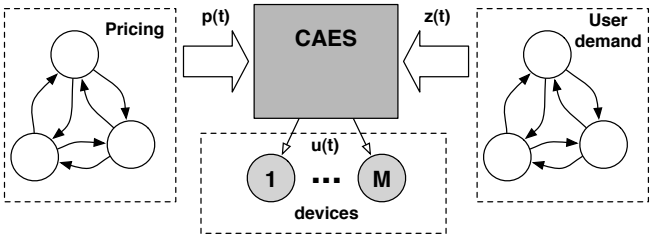

Fig. 1. CAES Energy Management System

period. When only a portion of the necessary energy is allocated to a device by the energy control policy, a pending backlog of needed energy is created. The *pending energy backlog* is written as $x(t) \in \mathbf{R}_+^M$ and

$$x(t+1) = x(t) + z(t) - u(t), \tag{1}$$

where we note $u(t) \leq x(t)$.

We define the state of the system to be

$$\Omega(t) = [x(t)^T, y(t)^T, z(t)^T, p(t)^T]^T, \tag{2}$$

where $\Omega \in \mathbf{R}_+^{3M+1}$.

The set of possible energy control policies is defined as the set of all stationary state feedback policies of the form $u(\Omega)$ such that $0 \leq u \leq x$. Note, at time t, the current values of the consumer selection process $z(t)$ and the market pricing process $p(t)$ are known to the energy control policy.

A. Utility Functions and Time Preferences

Utility functions are set functions that mathematically represent (unconstrained) consumer preferences in the economics and consumer preference literature [15], [16]. Utility functions are an abstraction used to numerically model that consumers prefer one product versus another or, as used here, that consumers have time preferences for things. Estimating individual utility functions is very difficult, with parameterized functional approximations often used in applications. We follow this approach here.

We assume consumers would prefer to have devices operate sooner rather than later, and that we can express this preference as a strictly concave utility function. For convenience, we work with the negative of this function and call it the *dis-utility function* $\bar{U}_m(y_m(t)) \in \mathbf{R}_+$. The auxiliary variable $y \in \mathbf{R}_+^M$ is defined as

$$y(t+1) = \theta y(t) + (I - \theta)x(t), \tag{3}$$

where I is the identity matrix and $0 < \theta < I$ is a diagonal matrix that parameterizes the utility function. Different residences can have different θ's. We call $y(t)$ the *average pending workload* of the system as it exponentially smooths pending energy backlog. More complicated averaging schemes, capturing more refined consumer dissatisfaction functions, can be modeled by choosing a (non-diagonal) θ as a filtering operator, but for clarity a simple scheme is used in this paper. As consumers wait longer for a device to complete its work, the dis-utility function increases in value, reflecting the consumers' dis-satisfaction of waiting. When $\theta \approx I$ consumers are sensitive to the average delay of a device

completing service. When $\theta = 0$, they are concerned about the delay associated with the device currently selected. Different home devices will have different dis-utility functions; Delaying cooking dinner is a problem, while delaying the dishwasher doing the resultant dishes probably is not. The approach can also incorporate individual device constraints in a simple and natural fashion in the form of modified dis-utility functions or as instantaneous energy allocation constraints. We assume the dis-utility function is strictly convex, positive and increasing.

III. Performance Metric

The consumer bears two types of costs: financial cost of consuming energy and the dis-utility of waiting for a device to operate and complete its work. The CAES performance metric is the average total discounted financial and dis-utility costs over an infinite horizon. The financial cost at time t is simply

$$\sum_{m}^{M} p(t)u_m(t), \tag{4}$$

and the dis-utility of delay at time t is

$$\sum_{m}^{M} \bar{U}_m(y_m(t)). \tag{5}$$

Because $\bar{U}_m$ is strictly convex, positive and increasing, every energy reservation will eventually be satisfied.

For a given state, $\Omega(t)$, and energy control policy, $u(t)$, the cost incurred is the sum of the financial cost and dis-utility cost

$$\Phi(\Omega(t), u(t)) = \sum_{m}^{M}(p(t)u_m(t) + \lambda\bar{U}_m(y_m(t))), \tag{6}$$

$\lambda \geq 0$ determines the trade-off between the financial cost and dis-utility cost and can be seen as the dollar cost of average delay. The average cost for period t is thus,

$$\mathbf{E}[\Phi(\Omega(t), u(t))] = \mathbf{E}[\sum_{m}^{M}(p(t)u_m(t) + \lambda\bar{U}_m(y_m(t)))], \tag{7}$$

where $\mathbf{E}[\,]$ is the expectation operator.

System performance is defined as the total discounted period costs over an infinite horizon

$$V_u(\Omega_0) = \lim_{T\to\infty} \mathbf{E}[\sum_{t}^{T} \gamma^t \Phi(\Omega(t), u(t))], \tag{8}$$

where Ω_0 is the initial state of the system and $0 \leq \gamma \leq 1$ is the discount factor. The discount factor reflects the fact that the consumer may give greater relevance to the immediate financial and dis-utility cost. The expectation operator is over $\{(z(t), p(t))\}_{t=0}^{\infty}$, given the system starts from Ω_0. The variable $V_u(\Omega_0)$ is the total discounted cost under the energy control policy u. Different policies can result in different total discounted costs.

The objective is to find the best stationary energy control policy u^* that minimizes (8). That is, we seek stationary functions u^* such that

$$V_{u^*}(\Omega_0) = \inf_{\pi\in\Pi} V_\pi(\Omega_0), \tag{9}$$

where π is any time sequence of stationary state feedback functions contained in the admissible class Π.[2]

Equations (8) and (9) describe a discounted cost Markov Decision Process, MDP, [17], [18]. The system state variable is Ω. The energy control policy is a function of the system state variable. Note that we are using the term policy to refer to both the function and (more properly) to the sequence of using the function repeatedly over the infinite horizon.

The model can be extended to include many specific consumer requirements by adding constraints to (8) or by restricting the admissible class of energy control policies. For instance, some devices may only operate at full power. The result would be a constrained MDP.

IV. Optimality Equations and Analysis

Bellman's equation [18] describes the optimality condition for a discounted cost MDP. The optimal energy policy $u(t)$ uses knowledge of the current energy backlog $x(t)$, and average energy backlog $y(t)$, current consumer selections $z(t)$ and current energy prices $p(t)$ to adapt to changing conditions. Knowledge of the values of $z(t)$ and $p(t)$ during the current period allows us to use the post-decision form of the Bellman equation [19]. Policy u^* is optimal if one can find a function $V(\Omega)$ such that

$$V(\Omega) \quad = \quad \min_{0\leq u\leq x}(\Phi(\Omega, u) + \mathbf{E}[V(\Omega_1)]), \tag{10}$$

where Ω_1 is the system state after using the energy control policy $u(\Omega)$. The function $V(\Omega)$ is called the value function and is the optimal cost when starting from state Ω. Unfortunately, Equation (10) is very difficult to solve analytically [18]. Numerical methods can be used when the state space $|\Omega|$ is small and the transition probability distributions are known. However, here the Markov transition probabilities for $z(t)$ and $p(t)$ are assumed unknown. Consequently, our approach is to to use an online learning approach, Q-learning [12], to estimate $V(\Omega)$. We first characterize the value function and the optimal control policy to gain insight into the structure and properties of the solution.

The cost function $\Phi(\Omega)$ is strictly convex and, consequently, it can be shown, using value iteration and elementary properties of limits, that $V(\Omega)$ is also convex in x and y. The expectation operator preserves convexity, so

$$L(x, y) = \mathbf{E}[V(\Omega)], \tag{11}$$

is also convex. Assuming L is differentiable and the optimal energy policy $0 < u^* < x$, then at u^*

$$\nabla_u\Phi - \gamma\nabla_u L = 0, \tag{12}$$

or equivalently,

$$p - \gamma\nabla_u L = 0. \tag{13}$$

Equation (13) has the interpretation of balancing current financial prices p against the discounted change in average future costs. The optimal energy policy simply allocates energy to devices to balance these costs. If $p \geq \gamma\nabla_u L$, then it is less

[2]By admissible class, we mean the set of all the admissible policy functions, which is defined by the action set and possible constraints.

expensive to defer using energy and conversely. Note, the effect of the dis-utility function $\bar{U}$ is captured solely in the average future cost term L.

In the more general case, in which L is not assumed to be differentiable, we form the Lagrangian and apply the Karush-Kuhn-Tucker (KKT) [20] conditions. We assume strong duality holds. The Lagrangian is

$$\Phi + \gamma L - \nu_1^T u + \nu_2^T (u - x) \tag{14}$$

where $\nu_1 \in \mathbf{R}_+^M$ and $\nu_2 \in \mathbf{R}_+^M$ are Lagrange multipliers. At the optimal $u = u^*$, the KKT conditions are

$$\begin{aligned}
\nabla_u \Phi - \gamma \nabla_u L - \nu_1 + \nu_2 &= 0 && (15)\\
u &\leq x && (16)\\
u &\geq 0 && (17)\\
\nu_i &\geq 0 \quad i = 1,2 && (18)\\
(\nu_1)_m u_m &= 0 \quad m = 1,...,M && (19)\\
(\nu_2)_m (u - x)_m &= 0 \quad m = 1,...,M. && (20)
\end{aligned}$$

Equation (15) requires the gradient of the Lagrangian to be zero and (16) through (18) require u, ν_1 and ν_2 to be primal and dual feasible. The complementary slackness conditions (19) and (20) require the Lagrange multipliers to be nonzero if the associated constraint is active. Because the per period cost function Φ is separable in u, Equation (15) can be rewritten component-wise as

$$p - \gamma \frac{\partial L}{\partial u_m} - (\nu_1)_m + (\nu_2)_m = 0 \quad \forall m = 1,...,M. \tag{21}$$

At most one of $(\nu_i)_m \quad i = 1,2$ can be nonzero. When the optimal energy control policy $u_i = 0$, then

$$(p - (\nu_1)_m) = \gamma \frac{\partial L}{\partial u_m}, \tag{22}$$

and the Lagrange multiplier acts to effectively decrease the price of energy. Similarly, when $u_m = x_m$, then

$$(p + (\nu_2)_m) = \gamma \frac{\partial L}{\partial u_m}, \tag{23}$$

and the Lagrange multiplier acts to effectively increase the price of energy.

V. Q-Learning and CAES

Solving (10) for $V(\Omega)$ and $u(\Omega)$ generally requires knowledge of the transition probability distributions for z and p. Unfortunately, the underlying structure and probability distributions for the consumer reservation process z are based on the preferences and behavior of individual consumers and are generally unknown. These preferences are idiosyncratic, resulting in different probability models for different consumers. Further, z is the result of decision processes internal to the consumer that are not directly observable. In this way, z can be thought of as a partially observable MDP. Similarly, the market price of energy p obeys a Markov model that is difficult to estimate and which may change with changing market conditions. To address these problems we use an on line learning algorithm. We use Q-learning [12] to illustrate the approach because of its relative ubiquity and its intuitive algorithm. Other approaches such as TD learning, and Least Squares TD learning can also be used, but at additional complexity, possible convergence issues and without providing much additional intuitive insight.

Q-learning is an on line, reinforcement learning method to approximate the value function $V(\Omega)$. Q-learning estimates the value of a Q-function for each state-action pair (Ω, u). The value functions and Q-functions are related by

$$V(\Omega) = \min_u Q(\Omega, u). \tag{24}$$

At state Ω, Q-learning samples the the behavior of the system in response to using control policy u and then updates an estimate of the Q-function $\hat{Q}(\Omega, u)$. The convergence properties of Q-learning can be described by using stochastic approximation techniques [17]. A necessary condition for convergence to the optimal Q-function is that all state action pairs are continuously updated.

The Q-learning algorithm for a discretized version of (10) is as follows:

1) At $k = 0$, initialize $\hat{Q}_k(\Omega, u)$ $\forall$ state-action pairs (Ω, u) and select initial state Ω.
2) Choose $u = \arg\min \hat{Q}_k(\Omega, u)$ with probability $1 - \alpha_t$, else let u be a random exploratory action.
3) Carry out action u. Let the next state be Ω', and the cost be $\Phi(\Omega, u)$. Update the $\hat{Q}$ values
$$\begin{aligned}
\hat{Q}_{k+1}(\Omega, u) &= \big((1 - \beta_k)\hat{Q}_k(\Omega, u) + && (25)\\
&= \beta_k(\Phi(\Omega, u) + \min_u \hat{Q}(\Omega', u))\big)
\end{aligned}$$
4) Set the current state to Ω', increment k and go to step 2.

The parameter sequence $0 < \beta_k < 1$ is called the learning rate and controls how new information is averaged with existing estimates of the Q-function. Step 2 ensures that all state-action pairs are continuously explored. Step 3 averages the current estimate of the Q-function $\hat{Q}_k(\Omega, u)$ with the cost of choosing action u and then following the best estimated optimal policy thereafter. This results in a new estimate $\hat{Q}_{k+1}(\Omega, u)$.

The sequence β_k strongly influences the rate and type of convergence of Q-learning. From stochastic approximation it is known that with $\beta_k = \frac{1}{k}$ and under conditions on the underlying Markov chains, the Q-learning algorithm will converge with probability one. But since the learning parameter is decreasing, the Q-function estimates become less and less responsive to new information. In practice, this means the Q estimates will not adapt to changes in the underlying probability distributions. An alternative used by CAES is to use a small fixed step size $\beta_k = b$. It can be shown [21] that $\hat{Q}_k(\Omega, u)$ converges to a region around the optimal Q values

$$P\Big[\Big\| Q - \hat{Q}_k \Big\| \geq \alpha \Big| \hat{Q}_0 \Big] \leq A_1(b) + A_2(\hat{Q}_0) \exp(-h(b)k) \tag{26}$$

where $\hat{Q}_0$ is the initial estimate of the Q-function. The error bound for the Q-function estimates decays exponentially with the iteration index, but does not decay to zero. The term $A_1(b)$ is a decreasing function of the parameter b and can be made arbitrarily small. The term $A_2(\hat{Q}_0) \exp(-h(b)k)$ reflects the error in the initial estimate Q_0 and decays exponentially. The

TABLE I

RELEVANT SIMULATION PARAMETERS

Θ	0.4
γ	0.8
β	0.05
b_1, τ_1, d_1	$10, 0.6, 1$
b_2, τ_2, d_2	$2, 0.2, 2$
b_3, τ_3, d_3	$2, 0.1, 8$

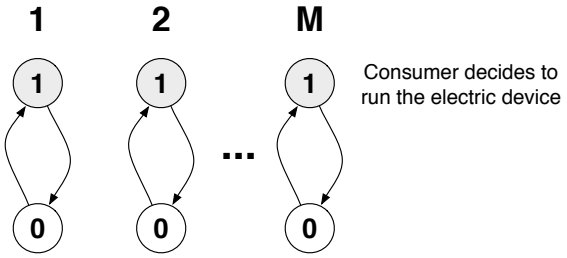

Fig. 2. Binary Markov processes modeling the electrical devices.

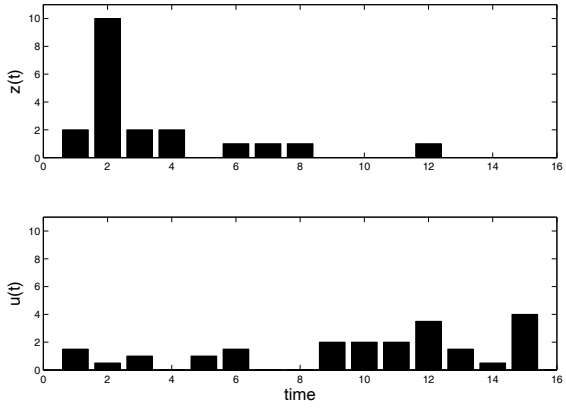

Fig. 3. Example of DR scheduling induced by CAES.

term $h(b)$ is decreasing in b. Thus, b controls both the rate of convergence and the accuracy of the estimates. The advantage of this approach is that if the underlying probability structure changes, then the algorithm will adapt to this change.

VI. NUMERICAL RESULTS

In this section, we present a numerical example of CAES. As expected, CAES has the desired DR behavior, shifting the time of operation and allocation of energy in response to pricing signals. We also explore the impact of different parameters on the behavior of the algorithm. For clarity, we consider the simplified case with $M = 3$ independent devices. Thus, devices are selected independently by consumers. We discretize the variables $x(t)$, $y(t)$, $z(t)$ and $p(t)$ so the state variable $\Omega(t)$ takes values on a bounded discrete state space. The learning rate parameter is set to a constant $\beta_k = \hat{\beta}$. Table I lists simulation parameters which are described below.

The dis-utility function captures the cost of waiting, see Section II-A, and is $U_m(y_m(t)) = y_m(t)^2$. This function captures the dis-utility of the consumer of having to wait for a device to finish its work. The longer a device is delayed, the larger y_m becomes, and the larger the dis-utility experienced by the consumer. The objective is then

$$\lim_{T\to\infty} \mathbf{E}\left[\sum_{t}^{T} \gamma^t \left(\sum_{m}^{M} p(t)u_m(t)) + y_m^2(t)\right)\right]. \qquad (27)$$

As shown in Figure 1, the CAES algorithm is driven by two exogenous process $z(t)$ and $p(t)$ and yields the energy control process $u(t)$. We first describe how the input processes were generated and then describe the performance of the algorithm. The consumer selection process $z(t)$ follows M independent bursty Markov chains taking values of zero or one. (See Fig. 2 for a graphical representation.) If $z_m(t)$=1, then d_m units of energy are requested by the system. The transition probabilities are defined as

$$P[z_m(t+1) = 0 | z_m(t) = 1] = \frac{1}{b_m}, \qquad (28)$$

$$P[z_m(t+1) = 1 | z_m(t) = 0] = \frac{\tau_m}{b_m(1-\tau_m)}, \qquad (29)$$

where b_m and τ_m are the burstiness and the steady-state probability of state 1. Likewise, the energy price process is driven by a binary Markov chain.

Figure 3 shows a representative aggregate consumer selection trace and associated aggregate energy control policy trace and illustrates the desired DR behavior. The vertical axis is in normalized energy units and the horizontal axis units of time. The upper plot is of $1^T z(t)$ the total energy selected by the consumer in time t. The lower plot is $1^T u(t)$ the total energy allocated by the CAES algorithm at time t. In time period one, a consumer selects a device requiring 2 units of energy and the optimal control policy allocates approximately this quantity, reflecting CAES's estimate that current energy prices are lower than future prices. In time period two, a consumer selects a device requiring 10 units of energy and CAES defers allocating this energy until several periods later, reflecting CAES's estimate that current prices are higher than estimated future prices. Similarly, energy use is time shifted from periods 3 and 4 to later periods. Note that the deferral of energy allocation increases the value of $y(t)$ in the next time-periods, reflecting a larger dissatisfaction of the consumer.

Figures 4 and 5 illustrate the behavior of CAES as parameters are adjusted. In Figure 4 average energy cost per period is graphed against different values of the parameter λ for three different values of θ. We recall that λ controls the tradeoff between the importance of the financial cost and dis-utility function. The larger λ, the more important the satisfaction of the consumer. The dotted line is the normalized energy cost if no scheduling is done and devices are run when selected $u(t) = z(t)$. As can be seen, CAES reduces energy costs for all values of the parameters selected. As λ is increased, the effective cost of delay is increased, causing the algorithm to less aggressively reschedule devices, resulting in greater energy costs. As θ is increased, the average pending energy increases. This has the effect of increasing the cost of delay and decreasing average energy usage.

While CAES has the effect of reducing energy costs, it does so at the cost of delaying selected device operation. An aggregate measure of this is the average pending energy backlog $\mathbf{E}[1^T x]$. Figure 5 shows the average pending energy

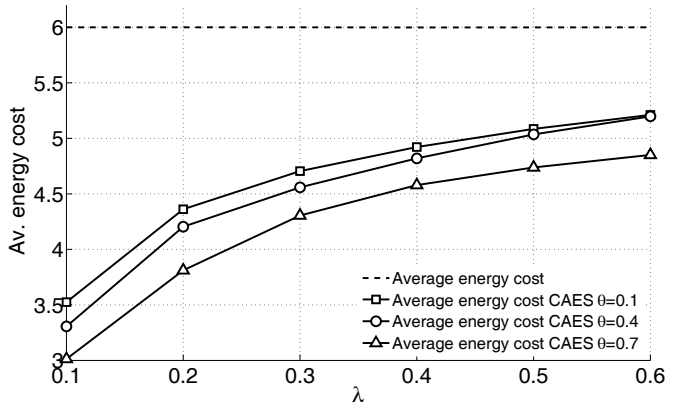

Fig. 4. Financial cost as a function of the tradeoff parameter λ. The parameter setting is reported in Table I.

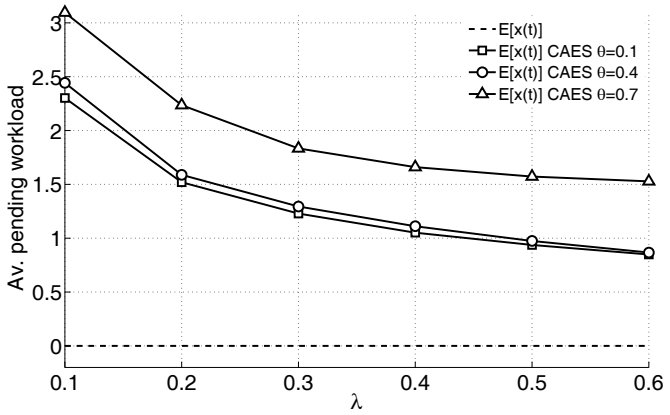

Fig. 5. Pending energy as a function of the tradeoff parameter λ. The parameter setting is reported in Table I.

backlog as a function of different values of λ. As before, the dotted line corresponds to the policy $u(t) = z(t)$. As λ increases, the average pending energy backlog decreases, reflecting the increased cost of delay and mirroring the results of figure 4. As θ is increased, the average pending backlog decreases, reflecting the impact of the smoothing parameter on the dis-utility functions.

VII. Conclusion

CAES is an energy management system for residential demand response applications. CAES reduces average residential energy costs and smooths energy usage. The approach is based on Q-learning and learns the impact of future energy prices and possible future consumer device selections on current energy decisions. The underlying probabilistic models are assumed Markov, but the underlying transition probabilities are not assumed known. Because of the constant learning parameter, CAES converges to a region around the optimum Q-functions. As a consequence, CAES adapts to changes in a consumer's preferences or the probabilistic structure of energy market pricing. Simulations demonstrate that the proposed algorithm enables a reduction of the average end-user financial cost up to 40% with respect to price-unaware energy allocation. Moreover, it effectively avoids pricing peaks and smooths the energy expense, with a beneficial effect to the overall utility network and the consumer.

References

[1] S. Borenstein, M. Jaske, and A. Rosenfeld, "Dynamic pricing, advanced metering, and demand response in electricity markets," *UC Berkeley: Center for the Study of Energy Markets*, Oct. 2002. [Online]. Available: http://www.escholarship.org/uc/item/11w8d6m4

[2] S. Braithwait and K. Eakin, "The role of demand response in electric power market design," *Edison Electric Institute*, 2002. [Online]. Available: http://www.eei.org/industry_issues/retail_services_and_delivery/wise_energy_use/demand_response/demandresponserole.pdf

[3] G. Barbose, C. Goldman, and B. Neenan, "A survey of utility experience with real time pricing," *Lawrence Berkeley National Laboratory: Lawrence Berkeley National Laboratory*, 2004. [Online]. Available: http://www.escholarship.org/uc/item/8685983c

[4] J. Roos and I. Lane, "Industrial power demand response analysis for one-part real-time pricing," *Power Systems, IEEE Transactions on*, vol. 13, no. 1, pp. 159 –164, feb 1998.

[5] M. A. Piette, O. Sezgen, D.Watson, N. Motegi, C. Shockman, and L. ten Hope, "Development and evaluation of fully automated demand response in large facilities," Jan. 2005. [Online]. Available: http://escholarship.org/uc/item/4r45b9zt

[6] K. Herter, "An exploratory analysis of California residential customer response to critical peak pricing of electricity," *Energy*, vol. 32, no. 1, pp. 25 – 34, Jan. 2007. [Online]. Available: http://www.sciencedirect.com/science/article/B6V2S-4JG5F91-2/2/bb70d546082f9f5483829aabeef5279e

[7] ——, "Residential implementation of critical-peak pricing of electricity," *Energy Policy*, vol. 35, no. 4, pp. 2121 – 2130, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/B6V2W-4KSSWHP-2/2/57823b87cba8355805b5896909d1f016

[8] A. Faruqui and S. George, "Quantifying customer response to dynamic pricing," *The Electricity Journal*, vol. 18, no. 4, pp. 53 – 63, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/B6VSS-4G1WY67-3/2/44466f47c4dd993cbf13c290bd91dc97

[9] E. Koch and M. Piette, "Architecture concepts and technical issues for an open, interoperable automated demand response infrastructure," in *Grid Interop Forum*, Albuquerque, NM, US, Nov. 2007.

[10] M. LeMay, R. Nelli, G. Gross, and C. A. Gunter, "An integrated architecture for demand response communications and control," *in Proc. of the 41st Hawaii International Conference on System Sciences*, 2008.

[11] M. A. Piette, D.Watson, N. Motegi, and S. Kiliccote, "Automated critical peak pricing field tests: 2006 pilot program description and results," in *LBNL Report 62218*, Albuquerque, NM, US, May 2007.

[12] R. Sutton and A. Barto, *Reinforcement learning*. MIT Press, 1998.

[13] S. Kiliccote, M. A. Piette, D. S. Watson, and G. Hughes, "Dynamic controls for energy efficiency and demand response: framework concepts and a new construction study case in New York," in *in Proc. of the 2006 ACEEE Summer Study on Energy Efficiency in Buildings*, Pacific Grove, CA, US, Aug. 2006.

[14] Q. Dam, S. Mohagheghi, and J. Stoupis, "Intelligent demand response scheme for customer side load management," in *Energy 2030 Conference, 2008. ENERGY 2008. IEEE*, 17-18 2008, pp. 1 –7.

[15] H. Varian, *Microeconomic Analysis*. Boston: W. W. Norton, 1984.

[16] D. Duffee, *Security Markets Stochastic Models*. San Diego: Academic Press, 1988.

[17] S. Meyn, *Control Techniques for Complex Networks*. New York, NY: Cambridge University Press, 2008.

[18] D. Bertsekas, *Dynamic Programming and Optimal Control*. Massachusetts: Athena Scientific, 2005.

[19] W. Powell, *Approximate Dynamic Programming*. Hoboken, NJ: John Wiley and Sons, 2007.

[20] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge Univ Pr, 2004.

[21] V. Borkar and S. Meyn, "The ode method for convergence of stochastic approximation and reinforcement learning," *SIAM Journal on Control and Optimization*, vol. 38, no. 2, pp. 447–469, 2000.