52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Index Policies for Demand Response Under Uncertainty

Joshua A. Taylor and Johanna L. Mathieu

Abstract— **Uncertainty is an intrinsic aspect of demand response because electrical loads are subject to many random factors and their capabilities are often not directly measurable until they have been deployed. Demand response algorithms must therefore balance utilizing well-characterized, good loads and learning about poorly characterized but potentially good loads; this is a manifestation of the classical tradeoff between exploration and exploitation.**

We address this tradeoff in a restless bandit framework, a generalization of the well-known multi-armed bandit problem. The formulation yields index policies, in which loads are ranked by a scalar index and those with the highest are deployed. The policy is particularly appropriate for demand response because the indices have explicit analytical expressions that may be evaluated separately for each load, making them both simple and scalable.

We numerically evaluate the performance of the index policy, and discuss implications of the policies in demand response.

I. INTRODUCTION

Aggregations of flexible electric loads can enhance power system efficiency and reliability through demand response programs, wherein loads are financially rewarded for adjusting their power consumption in response to signals from system operators, utilities, or aggregators [1] (henceforth "load managers"). A fundamental challenge to demand response is uncertainty; due to communication constraints and inputs from complex sources such as weather and human behavior, it is difficult to know the capability of a resource until it has been deployed [2]. Should the system operator then focus on loads with well-known, good characteristics, or invest in learning about lesser known loads? This is a manifestation of the tradeoff between exploration and exploitation.

When engaging large number of electric loads in demand response programs it may not be economically feasible to equip each load with a high fidelity two-way real-time communication system. On the other hand, it is important for load managers to have accurate information about individual loads so that they can deploy them in ways that maximize their response subject to local constraints. Recent research has focused on ways to minimize communications infrastructure and real-time data transfer by using state estimation techniques [3], [4], [5]. In these approaches, state estimation and control are solved as separate problems.

We propose a new method for deploying demand response resources given limited communication between the loads and system operator. We pose this scenario as a restless bandit problem [6], a popular generalization of the multi-armed bandit problem [7]. The multi-armed bandit problem refers to a scenario in which a decision maker sequentially chooses single resources from a collection, and receives a random reward and information about the selected resource. The restless bandit allows the resources to dynamically evolve whether or not they are selected, and for multiple resources to be selected at each time stage. The general restless bandit problem is PSPACE-hard [8], but admits a polynomial time relaxation if a certain *indexability* property is satisfied. Mechanistic approaches to the relaxation are given in [9], [10]. The technical content of our work rather resembles the direct formulations of [11] for vehicle routing and [12] for multichannel access; however, to our knowledge, demand response has not been formulated as a restless bandit problem before. Moreover, our work differs from existing restless bandit formulations in that each resource's evolution depends on whether it is deployed or not.

The motivation for formulating demand response in this way is that multi-armed and relaxed restless bandit problems admit optimal *index policies*, in which each resource is assigned a scalar index and the highest are dexployed. We assume the system operator must pay for each deployed resource and so he wishes to limit the number deployed to limit his costs. Because indices may be individually computed for each load, index policies are both extremely simple and scalable. This makes them well-suited for demand response, in which an aggregation may contain thousands to *millions* of flexible loads.

We specifically consider load curtailment, in which loads reduce their electricity consumption when the system is operating near its capacity limits. We assume the loads can be described by two state systems, which is a valid approximation for many types of loads, including thermostatically controlled loads [13] that modulate temperature by switching between on and off states. Moreover, it is often justifiable to approximate multi-state loads as two-state systems when considering large aggregations. We model each load as a separate, partially observable Markov process, which we show satisfies the indexability condition. We then analytically derive the associated indices, which have simple, explicit forms, and outperform the greedy policy in numerical examples.

The novel contributions of our paper are summarized as follows:

- formulation of demand response as a restless bandit problem,
- theoretical analysis of the resulting model's indexability, and

J.A. Taylor is with the Department of Electrical and Computer Engineering at the University of Toronto, Canada josh.taylor@utoronto.ca

J.L. Mathieu is with the Power Systems Laboratory at ETH Zürich, Switzerland jmathieu@eeh.ee.ethz.ch

978-1-4673-5716-6/13/$31.00 ©2013 IEEE

- construction of analytical index policies, which enable highly scalable, low-communication demand response.

The rest of this paper is organized as follows. Section II details our load models and in Section III we prove indexability and then derive the index policy. Section IV provides a discussion of the results, and in Section V we discuss future research directions.

II. LOAD MODELING

A load has binary state $x_i(t) = 1$ if it is available and $x_i(t) = 0$ if it is unavailable for demand response at time t, and fixed capacity $c_i > 0$, $i \in \{1, ..., n\}$. At each time stage, we assume the load manager wishes to limit the number of loads participating in demand response because he has a fixed budget and must pay each load each time it is deployed. We assume the payment is uniform across the loads and over time. Therefore, at each time stage, he chooses $m < n$ loads, which are designated by the control vector $u(t) \in \{0, 1\}^n$. We refer to a load with $u_i(t) = 0$ or $u_i(t) = 1$ as *passive* and *active* at time t, respectively. We model the state evolution as the time-invariant Markov process

$$\begin{aligned}
p\left(x_i(t+1) = 1 | x_i(t) = 0, u_i(t) = 1\right) &= \psi_i, \\
p\left(x_i(t+1) = 1 | x_i(t) = 1, u_i(t) = 1\right) &= \gamma_i, \\
p\left(x_i(t+1) = 1 | x_i(t) = 0, u_i(t) = 0\right) &= \rho_i, \\
p\left(x_i(t+1) = 1 | x_i(t) = 1, u_i(t) = 0\right) &= \beta_i.
\end{aligned}$$

We may interpret these parameters as follows. ψ_i and γ_i are respectively the probabilities that unavailable and available loads will subsequently be able to provide demand response given that they are currently active. ρ_i and β_i are the probabilities that unavailable and available loads will subsequently be able to provide demand response given that they are currently passive. Note that in [11], $\psi_i = \rho_i$ and $\gamma_i = \beta_i$, i.e. the state evolution does not depend on the control action.

Our assumption that the resource models are time invariant makes sense for certain types of resources, for example, refrigerators. Time varying resources can also be modeled as time invariant over certain timescales, for example, heaters and air conditioners are approximately time invariant over timescales of minutes to a couple hours. We assume that each load's parameters are known to the load manager. These values can be computed locally and acquired by the load manager either 1) a-priori, when a load signs up to participate in the program, or 2) via a communication link, which may only be active after a load is dispatched, as described below.

We assume that a load's state is not directly observable until it is dispatched. This means that a load communicates its state to the system operator only directly after the system operator dispatches the load (in Section IV-B we discuss the realistic scenario in which, after dispatch, only the aggregate system state is known). We use $y_i(t)$ to denote the probability distribution of $x_i(t)$ conditioned on all previous observations and actions, which is a sufficient statistic for optimal control [14]. The goal is to maximize the expected discounted infinite horizon capacity, i.e.

$$\max_{u(t) \in \{0,1\}^n} J, \tag{1}$$

where

$$J = \mathbb{E} \sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{n} c_i u_i(t) x_i(t) = \sum_{t=0}^{\infty} \alpha^t \sum_{i=1}^{n} c_i u_i(t) y_i(t), \tag{2}$$

$0 \leq \alpha < 1$ is the discount factor, and $\sum_{i=1}^{n} u_i(t) = m$.

Define the mapping

$$\phi_i y_i = y_i \beta_i + (1 - y_i) \rho_i,$$

and denote the k times repeated application of ϕ_i to y_i by $\phi_i^k y_i$. Note that for $k \geq 0$,

$$\phi_i^k y_i = (\beta_i - \rho_i)^k y_i + \rho_i \frac{1 - (\beta_i - \rho_i)^k}{1 - (\beta_i - \rho_i)}, \tag{3}$$

and denote $\lim_{k \to \infty} \phi_i^k y_i$ by

$$\chi_i = \frac{\rho_i}{1 - (\beta_i - \rho_i)}.$$

The evolution of y_i is given by

$$y_i(t+1) = \begin{cases} \psi_i, & u_i(t) = 1, x_i(t) = 0 \\ \gamma_i, & u_i(t) = 1, x_i(t) = 1 \\ \phi_i y_i(t), & u_i(t) = 0 \end{cases}. \tag{4}$$

We will omit the time index t when there is no risk of ambiguity.

We assume that the transition probabilities satisfy $\psi_i \leq \gamma_i$, $\psi_i \leq \rho_i$, $\rho_i < \beta_i$, $\gamma_i \leq \chi_i$. The rationale for the first three is clear, because, all else being equal, an active load is less likely to be subsequently available than a passive load. Since χ_i is the steady state of an always passive load, the last assumption enforces that an unavailable, passive load must eventually become available, upon which it is more likely to remain available than an active, available load.

Remark 1 (Learning): The belief state y represents the quality of the loads. Its evolution of y encodes the tradeoff between exploration and exploitation in that the actual state of a particular load, x_i, is unknown to the load manager until it is tapped for demand response.

We comment that this setup and our subsequent approach resemble [11] and [12], the difference being that both the belief and actual state evolution depend on $u(t)$.

Remark 2 (Curtailment and transition probabilities): As the objective is to induce the maximum power reduction, which is consistent with current curtailment frameworks [15], we assume a load's transition probabilities are independent of the actions of other deployed loads. Our formulation would be less appropriate to energy balancing, in which a group of loads works together to counteract an energy imbalance. In this case, the amount counteracted by a single load depends on the energy imbalance size and the other loads' capabilities. Moreover, optimally matching loads to an energy imbalance is difficult even as a one time task because it entails solving an NP-hard knapsack problem [16], [17], rendering it unlikely that a computationally tractable solution to a multistage, stochastic formulation exists.

III. INDEX POLICY

We seek *index policies*, in which each load has a real-valued, scalar index, and the control is simply to select the m largest indices for activity. The myopic policy is to set each load's index to its expected current reward, $c_i y_i$. While simple, it is easy to show that this is suboptimal.

A. Restless bandit relaxation

We can do much better than the myopic policy by recognizing that the problem at hand is a restless bandit problem, for which we now summarize the standard approach; details may be found in [6], [7].

The nominal problem of solving (1) subject to (4) is PSPACE-hard [8]. We rather solve the relaxation posed in [6], which is obtained by replacing the constraint $\sum_{i=1}^n u_i(t) = m$ for all t with the time averaged constraint $\sum_{t=0}^\infty \alpha^t \sum_{i=1}^n u_i(t) = \frac{m}{1-\alpha}$. This relaxation admits an index policy, which may be adapted to a feasible policy by enforcing the original constraint, i.e. choosing the m largest indices at each time.

We construct the restless bandit relaxation's optimal index policy for the above described demand response problem in analytical form. Each load must be *indexable*, the name given to the condition from [6] guaranteeing that index policies are optimal for the restless bandit relaxation.

B. Indexability

In this section, we prove that each load is indexable, so that we may subsequently derive and apply an index policy. Since we consider a single generic load, we drop the subscript i and set $c = 1$ in this section. Consider an augmented scenario in which each load receives a subsidy θ whenever it is passive. Denote the load's value function $J(y,\theta)$, and define the passive and active value

$$\begin{aligned} J^p(y,\theta) &= \theta + \alpha J(\phi y, \theta) && (5)\\ J^a(y,\theta) &= y + \alpha\left(yJ(\gamma,\theta) + (1-y)J(\psi,\theta)\right). && (6)\end{aligned}$$

The dynamic program for this augmented problem is given by

$$J(y,\theta) = \max\{J^p(y,\theta), J^a(y,\theta)\}. \qquad (7)$$

Define the set $I(\theta)$ to be the values of y for which passivity is optimal, i.e. $J^p(y,\theta) \le J^a(y,\theta)$.

Definition 1: A load is indexable if as θ goes from $-\infty$ to ∞, $I(\theta)$ monotonically increases from $\emptyset$ to the entire state space, in this case the interval $[0,1]$.

Lemma 1: $J(y,\theta)$ is a nondecreasing function of y.

Proof: We adapt the argument from [11]. Define the finite horizon value function

$$\begin{aligned} J_{k-1}(y,\theta) &= \max\{\theta + \alpha J_k(\phi y,\theta),\\ &\qquad y + \alpha\left(yJ_k(\gamma,\theta) + (1-y)J_k(\psi,\theta)\right)\},\\ J_K(y,\theta) &= 0. \end{aligned}$$

By rewriting the second argument of the maximization as $y + \alpha\left(y\left(J_k(\gamma,\theta) - J_k(\psi,\theta)\right) + J_k(\psi,\theta)\right)$, it is also evident that each $J_k(y,\theta)$ is nondecreasing in y due to the assumption that $\psi \le \gamma$; we omit the induction argument because it is standard. We can also see that because both arguments are affine in y, each $J_k(y,\theta)$ is piecewise linear and convex. Because $\alpha < 1$ and $y \in [0,1]$, the dynamic programming operator is a contraction, and we may define the infinite horizon value function as the uniformly convergent limit, $J(y,\theta) = \lim_{K\to\infty} J_k(y,\theta)$ [18], [14]. Therefore, $J(y,\theta)$ is continuous, convex, and nondecreasing in y. ■

Lemma 2: $I(\theta)$ is of the form $[0,\xi(\theta)]$.

Proof: For the sake of contradiction, suppose that $y_1 < y_2 \le 1$ and that activity is optimal at y_1 and passivity is optimal at y_2. Let τ be the smallest positive integer such that activity is optimal at $\phi^\tau y_2$. Define the affine function

$$\hat{J}(y,\theta) = \frac{1-\alpha^\tau}{1-\alpha}\theta + \alpha^\tau J^a(\phi^\tau y, \theta).$$

Note that $\tau \ge 1$ because y_2 is passive and that the case in which activity never becomes optimal is captured by $\tau = \infty$. From (7) and the continuity of $J(y,\theta)$,

$$J(y_2,\theta) = J^p(y_2,\theta) = \hat{J}(y_2,\theta).$$

Since $J(\gamma,\theta) \ge J(\psi,\theta)$ by Lemma 1 and $\alpha(\beta-\rho) < 1$,

$$\begin{aligned} \frac{\partial \hat{J}(y,\theta)}{\partial y} &= \alpha^\tau(\beta-\rho)^\tau(1 + \alpha(J(\gamma,\theta) - J(\psi,\theta)))\\ &< 1 + \alpha(J(\gamma,\theta) - J(\psi,\theta))\\ &= \frac{\partial J^a(y,\theta)}{\partial y}. \end{aligned}$$

Since $\hat{J}(y,\theta)$ and $J^a(y,\theta)$ are affine and (by construction) $\hat{J}(y_1,\theta) \le J^a(y_1,\theta)$, we have that

$$\begin{aligned} J^p(y_2,\theta) &= \hat{J}(y_1,\theta) + \frac{\partial \hat{J}(y,\theta)}{\partial y}(y_2 - y_1)\\ &< J^a(y_1,\theta) + \frac{\partial J^a(y,\theta)}{\partial y}(y_2 - y_1)\\ &= J^a(y_2,\theta). \end{aligned}$$

This contradicts the assumption that passivity is optimal at y_2, proving that a passive state cannot be larger than an active state. This implies that the passivity is optimal on an interval containing zero; denoting its upper endpoint $\xi(\theta)$ establishes the claim. ■

These results imply that for any $y \in [0,1]$, there is a $\theta(y)$ for which the load is indifferent between passivity and activity, so that

$$J(y,\theta(y)) = J^p(y,\theta(y)) = J^a(y,\theta(y)). \qquad (8)$$

We now use (8) to explicitly determine the function $\theta(y)$, which will comprise the index in the restless bandit index policy. We again separately treat the case that $y < \chi$ and $y \ge \chi$.

1) $y < \chi$: Define

$$\begin{aligned}\tau_1 &= \min\{\tau \geq 0 \mid y \leq \phi^\tau \psi\} \\ &= \left\lceil \log_{\beta-\rho} \frac{\rho - y(1-(\beta-\rho))}{\rho - \psi(1-(\beta-\rho))} \right\rceil, \\ \tau_2 &= \min\{\tau \geq 0 \mid y \leq \phi^\tau \gamma\} \\ &= \left\lceil \log_{\beta-\rho} \frac{\rho - y(1-(\beta-\rho))}{\rho - \gamma(1-(\beta-\rho))} \right\rceil.\end{aligned}$$

We first evaluate the $J(\psi)$ and $J(\gamma)$. Observe that

$$\begin{aligned}J(\psi) &= \theta(y) + \alpha J(\phi\psi) \\ &= \frac{1-\alpha^{\tau_1}}{1-\alpha}\theta(y) + \alpha^{\tau_1}\left(\phi^{\tau_1}\psi + \alpha(\phi^{\tau_1}\psi J(\gamma)\right. \\ &\quad \left. +(1-\phi^{\tau_1}\psi)J(\psi))\right).\end{aligned}$$

We write $J(\gamma)$ similarly to obtain the following implicit equation:

$$\begin{aligned}\left[\begin{array}{c} J(\psi) \\ J(\gamma) \end{array}\right] &= \frac{\theta(y)}{1-\alpha}\left[\begin{array}{c} 1-\alpha^{\tau_1} \\ 1-\alpha^{\tau_2} \end{array}\right] \\ &\quad + \left[\begin{array}{cc} \alpha^{\tau_1} & 0 \\ 0 & \alpha^{\tau_2} \end{array}\right]\left(\left[\begin{array}{c} \phi^{\tau_1}\psi \\ \phi^{\tau_2}\gamma \end{array}\right]\right. \\ &\quad \left. +\alpha\left[\begin{array}{cc} 1-\phi^{\tau_1}\psi & \phi^{\tau_1}\psi \\ 1-\phi^{\tau_2}\gamma & \phi^{\tau_2}\gamma \end{array}\right]\left[\begin{array}{c} J(\psi) \\ J(\gamma) \end{array}\right]\right).\end{aligned}$$

Define

$$\begin{aligned}\Delta &= 1 - \alpha\left(\alpha^{\tau_1} + \alpha^{\tau_2}\left(1-\alpha^{\tau_1+1}\right)\phi^{\tau_2}\gamma\right. \\ &\quad \left. -\alpha^{\tau_1}\left(1-\alpha^{\tau_2+1}\right)\phi^{\tau_1}\psi\right), \\ \Gamma_1 &= 1-\alpha^{\tau_1}-\alpha^{\tau_2+1}\left(1-\alpha^{\tau_1}\right)\phi^{\tau_2}\gamma \\ &\quad +\alpha^{\tau_1+1}\left(1-\alpha^{\tau_2}\right)\phi^{\tau_1}\psi, \\ \Gamma_2 &= 1-\alpha^{\tau_2}-\alpha^{\tau_2+1}\left(\left(1-\alpha^{\tau_1}\right)\phi^{\tau_2}\gamma - 1\right) \\ &\quad +\alpha^{\tau_1+1}\left(\left(1-\alpha^{\tau_2}\right)\phi^{\tau_1}\psi - 1\right), \\ \Omega_1 &= \alpha^{\tau_1}\phi^{\tau_1}\psi, \\ \Omega_2 &= \alpha^{\tau_2}\left(\phi^{\tau_2}\gamma + \alpha^{\tau_1+1}\left(\phi^{\tau_1}\psi - \phi^{\tau_2}\gamma\right)\right).\end{aligned}$$

The solution to the above equation is then given by

$$\left[\begin{array}{c} J(\psi) \\ J(\gamma) \end{array}\right] = \frac{1}{\Delta}\left(\frac{\theta(y)}{(1-\alpha)}\left[\begin{array}{c} \Gamma_1 \\ \Gamma_2 \end{array}\right] + \left[\begin{array}{c} \Omega_1 \\ \Omega_2 \end{array}\right]\right).$$

We now apply (8) to solve for $\theta(y)$. We have

$$\begin{aligned}J(y) &= y + \alpha\left(yJ(\gamma) + (1-y)J(\psi)\right) \\ &= \theta(y) + \alpha J^a(\phi y) \\ &= \theta(y) + \alpha\left(\phi y + \alpha\left(\phi y J(\gamma) + (1-\phi y)J(\psi)\right)\right)\end{aligned}$$

from which we obtain

$$\theta^1(y) = \frac{(y-\alpha\phi y)\left(1-\alpha^{\tau_1+1}\right) + (1-\alpha)\alpha^{\tau_1+1}\phi^{\tau_1}\psi}{\begin{array}{c}(y-\alpha\phi y)\left(\alpha^{\tau_2+1}-\alpha^{\tau_1+1}\right) \\ +(1-\alpha)\left(1+\alpha^{\tau_1+1}\phi^{\tau_1}\psi - \alpha^{\tau_2+1}\phi^{\tau_2}\gamma\right)\end{array}}.$$

2) $y \geq \chi$: In this case,

$$\begin{aligned}J(y,\theta(y)) &= \theta + \alpha J(\phi y, \theta(y)) \\ &= \frac{\theta(y)}{1-\alpha}.\end{aligned}$$

Recall that since the passive set is of the form $I(\theta(y)) = [0, y]$, passivity is optimal at any state less than y. By our initial assumptions in Section II, γ and ψ are less than χ and hence y, so that

$$\begin{aligned}J(y,\theta(y)) &= y + \alpha(yJ(\gamma,\theta(y)) \\ &\quad +(1-y)J(\psi,\theta(y))) \\ &= y + \frac{\alpha\theta(y)}{1-\alpha}.\end{aligned}$$

Combining these equations, we have $\theta(y) = y$.

The load's index is then

$$\theta(y) = \begin{cases} \theta^1(y) & y < \chi \\ y & y \geq \chi \end{cases}.$$

Theorem 1: The load is indexable.

Proof: Since $\theta(-\infty) = 0$ and $\theta(\infty) = 1$, it is sufficient to show that $\theta(y)$ is non-decreasing. For $y \geq \chi$, $\theta(y) = y$ and so it is increasing. For $y < \chi$, $\theta(y) = \theta^1(y)$, which is composed of increasing segments corresponding to fixed values of τ_1 and τ_2, which increment when $y = \phi^\tau\psi$ or $y = \phi^\tau\gamma$ for some integer τ. It is straightforward to verify that $\theta^1(y)$ is continuous at these values of y, and is hence monotonically increasing. Since $\theta^1(\chi) = \chi$, $\theta(y)$ is continuous and monotonically increasing as well. ■

C. Policy

We now restore load indices and allow for any $c_i > 0$. The policy at a particular time step is as follows:

1) For each i, set

$$\theta_i(y_i) = \begin{cases} c_i\theta_i^1(y_i) & y_i < \chi_i \\ c_i y_i & y_i \geq \chi_i \end{cases}.$$

2) Dispatch the m loads with the largest θ_i's.
3) Observe the actual state x_i of the active loads, and update the belief state y according to (4).

The greedy policy and $\theta_i(y_i)$ are illustrated in Fig. 1 for a range of parameters. In Fig. 2, we compare total discounted actual and expected capacities resulting from greedy and restless bandit index policies on an example with 1000 randomly generated loads, 200 of which are deployed at each time stage. We see that at time stage 50, the restless bandit index policy outperforms the greedy policy by about 30 capacity units in expectation and 50 in the actual trajectories. In other experiments, we similarly observed 5% to 10% improvements with the restless bandit policy.

Remark 3 (Identical loads): Because each $\theta_i(y)$ is monotonically increasing, this policy and the myopic policy are identical if all loads have identical transition probabilities, as observed in [12].

IV. DISCUSSION

We now discuss several implementation aspects of our development.

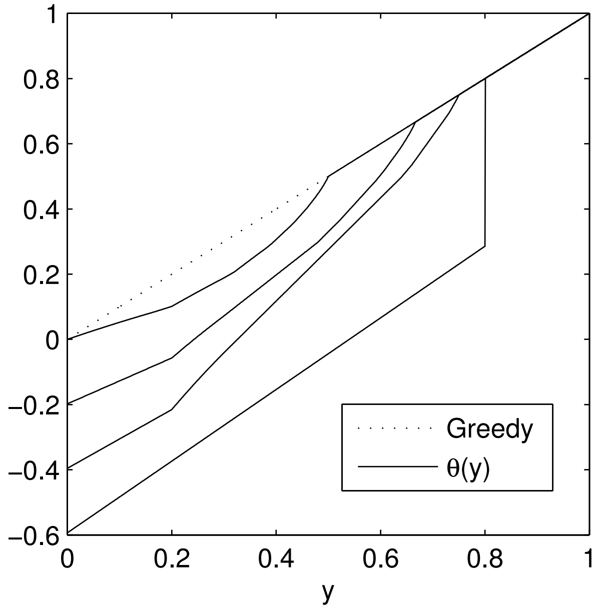

Fig. 1. The greedy policy and $\theta(y)$ for $c = 1$ $\alpha = 0.9$, $\psi = 0.2$, $\gamma = 0.3$, $\beta = 0.8$, and $\rho \in \{0.2, .04, 0.6, 0.8\}$ (the upper and lower most $\theta(y)$ curves correspond to $\rho = 0.2$ and $\rho = 0.8$, respectively).

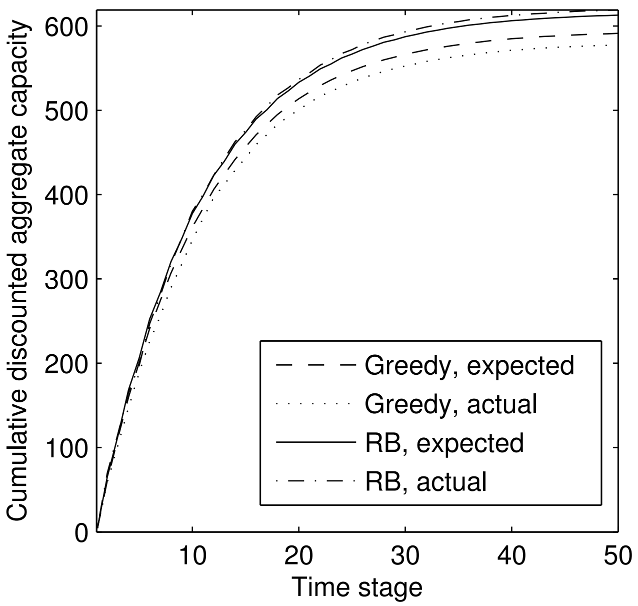

Fig. 2. The total discounted actual and expected capacities $\left(\sum_{t=0}^{T} \alpha^t \sum_{i=1}^{n} c_i u_i(t) x_i(t) \text{ and } \sum_{t=0}^{T} \alpha^t \sum_{i=1}^{n} c_i u_i(t) y_i(t)\right)$ for greedy and restless bandit index policies.

A. Alternative policies

Observe that the $\theta_i(y_i)$ do not depend on m; indeed, the restriction that exactly m loads be deployed is an assumption of the basic restless bandit formulation [6], but is not fundamental to demand response. Moreover, recall that deploying exactly m loads is a *heuristic* application of the indices, which are actually optimal for the relaxed case that m loads be dispatched on average.

These observations motivate us to suggest that $\theta_i(y_i)$ may be more generally useful as a ranking, i.e. when trying to choose which subset of loads to dispatch at each time while subject to other constraints. We note, however, that such application may compromise the validity of the Markov transition model, for reasons discussed in Remark 2.

B. Disaggregation

In this paper, we have assumed that the states of deployed loads, x_i, are observable. While this is realistic in some setups, a more likely scenario is that the aggregate effect of demand response is observed, which is of the form

$$\nu + \sum_i c_i x_i u_i,$$

where ν is a random observation error. This could be addressed by directly incorporating the aggregate description into the belief state transition model; however, this would complicate the current development by coupling the evolution of y_i with that of other loads, potentially violating the restless bandit assumptions.

Alternatively, estimating x could be decoupled from the learning and control problem, e.g. via a filter, and the resulting estimated values could then be fed into the policy in Section III-C.

V. CONCLUDING REMARKS

Resource uncertainty is a fundamental challenge in demand response. Because learning the characteristics of loads is coupled to their deployment, demand response strategies must balance exploring lesser known loads with utilizing well-known, good loads. We have addressed this tradeoff within the restless bandit framework, and obtained index policies. Because the indices are analytical expressions that may be computed separately for each load, they are extremely simple and scalable, making them well-suited for demand response. In a numerical example, the index policy was shown to outperform the greedy policy.

There are many potential venues for future research, motivated by the crudeness of the model and the variety of alternate demand response scenarios. In particular, it would be of interest to develop computational rather than analytical policies for loads with more complex but indexable Markovian evolutions.

REFERENCES

[1] DOE, "Benefits of demand response in electricity markets and recommendations for achieving them," Department of Energy Report to the US Congress, Tech. Rep., 2006. [Online]. Available: http://www.oe.energy.gov/DocumentsandMedia/congress_1252d.pdf

[2] J. Mathieu, D. Callaway, and S. Kiliccote, "Variability in automated responses of commercial buildings and industrial facilities to dynamic electricity prices," *Energy and Buildings*, vol. 43, pp. 3322–3330, 2011.

[3] J. Mathieu, S. Koch, and D. Callaway, "State estimation and control of electric loads to manage real-time energy imbalance," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 430–440, 2013.

[4] T. Borsche, F. Oldewurtel, and G. Andersson, “Minimizing communication cost for demand response using state estimation,” in *Proceedings of PowerTech Conference*, Grenoble, France, 2013.

[5] E. Kara, Z. Kolter, M. Berges, B. Krogh, G. Hug, and T. Yuksel, “A moving window state estimator in the control of thermostatically controlled loads for demand response,” in *Proceedings of SmartGrid-Comm*, Vancouver, BC, 2013.

[6] P. Whittle, “Restless bandits: activity allocation in a changing world,” *Journal of Applied Probability*, vol. 25A, pp. 287–298, 1988.

[7] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices*. Wiley, 2011.

[8] C. H. Papadimitriou and J. N. Tsitsiklis, “The complexity of optimal queuing network control,” *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, 1999.

[9] D. Bertsimas and J. Niño Mora, “Restless bandits, linear programming relaxations, and a primal-dual index heuristic,” *Operations Research*, vol. 48, no. 1, pp. 80–90, Jan./Feb. 2000.

[10] J. Niño Mora, “Restless bandits, partial conservation laws and indexability,” *Advances in Applied Probability*, vol. 33, no. 1, pp. 76–98, 2001.

[11] J. Le Ny, M. Dahleh, and E. Feron, “Multi-UAV dynamic routing with partial observations using restless bandit allocation indices,” in *American Control Conference, 2008*, Jun. 2008, pp. 4220 –4225.

[12] K. Liu and Q. Zhao, “Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access,” *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5547 –5567, Nov. 2010.

[13] D. Callaway, “Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy,” *Energy Conversion and Management*, vol. 50, pp. 1389–1400, 2009.

[14] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Two Volume Set*. Athena Scientific, 2005.

[15] PG&E, “Peak day pricing,” Pacific Gas and Electric Company, 2012. [Online]. Available: http://www.pge.com/pdp/

[16] V. V. Vazirani, *Approximation Algorithms*. Springer, March 2004.

[17] G. Xiong, C. Chen, S. Kishore, and A. Yener, “Smart (in-home) power scheduling for demand response on the smart grid,” in *Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES*, Jan. 2011, pp. 1 –7.

[18] R. D. Smallwood and E. J. Sondik, “The optimal control of partially observable markov processes over a finite horizon,” *Operations Research*, vol. 21, no. 5, pp. 1071–1088, Sept./Oct. 1973.