Optimal Demand Response Using Device Based Reinforcement Learning

Zheng Wen, Hamid Reza Maei and Daniel O'Neill
Dept. of Electrical Engineering, Stanford University, Stanford, CA 94305 USA
e-mail: zhengwen, maei, dconeill@stanford.edu

***Abstract*— We present a reinforcement learning based approach to Demand Response. Demand Response (DR) adjusts electrical energy demand in response to dynamic energy pricing or other electrical grid signals. By suitably adjusting energy prices, load can be shifted from peak energy consumption periods to other periods, either by delaying energy usage or, in some predictable situations, to an earlier time period, improving operational efficiency and reducing emissions. The variability of renewables can create an additional need to shift energy consumption. Critical to a successful DR approach is learning the consequences of deferring energy consumption on consumer satisfaction, cost, and future energy behavior. This is a particular problem for residential or small commercial building energy consumers where it is often difficult to cost effectively model the behavior of consumers, to explicitly illicit their energy time of use preferences, or explicitly understand the interaction of energy consuming devices. We use reinforcement learning methods to address these issues. The approach learns the energy behavior of consumers and optimally schedules the operation of devices to minimize the tradeoff between consumer dis-satisfaction with energy delay and energy cost. The approach comprehends both consumer driven energy requests (reservations) but also, using exploration methods, can explicitly anticipate requests, where this will reduce cost or improve satisfaction such as heating or cooling. We explore Q-Learning with eligibility traces and importance-weighting to improve sample efficiency.**

I. Introduction

Demand response (DR) systems [1]–[3] dynamically adjusts electrical demand in response to changing electrical energy prices or other grid signals. DR offers several benefits. By suitably adjusting energy prices, load can be shifted from peak energy consumption periods to other periods, either by delaying energy usage or, in some predictable situations, to an earlier time period, improving operational efficiency and reducing emissions. This, in turn, can improve operational efficiency, reduce operating costs, improve capital efficiency, and reduce harmful emissions and risk of outages. The variability of renewables can create an additional need to shift energy consumption in order to better match energy demand with unforecasted changes in electrical energy generation. The benefit is a possible reduction in backup (ancillary) generation frequently used to hedge renewable sources. DR has been extensively investigated for larger energy users and has been implemented in many areas (*e.g.*, [4], [5]).

Residential and small building DR [6]–[9] offers similar potential benefits. DR for residential and small commercial buildings was estimated to account for as much as 65% of the total energy savings potential of DR. But residential and commercial buildings face several challenges. Technical challenges include the deployment of an infrastructure supplying real-time pricing information to energy consumers in a useful way, networking devices, ensuring security, and advanced metering [9], [10]. There are also challenges in capturing the consequences of deferring energy consumption in small commercial buildings, where detailed energy usage and energy flow models are frequently not available, and in relieving energy consumers from having to make a long sequence of explicit energy consumption decisions. In small commercial buildings it can be impractical to fully model the behavior of heating and cooling systems and their perceived impact on occupant comfort. 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, especially when many of these decisions will have limited short term financial impact on the consumer [11]. As a consequence, we believe developing fully-automated Energy Management Systems (EMS) [9], [12] are a necessary prerequisite to DR in residential and small building settings.

Critical to a successful DR EMS approach is learning the consequences of deferring energy consumption on consumer satisfaction, cost, and future energy behavior. In [11] we present a residential EMS approach that used reinforcement learning to learn energy consumer's behavior and automatically make optimal energy scheduling and allocation decisions in the face of uncertain future energy prices. This approach assumed consumer dis-satisfaction with delay could be captured by (known) dis-utility functions and that all energy usage was explicitly initiated by a consumer energy request or reservation (e.g. press of a button). Both energy prices and consumer energy requests are modeled a Markov processes with unknown distributions.

Here, we extend this work by removing these two assumptions and extend it to residential and small commercial buildings applications. The approach samples consumer dis-utility with different scheduling policy options and learns the costs associated with each possible policy. We also consider a device centered point of view. First a device can receive a request from a consumer and schedule an optimal time to run it and second the device can initiate a request itself, speculating that the device will be needed

at a forecastable time. Probing is used to allow the device initiated patterns to find the best time to run the device to maximize consumer satisfaction. We model the problem as a discounted cost infinite horizon Markov decision problem with unknown transition probabilities and with sampled delay dis-utility values. We use Q-learning [13], a type of temporal-difference learning, to allow the algorithm to learn the behaviors of consumers and to optimally make energy consumption scheduling decisions. We explore Q-Learning with eligibility traces and importance-weighting to improve sample efficiency. The simulation result in Section IV indicates that our proposed Q-learning algorithm reduces the consumer's cost by 56% in an illustrative example.

The remainder of the paper is organized as follows: in Section II, we pose the optimal demand response problem as an infinite horizon discounted Markov decision process (MDP), and decomposes this high-dimensional MDP into a collection of low-dimensional MDPs under suitable assumptions; in Section III, we propose a Q-learning algorithm with eligibility traces and importance-weighting as the RL-EMS algorithm; simulation results are shown in Section IV and we conclude in Section V.

II. Device Based MDP Model

This section is organized as follows: Subsection II-A briefly discusses the Reinforcement Learning based Energy Management System (RL-EMS) and consumer requests; Subsection II-B reviews the concept of utility function and discusses the functional form of the "dis-utility function" of the electricity consumer; Subsection II-D formulates the optimal demand response problem as a collection of device based MDPs, the performance metric is described in II-C and the dynamic programming (DP) solution to such MDPs are presented in Subsection II-E. In this section we assume that the probabilistic properties of consumer behavior and electricity pricing is known as are the precise forms of consumer utility functions. In Section III, we relax these unrealistic assumptions using reinforcement learning techniques.

A. RL-EMS and Consumer Requests

Demand response (DR) is a key component of the smart grid and enables the dynamic adjustment of electrical demand in response to pricing signals. It is well-known that DR offers benefits in both the consumer level and the power system level. Specifically, DR not only can optimize an electricity consumer's utility by shifting her requests from periods of high electricity price to other periods, but also has the potential to improve the "social welfare" of the whole power system if the electricity price is suitably adjusted. As a result, DR has been widely investigated and implemented in many areas.

As has been pointed out in O'Neill et al. [11], the "decision fatigue" of electricity consumers in the residential sector necessitates the development of Reinforcement Learning based Energy Management Systems (RL-EMS), which are algorithms that learn a consumer's behavior and then automatically make optimal consumer request scheduling and prediction decisions for smart devices. Specifically, in the vision of smart grid, we conjecture that future RL-EMS should perform the following functions:

- RL-EMS receives requests from the consumers, and then schedules when to fulfill the received requests. We henceforth refer to this case as a *requested job*.
- If a smart device managed by the RL-EMS is idle (i.e. currently there is no request for that device), RL-EMS could speculatively power a device. We henceforth refer to this case as a *speculative job*. For instance, in a small commercial building, the RL-EMS algorithm might speculatively turn on the building's air conditioning in advance of the tenets arrival to capture early morning lower energy costs or to mask the latency of cooling the building. Notice that we should not allow RL-EMS to do speculative jobs on all the smart devices (such as dishwasher).

We assume time is discrete $t = 0, 1, \cdots$ and that there are N smart devices managed by the RL-EMS and numbered $n = 1, 2, \cdots, N$. To simplify exposition, we assume that all the jobs done by device n are standardized and hence they can be completed in one time step and consume a constant energy $C(n)$, which only depends on the type of the smart device. This assumption can be readily relaxed to devices with different operating periods. We further assume that the RL-EMS will ignore all the consumer requests to device n if device n currently has an unsatisfied request, but the RL-EMS allows a consumer to cancel an existent unsatisfied request. Specifically, if the consumer wants to replace an existent request with a new request, he must cancel the existent request first, and then start the new request. Notice under this assumption, at each time a smart device has at most one unsatisfied request.

Each consumer request is represented by a four-tuple $J = (n, \tau_r, \tau_g, g)$, where

- n denotes the requested device;
- τ_r is the *request time* and denotes when the RL-EMS receives this request;
- τ_g is the *target time* and denotes when the consumers prefer this request to be satisfied;
- g denotes the *priority* of this request, and higher priority implies the "stronger preference" of the consumer that he wants the request to be satisfied at a time close to the target time τ_g.

Notice that the target time τ_g is not necessarily equal to the request time τ_r, specifically, the consumers might request to use a device in a later time. Thus, we only require that $\tau_g \geq \tau_r$.

We observe that in practice, the consumer does not allow a request to be delayed for an arbitrarily long time. In addition, for a request $J = (n, \tau_r, \tau_g, g)$, it is unreasonable to assume that $\tau_g - \tau_r$, the difference between the target time and the request time, can be arbitrarily large. Thus, in this paper, we assume that (1) for any request $J = (n, \tau_r, \tau_g, g)$, its target time τ_g must satisfy $\tau_r \leq \tau_g \leq \tau_r + W(n)$, and (2) if the request $J = (n, \tau_r, \tau_g, g)$ is not fulfilled by time $\tau_r + W(n)$,

then it will be cancelled by the consumer for sure, where $W(n)$ is a known time window and only depends on the type of device.

The difference between our RL-EMS model and those in previous literatures (e.g. [11]). To the best of our knowledge, our RL-EMS model is more general in the following aspects: (1) RL-EMS is allowed to perform speculative jobs; (2) a request's target time can be different from its request time, and requests are allowed to have different priorities; (3) an unsatisfied request can be canceled by the consumer.

B. Consumer Preference and Dis-utility Function

To formalize the notion of optimal demand response, we define a performance metric which captures the electricity consumer's preferences called *Utility functions*, often used in the economics literature [14]. We assume utility functions are concave, strictly increasing and satisfy the following properties:

- A rational consumer prefers low electricity price to high electricity price.
- For a requested job sent to the RL-EMS, a rational consumer prefers that job is completed at a time close to the target time, and the higher the request's priority is, the "stronger" this preference is.
- For a speculative job done by the RL-EMS, a rational consumer will be "happy" if the RL-EMS has done a "good" job, and will be "unhappy" if it has done a "bad" job scheduling a device.
- When a rational consumer decides to cancel an unsatisfied requested job, he is usually "unhappy" if the target time has already passed.
- A rational consumer usually cares more about her spending and feeling at current time than those in the future.

For notational convenience we work with the negative of utility functions and call them dis-utility functions

$$\bar{U}^{(t)}\left[(\mathbf{z}(s),\mathcal{J}(s)),\,\forall 0\le s\le t\right]. \tag{1}$$

We use the following notation for the arguments of the dis-utility function. For each time t, we define disjoint sets of smart devices $\mathcal{D}(t)\subseteq\{1,2,\cdots,N\}$ as $\mathcal{D}(t)=\{\text{devices that do a job at time } t\}$. Furthermore, we use $\mathbf{z}(t)\in\{0,1\}^N$ to denote the status of the set of devices at time t. Specifically, let $z(t,n)$ denote the nth component of $\mathbf{z}(t)$, then $z(t,n)=1$ indicates that device n is on at time t and $z(t,n)=0$ indicates that device n is off at time t. Let $\mathcal{J}(t)$ denote the set of unsatisfied requests at time t. Recall that at time t device n has at most one unsatisfied request, thus we use $J(t,n)$ to denote the unsatisfied request of device n at time t if that request exists; otherwise, we set $J(t,n)=$ NULL.Notice that $(\mathbf{z}(s),\mathcal{J}(s)),\,\forall 0\le s\le t$ completely specifies the "history" of device status, consumer behavior and the RL-EMS decisions until time t.

We make the following assumption:

Assumption 1: *For any $t\ge 0$, the dis-utility function $\bar{U}^{(t)}$ is additive over the devices, that is*

$$\begin{aligned}&\bar{U}^{(t)}\left\{(\mathbf{z}(s),\mathcal{J}(s)),\,\forall 0\le s\le t\right\}\\ =&\sum_{n=1}^{N}\bar{U}^{(t,n)}\left[(z(s,n),J(s,n)),\,\forall 0\le s\le t\right],\end{aligned} \tag{2}$$

where $\bar{U}^{(t,n)}$ captures the consumer's dis-utility at time t for Device n and $z(s,n)$, $J(s,n)$ is defined above.

Furthermore, in this paper, we assume the dis-utility function $\bar{U}^{(t,n)}$ takes the following forms:

- If Device n satisfies request $J(t,n)=(n,\tau_r,\tau_g,g)$ at time t, we assume

$$\bar{U}^{(t,n)}\left[(z(s,n),J(s,n)),\,\forall 0\le s\le t\right]=\tilde{U}_r^{(n)}(t-\tau_g,g), \tag{3}$$

  where the subscript "r" denotes that it is the dis-utility incurred when a request is satisfied, and g is the priority of the request. Note $t-\tau_g$ captures not only the distance between the current time t and the request's target time τ_g, but also whether or not the target time has passed. In practice, $\tilde{U}_r^{(n)}$ should be small when $t-\tau_g$ is close to 0 and increases as $t-\tau_g$ deviates from 0; furthermore, the higher the priority g, the higher this increase rate will be.
- Similarly, if the request $J(t,n)=(n,\tau_r,\tau_g,g)$ is cancelled by the consumer at time t, we assume

$$\bar{U}^{(t,n)}\left[(z(s,n),J(s,n)),\,\forall 0\le s\le t\right]=\tilde{U}_c^{(n)}(t-\tau_g,g), \tag{4}$$

  where the subscript "c" denotes that it is the dis-utility incurred when a request is cancelled. In practice, $\tilde{U}_c^{(n)}$ should be very small if $t<\tau_g$, and it will increase with t when $t\ge\tau_g$; furthermore, the higher the priority g, the higher this increase rate will be.
- If Device n does a speculative job at time t, we assume that

$$\bar{U}^{(t,n)}\left[(z(s,n),J(s,n)),\,\forall 0\le s\le t\right]=\tilde{U}_s^{(n)}(t-\tau_p), \tag{5}$$

  where the subscript "s" denotes that it is the dis-utility incurred when a speculative job is done, and τ_p-1 is the time when the previous job on Device n (either requested or speculative) is completed or cancelled. In practice, $\tilde{U}_s^{(n)}$ should decrease as $t-\tau_p$ increases; the large $\tilde{U}_s^{(n)}$ for small $t-\tau_p$ prevents too frequent speculative jobs.
- Otherwise, we assume that

$$\bar{U}^{(t,n)}\left[(z(s,n),J(s,n)),\,\forall 0\le s\le t\right]=0.$$

It is worth pointing out that the dis-utility function described in this section is quite general, and it is very challenging to derive the specific functional forms of $\tilde{U}_r^{(n)}$, $\tilde{U}_c^{(n)}$ and $\tilde{U}_s^{(n)}$.

C. Performance Metric

We assume the cost function of the consumer at time t has the following form:

$$P(t)\sum_{n\in\mathcal{D}(t)}C(n)+\gamma\bar{U}^{(t)}\left[(\mathbf{z}(s),\mathcal{J}(s)),\,\forall 0\le s\le t\right], \tag{6}$$

where $P(t)$ is the electricity price at time t and $\bar{U}$ is a dis-utility function capturing the consumer's "unhappiness" at time t. Specifically, notice that $\sum_{n\in\mathcal{D}(t)}C(n)$ is the

total electricity energy consumed at time t, and hence $P(t)\sum_{n\in\mathcal{D}(t)} C(n)$ is the electricity bill the consumer pays at time t. We assume the consumer's dis-utility at time t depends on the "history" $(\mathbf{z}(s), \mathcal{J}(s))\,, \forall 0 \leq s \leq t$ and $\gamma > 0$ represents the tradeoff between the electricity bill paid and the consumer's dis-utility.

Notice that from the RL-EMS's perspective, both the electricity price and the consumer behavior are exogenous and stochastic, thus, in this paper, we assume that RL-EMS aims to minimize the expected infinite-horizon discounted cost:

$$\mathbb{E}\sum_{t=0}^{\infty}\alpha^t\left[\sum_{n\in\mathcal{D}(t)} P(t)C(n) + \gamma\bar{U}^{(t)}\left[(\mathbf{z}(s),\mathcal{J}(s))\,,\forall 0\leq s\leq t\right]\right],$$

where $0 < \alpha < 1$ is the discrete-time discount and captures the assumption that a rational consumer cares more about her spending/feeling at current time than those in the future. Notice that in this problem formulation, the state at time t is $(P(t), \mathbf{z}(s)\,\forall 0 \leq s \leq t-1, \mathcal{J}(s)\,\forall 0 \leq s \leq t)$, and the action is $\mathbf{z}(t)$.

Under Assumption 1, the dis-utility function is decomposable and the cost function can be written as

$$\sum_{n=1}^{N}\left[\mathbb{E}\left\{\sum_{t=0}^{\infty}\alpha^t\left[P(t)C(n)\mathbf{1}(n\in\mathcal{D}(t))\right.\right.\right.$$
$$\left.\left.\left. +\ \gamma\bar{U}^{(t,n)}\left[(z(s,n),J(s,n))\,,\forall 0\leq s\leq t\right]\right]\right\}\right]. \quad (7)$$

D. Device Based MDP Model

In this subsection we formulate the problem as Markov Decision Process, MDP. We make the following probabilistic assumption:

Assumption 2: *Both the electricity price $P(t)$ and the consumer requests to the RL-EMS follow exogenous Markov chains. Furthermore, we assume that*

- *Electricity price process is independent of the consumer requests process.*
- *Consumer requests to different devices are independent.*

Under both Assumptions 1 and 2, our objective is to find the optimal scheduling policy to minimize (7). We can reformulate this as an infinite-horizon MDP,

$$\sum_{n=1}^{N}\left[\min\mathbb{E}\left\{\sum_{t=0}^{\infty}\alpha^t\left[P(t)C(n)\mathbf{1}(n\in\mathcal{D}(t))\right.\right.\right.$$
$$\left.\left.\left. +\ \gamma\bar{U}^{(t,n)}\left[(z(s,n),J(s,n))\,,\forall 0\leq s\leq t\right]\right]\right\}\right], \quad (8)$$

which decomposes over devices.

In the remainder of this paper, we focus on deriving the optimal scheduling/prediction policy for a single device and will drop the superscript n. For example, we will use W instead of $W(n)$ to denote the time window and represent a request as $J = (\tau_r, \tau_g, g)$. We also use the term "smart device" and "RL-EMS" interchangably henceforth, since due to the decomposition of the problem, one can think each smart device has its own RL-EMS.

We further assume that the timeline for a smart device can be divided into "episodes". Specifically, we assume that whenever the smart device completes a job (either speculative or requested) or the current unsatisfied request is canceled by the consumer, the current episode terminates. In the next time step, the smart device "regenerate" its state according to a fixed distribution π_0 and a new episode starts. Thus, each episode in the timeline corresponds to a finite-horizon MDP. The notion of episode is illustrated in Figure 1.

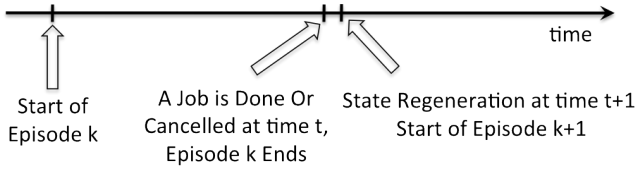


Fig. 1. Illustration of the notion of episode

The state of the finite-horizon MDP at time t is

$$x(t) = [P(t), s(t), g(t)]^T \in \mathcal{S},$$

where $P(t)$ is the exogenous electricity price at time t, $s(t)$ is the *elapsed time* at time t, $g(t)$ is the priority of request at time t and $\mathcal{S}$ is the state space. Specifically, we define

$$s(t) = \begin{cases} t-\tau_p & \text{no request received in the current episode} \\ t-\tau_g & \text{otherwise} \end{cases}$$

where τ_p is the start time of the current episode, and τ_g is the target time of the received request. Furthermore, once a request is received in the current episode, we assume its priority $g(t) \in \{1, 2, \cdots, g_{max}\}$; on the other hand, we use $g(t) = 0$ to denote that no request has yet been received in the current episode.

Since the electricity price is exogenous, we can partition the state $x(t) = [P(t), s(t), g(t)]^T$ as the "price portion" $P(t)$ and "device portion" $[s(t), g(t)]^T$. The "device portion" of the MDP state transition model is summarized in Figure 2, notice that there are $(2W+1)g_{max}+\hat{W}+1$ "device portion" states. If the price Markov chain has P_{max} state variables, the cardinality of the state space for the device based MDP is $|\mathcal{S}| = P_{max}\left[(2W+1)g_{max}+\hat{W}+1\right]$, which is polynomial in P_{max}, W, $\hat{W}$ and g_{max}. We now describe the device based MDP model in detail:

- Recall $P(t)$ is assumed to follow an exogenous Markov Chain with P_{max} state variables. The transition probability from $P(t)$ to $P(t+1)$ is denoted as $Pr(P(t+1)|P(t))$.
- If the smart device has not received a consumer request in the current episode, recall that we set $g(t) = 0$ and $s(t) = t - \tau_p$. The current action space is $\mathcal{A}(x(t)) = \{\text{off}, \text{on}\}$. Notice that:
  1) If action "off" is selected, the current cost is $\Phi(x(t), a(t), x(t+1)) = 0$. Then the smart device receives a consumer request $(t+1, \tau_g, g)$ at the next time step (time $t+1$) with probability $p_{s(t),t+1-\tau_g,g}$, where $t+1 \leq \tau_g \leq t+1+W$ and $g \in \{1, 2, \cdots, g_{max}\}$. In other words, the MDP transits to the state $x(t+1) = [P(t+1), t+1-\tau_g, g]^T$ with probability $p_{s(t),t+1-\tau_g,g}$, for any $t+1 \leq \tau_g \leq t+1+W$ and any $g \in \{1, 2, \cdots, g_{max}\}$ (notice that the target time and priority together specify the

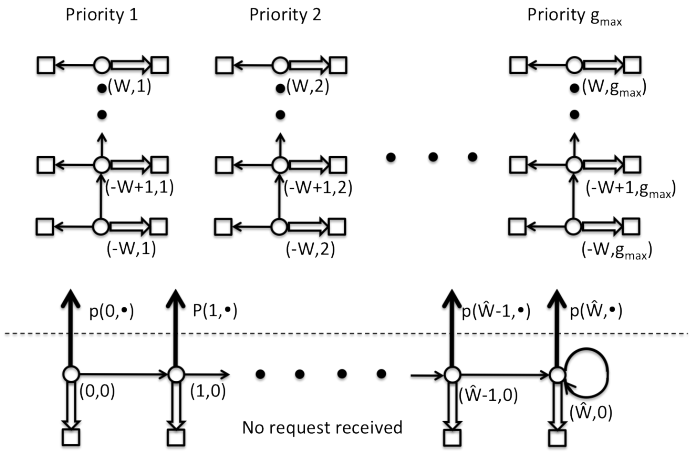


Fig. 2. The state transition model of the "device portion". Notice that each circle corresponds to a "device portion" state $[s(t), g(t)]^T$ and each square corresponds to the termination of the current episode and regeneration in the next time step. The hallow arrows indicate the state transitions under action "on", while the line arrows indicate the state transitions under action "off". The bold line arrows across the dotted line indicate the fact that the states below the dotted line can transit to many states above the dotted line, since there are $(W+1)g_{max}$ types of requests.

type of the received request). On the other hand, with probability $1-\sum_{\tau_g=t+1}^{t+1+W}\sum_{g=1}^{g_{max}} p_{s(t),t+1-\tau_g,g}$, the smart device does not receive the consumer request at time $t+1$ and transits to the state $x(t+1)=[P(t+1), s(t)+1, 0]^T$.

2) If action "on" is selected, the current cost is $\Phi\left(x(t), a(t), x(t+1)\right) = P(t)C(n)+\gamma\tilde{U}_s\left(s(t)\right)$, where $\tilde{U}_s\left(s(t)\right)$ is defined in (5). Then the current episode terminates and the smart device regenerates its state based on distribution π_0.

Notice that the transition probability $p_{s(t),t+1-\tau_g,g}$ and the dis-utility $\tilde{U}_s\left(s(t)\right)$ depends on $s(t)$, in order to ensure the state space is finite, we assume that if $s(t)\geq\hat{W}$, we have $p_{s(t),t+1-\tau_g,g}=p_{\hat{W},t+1-\tau_g,g}\,\forall\tau_g\,\forall g$, and $\tilde{U}_s\left(s(t)\right)=\tilde{U}_s\left(\hat{W}\right)$. [1]

- If the smart device has already received a consumer request in the current episode but has not satisfied this request, recall we set $s(t)=t-\tau_g$, where τ_g is the target time of this request. Notice that the action space $\mathcal{A}\left(x(t)\right)=\{\text{off, on}\}$. Notice that:
  1) If action "off" is selected, then with probability $\tilde{p}_{s(t),g(t)}$, nothing will occur and the MDP transits to the state $x(t+1)=[P(t+1), s(t)+1, g(t)]^T$ and the cost associated with this transition is $\Phi\left(x(t), a(t), x(t+1)\right)=0$. On the other hand, with probability $1-\tilde{p}_{s(t),g(t)}$, the consumer will cancel this request, then the current episode terminates and the smart device regenerates its state based on distribution π_0. The cost associated with this transition is $\Phi\left(x(t), a(t), x(t+1)\right)=\gamma\tilde{U}_c(s(t), g(t))$, where $\tilde{U}_c$ is defined in (4). Notice that if $s(t)=W$, $\tilde{p}_{s(t),g(t)}=0$.
  2) If action "on" is selected, the current cost is $\Phi\left(x(t), a(t), x(t+1)\right) = P(t)C(n)+\gamma\tilde{U}_r\left(s(t), g(t)\right)$, where $\tilde{U}_r\left(s(t), g(t)\right)$ is defined in (3). Then the current episode terminates and the smart device regenerates its state based on distribution π_0.

[1] That is, when $s(t)\geq\hat{W}$ and action "off" is selected, with probability $1-\sum_{\tau_g=t+1}^{t+1+W}\sum_{g=1}^{g_{max}} p_{\hat{W},t+1-\tau_g,g}$, the smart device will stay at the same state.

E. Dynamic Programming Solution

If the transition model of the device based MDP and the dis-utility function of the consumer are known, the device based MDP in each episode can be solved by finite-horizon dynamic programming (DP). Specifically, once an action "on" is selected, the current episode terminates and the smart device regenerates its state and starts a new episode in the next time step. Thus, in each episode, the device based MDP is an optimal stopping problem. Following ideas in classical DP, in this subsection, we compute the optimal Q-function based on backward induction.

Recall that $x(t)=[P(t), s(t), g(t)]^T$ and $\mathcal{A}(x(t))=\mathcal{A}=\{\text{off, on}\}$, we have

- If the smart device has not received a consumer request in the current episode, we have

$$Q^*\left(x(t), \text{on}\right)=P(t)C(n)+\gamma\tilde{U}_s\left(s(t)\right),$$

and $Q^*\left(x(t), \text{off}\right)=$

$$\alpha\mathbb{E}_{P(t+1)}\left[\sum_{s'=-W}^{0}\sum_{g=1}^{g_{max}} p_{s(t),s',g}\min_{a\in\mathcal{A}}Q^*\left(\left[P(t+1), s', g\right]^T, a\right) + \left(1-p_{s(t)}\right)\min_{a\in\mathcal{A}}Q^*\left(\left[P(t+1), \tilde{s}(t+1), 0\right]^T, a\right)\right],$$

where $p_{s(t)}=\sum_{s'=-W}^{0}\sum_{g=1}^{g_{max}} p_{s(t),s',g}$ is the probability that a consumer request will be received in the next time step, and $\tilde{s}(t+1)=s(t)+1$ if $s(t)<\hat{W}$ and $\tilde{s}(t+1)=\hat{W}$ if $s(t)=\hat{W}$.

- If the smart device has already received a consumer request in the current episode, we have

$$Q^*\left(x(t), \text{on}\right)=P(t)C(n)+\gamma\tilde{U}_r\left(s(t), g(t)\right),$$

and $Q^*\left(x(t), \text{off}\right)=$

$$\alpha\tilde{p}_{s(t),g(t)}\mathbb{E}_{P(t+1)}\left[\min_{a\in\mathcal{A}}Q^*\left(\left[P(t+1), s(t)+1, g(t)\right]^T, a\right)\right] + (1-\tilde{p}_{s(t),g(t)})\gamma\tilde{U}_c\left(s(t), g(t)\right)$$

if $s(t)<W$ and $Q^*\left(x(t), \text{off}\right)=\gamma\tilde{U}_c\left(W, g(t)\right)$ if $s(t)=W$.

From the above Bellman equations, it is obvious that Q^* can be exactly computed based on backward induction. We observe that this computation is tractable since the cardinality of the state space in a device based MDP is usually small. Once Q^* is available, one optimal policy μ^* is

$$\mu^*(x(t))\in\operatorname*{argmin}_{a\in A}Q^*(x(t), a).$$

III. RL-EMS ALGORITHM

In Section II, the optimal DR problem of an electricity consumer is formulated as an infinite-horizon MDP of the RL-EMS, and this MDP is decomposed over devices under suitable assumptions. Furthermore, we show that if the transition model and the cost Φ of that device based MDP is known, the optimal scheduling/prediction policy can be derived based on a finite-horizon DP algorithm and this computation is tractable.

However, in practice, the transition probabilities and dis-utility functions are not known and can be different for different devices or people. To solve this problem, we introduce an reinforcement learning (RL) algorithm. Specifically, in the classical RL literature, the *environment* is consist of an an unknown MDP, and the agent (decision-maker) learns how to make decisions from consequences of actions while interacting with the environment. Notice in the optimal DR problem, the "agent" is the RL-EMS and the "environment" includes both the exogenous electricity price and the electricity consumer (see Figure 3). We assume the RL-EMS knows the state space $\mathcal{S}$, action space $\mathcal{A}$, but needs to learn the transition model and the dis-utility of the consumer based on its experience. There are various RL algorithms

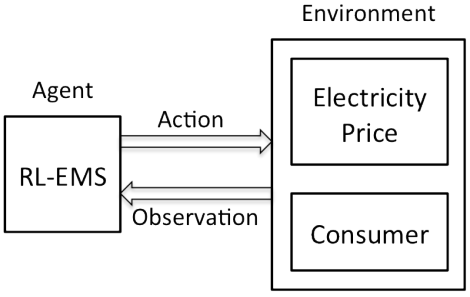


Fig. 3. Illustration of the Reinforcement Learning Model

(see [13]), and we conjecture that many of them can be implemented in the RL-EMS. In this section, we propose to implement a variant of $Q(\lambda)$ algorithm (see [15]) in the RL-EMS. The main advantage of the $Q(\lambda)$ algorithm is that it incorporates the *eligible trace* and *importance sampling* into consideration, and hence is expected to learn the optimal policy more efficiently than the classical Q-learning algorithm.

We briefly motivate our $Q(\lambda)$ algorithm. As is in the classical Q-learning, the temporal difference (TD) error at time t is

$$\begin{aligned}\delta_t &\stackrel{\text{def}}{=} \Phi\left(x(t), a(t), x(t+1)\right) \\ &+ \alpha \min_{a'} Q_t\left(x(t+1), a'\right) - Q_t\left(x(t), a(t)\right),\end{aligned}$$

where Φ is the instantaneous cost function and $Q_t \in \Re^{|\mathcal{S}||\mathcal{A}|}$ represents (vector) Q-value function estimate at time t. The update in the classical Q-learning algorithm is as follows:

$$Q_{t+1}(x(t), a(t)) = Q_t(x(t), a(t)) + \beta_t \delta_t,$$

where, $\beta_t > 0$ denotes step-size at time t. Under proper step-size conditions, Q-learning is guaranteed to converge to the optimal solution if all states are visited infinitely often (see [13]).

We can combine Q-learning with eligibility traces (see [13]). Eligibility traces are essential when data is generated in a temporally fine resolution, and they carry information about previously seen states in cause and effect. We use a version of eligibility traces that is derived in [15] (also see [16]). Algorithm 1 shows how to use Q-learning with eligibility traces. Note for the case where $\lambda = 0$ we get classical Q-learning.

Here, we have considered $e_t = \psi_t + \rho_t \alpha \lambda e_{t-1}$, where e_t is the eligibility trace vector, ψ_t is a binary vector whose only nonzero element is $(x(t), a(t))$, α is the discrete-time discount rate, λ is a pre-specified parameter in the $Q(\lambda)$ algorithm and ρ_t captures the notion of importance sampling. Notice, Q-learning is an off-policy learning algorithm, that is while the agent is following its own behavior policy it can learn about the greedy target policy (here, greedy target policy refers to the policy which is greedy with respect to the negative value of the current estimate for Q-functions). (Thus, the importance sampling ration, ρ_t, refers to the ratio between the greedy target policy and the agent's behavior policy for choosing action a_t at time t.) Specifically, let $\mu_b : \mathcal{S} \times \mathcal{A} \to [0, 1]$ be the (randomized) behavioral policy used in the $Q(\lambda)$ algorithm, we define

$$\rho_t = \begin{cases} \frac{1}{\mu_b(a(t)|x(t))} & \text{if } a(t) \in \operatorname{argmin}_{a \in \mathcal{A}} Q_t(x(t), a) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

Please refer to [15] for complete derivation of this algorithm. Notice that in our algorithm, the eligibility traces are updated according to an importance-weight scenario, which is different than Watkins's $Q(\lambda)$ algorithm (see [13]), where ρ_t is considered either 1 or 0.

Finally, we specify the behavioral policy μ_b in this Q-learning algorithm. There are many choices of μ_b, in this paper, we choose μ_b as the *ϵ-softmin* policy. That is, with probability $1 - \epsilon$, we choose $a(t) = a^* \in \operatorname{argmin}_{a \in \mathcal{A}(x(t))} Q_t\left(x(t), a\right)$ and with probability ϵ, we choose $a(t)$ according to a randomized softmin policy

$$\mu_b(a(t)|x(t)) = \frac{\exp\left[-Q_t(x(t), a(t))/\eta\right]}{\sum_{a' \in \mathcal{A}(x(t))} \exp\left[-Q_t(x(t), a')/\eta\right]},$$

where ϵ is the "exploration" probability and $\eta > 0$ is the "temperature" of this softmin policy. The $Q(\lambda)$ algorithm is described in Algorithm 1.

IV. SIMULATION RESULT

In this section, we present the simulation result of the proposed RL-EMS algorithm on an illustrative example. As is expected, the simulation result indicates that the proposed RL-EMS algorithm learns a near-optimal request scheduling/prediction policy after a finite number of episodes.

Specifically, in this numerical example, we assume the exogenous price Markov chain has $P_{max} = 4$ states, and the consumer requests have two different priorities, "high" and "normal". We set the time window $W = 4$, $\hat{W} = 5$, the discrete-time discount $\alpha = 0.995$ and the "tradeoff" $\gamma = 0.05$. Thus, there are $|\mathcal{S}| = 96$ states in this example.

To simplify exposition, we assume that this example is normalized so that both the highest electricity price and

Algorithm 1 Q(λ): Q-learning with eligibility traces

1: **Initialize** Q_0 arbitrarily, set eligibility parameter $\lambda \in [0,1]$.
2: **Repeat** for each episode:
3: **Choose** a small constant step-size $\beta > 0$ for each episode.
4: **Initialize** eligibility trace vector $e_{t-1} = 0$.
5: **Take** $a(t)$ from $x(t)$ according to μ_b (e.g. ϵ-softmin policy), and arrive at $x(t+1)$.
6: **for** each time step in an episode **do**
7: Observe sample, $(x(t), a(t), x(t+1), \Phi_t)$ at time step t, where Φ_t is the instantaneous cost.
8: $\delta_t \stackrel{\text{def}}{=} \Phi(x(t), a(t), x(t+1)) + \alpha \min_{a'} Q_t(x(t+1), a') - Q_t(x(t), a(t))$.
9: If $a(t) \in \text{argmin}_a Q_t(x(t), a)$, then $\rho_t \leftarrow \frac{1}{\mu_b(a(t)|x(t))}$; otherwise $\rho_t \leftarrow 0$.
10: $e_t = \psi_t + \rho_t \alpha \lambda e_{t-1}$, where ψ_t is a binary vector whose only nonzero element is $(x(t), a(t))$.
11: $Q_{t+1} \leftarrow Q_t + \beta \delta_t e_t$.
12: **end for**

the energy consumed by a standardized job are 1. The dis-utility functions $\tilde{U}_r$, $\tilde{U}_c$ and $\tilde{U}_s$ are illustrated in Figure 4(a), 4(b) and 4(c). Notice that these dis-utility functions satisfy the convexity and monotonicity properties discussed in Subsection II-B.

As to the transition model, we assume that if the smart device has not received a consumer request in the current episode, then under action "off", it will receive a consumer request in the next time step with probability $p_{s(t)}$. Notice that $p_{s(t)}$ is chosen to be an increasing function of $s(t)$ (see Figure 4(d)). We further notice that there are $(W+1)g_{max} = 10$ types of consumer requests (with different target times and priorities), for simplicity, we assume these 10 types of requests are equally likely.

Furthermore, if the smart device has received a consumer request in the current episode, then under action "off", the unsatisfied request will be cancelled with probability $\hat{p}_{s(t)}$. In this example, we assume the "cancellation probability" $\hat{p}$ only depends on the elapsed time $s(t)$ and is independent of the priority $g(t)$. We choose $\hat{p}_{s(t)}$ as an increasing function of $s(t)$ (see Figure 4(e)). Finally, we assume that when the smart device regenerates its state, with probability 1, the regenerated "device portion" state is $[s(t+1) = 0, g(t+1) = 0]^T$.

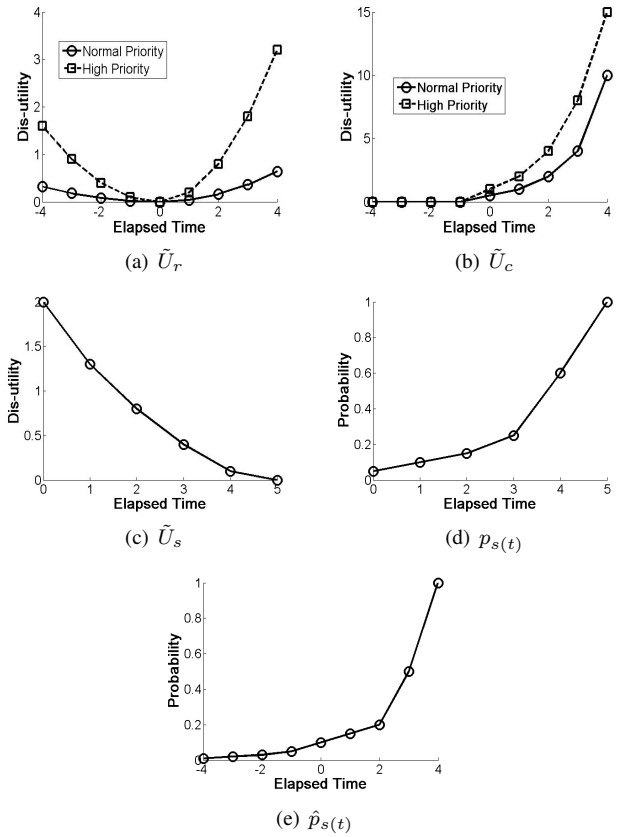
Before describing the implementation of the RL-EMS algorithm for this illustrative example, we first define the performance metric. Let Q^* denote the optimal Q-function of the finite-horizon device based MDP, and π_P^* be the stationary distribution of the exogenous price Markov chain, since the smart device regenerates its state to $[s(t+1) = 0, g(t+1) = 0]^T$, we define the "optimal performance" V^* as

$$V^* = \mathbb{E}_{P \sim \pi_P^*} \left[\min_{a \in \mathcal{A}} Q^* \left([P, 0, 0]^T, a \right) \right].$$

For any policy $\mu : \mathcal{S} \times \mathcal{A} \to [0,1]$, we define $V(\mu)$, the performance of policy μ, as

$$V(\mu) = \mathbb{E}_{P \sim \pi_P^*} \left[\min_{a \in \mathcal{A}} Q_\mu \left([P, 0, 0]^T, a \right) \right],$$

where Q_μ is the Q-function of the finite-horizon device based MDP under policy μ. Notice that Q_μ can also be computed based on backward induction, and due to the optimality of Q^*, we have $Q_\mu(x,a) \geq Q^*(x,a)$ for any $(x,a) \in \mathcal{S} \times \mathcal{A}$, which implies that $V(\mu) \geq V^*$. We define the normalized performance of policy μ as $\bar{V}(\mu) = V(\mu)/V^*$, thus $\bar{V}(\mu) \geq 1$ and obviously $\bar{V}(\mu^*) = 1$.



(a) $\tilde{U}_r$ (b) $\tilde{U}_c$ (c) $\tilde{U}_s$ (d) $p_{s(t)}$ (e) $\hat{p}_{s(t)}$

Fig. 4. Dis-utility Functions and Transition Model

With the above-defined performance metric, we first quantify the DR potential in this example. Specifically, let μ_{base} denote the default policy without DR (i.e. the device will never do a speculative job and all the requested jobs will be done at their target times). We use this default policy as the baseline in this simulation example. Notice $\bar{V}(\mu_{base}) = 2.2754$ in this example (see Figure 5); in other words, DR has the potential to reduce the consumer's cost by 56% in each episode.

Furthermore, for each function $Q : \mathcal{S} \times \mathcal{A} \to \Re$, we use μ_Q to denote a policy greedy to Q. [2] We define the normalized performance of Q as $\bar{V}(\mu_Q)$.

We now describe how we implement the proposed RL-EMS algorithm (Q(λ) algorithm). We choose the eligibility parameter $\lambda = 0.6$, "exploration probability" $\epsilon = 0.05$, and

[2] If there are multiple greedy policies, choose μ_Q as an arbitrary greedy policy.

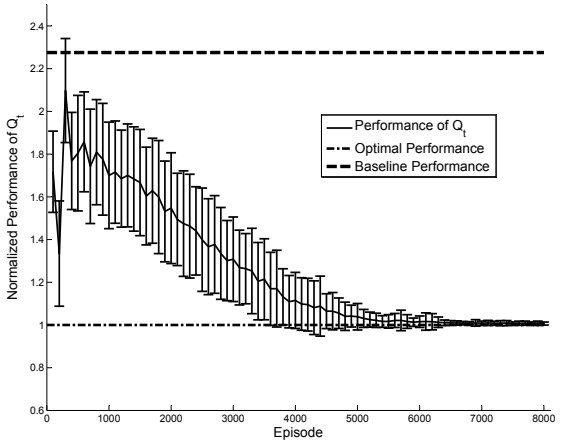


Fig. 5. Simulation Result, where the curve denotes the sample average normalized performance $\bar{V}_k^{avg}$, and the lengths of error bars denote the standard deviation.

the "temperature" of the softmin policy $\eta = 0.1$. For episode k, we choose the step-size $\beta_k = max\left\{\frac{1000}{1000+k}, 0.05\right\}$. We initialize the RL-EMS algorithm by setting $Q_0 = 0$.

The simulation result is summarized in Figure 5. Specifically, we run the proposed RL-EMS algorithm for $8,000$ episodes, and repeat the simulation for 100 times. Let $Q_{k,i}$ denote the Q-function learned after episode k in the ith simulation, we define $\bar{V}_k^{(avg)}$, the sample average normalized performance after episode k as

$$\bar{V}_k^{(avg)} = \frac{1}{100}\sum_{i=1}^{100} \bar{V}\left(\mu_{Q_{k,i}}\right),$$

and the associated standard deviation is

$$std\left(\bar{V}_k\right) = \sqrt{\frac{1}{99}\sum_{i=1}^{100}\left[\bar{V}\left(\mu_{Q_{k,i}}\right) - \bar{V}_k^{(avg)}\right]^2}$$

We plot $\bar{V}_k^{(avg)}$ and $std\left(\bar{V}_k\right)$ against the number of episodes in Figure 5.

From Figure 5, we see that in this illustrative example, the $\bar{V}_k^{(avg)}$, the sample average normalized performance of the RL-EMS algorithm after episode k outperforms the baseline $\bar{V}(\mu_{base})$ with $k < 100$, which suggests that in expectation the RL-EMS algorithm will outperform the baseline very quickly. We further notice that for $k \geq 7000$, $\bar{V}_k^{(avg)}$ is very close to 1 (the optimal performance) and the standard deviation approaches to 0, which suggests that if the proposed RL-EMS algorithm has been run for a sufficiently long time, it will achieve a near-optimal performance almost surely.

V. Conclusion

We present a reinforcement learning approach to DR for residential and small commercial buildings. The approach reduces average energy costs by shifting the time of operation of energy consuming devices either by delaying their operation or by anticipating their future use and operating them at an optimal earlier time (e.g. HVAC). The algorithm selects operating times that balance consumer dissatisfaction with energy costs and learns consumer choices and preferences, but without prior knowledge of the distribution of energy prices or consumer utility functions. The algorithm uses Q-learning with eligibility traces to learn consumer choices and time preferences. The simulation result indicates that this algorithm reduces the consumer's cost by 56% in an illustrative example. Future work includes investigating other RL algorithms with improved sample efficiency.

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] D. ONeill, M. Levorato, A. J. Goldsmith, and U. Mitra, "Residential demand response using reinforcement learning," in *IEEE SmartGridComm*, Gaithersburg, Maryland, USA, OCT 2010.

[12] 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.

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

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

[15] H. R. Maei and R. S. Sutton, "GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces," in *Proceedings of the Third Conference on Artificial General Intelligence*. Atlantis Press, 2010, pp. 91–96.

[16] H. R. Maei, "Gradient temporal-difference learning algorithms," Ph.D. dissertation, University of Alberta, 2011.