A Near-Optimal Maintenance Policy for Automated DR Devices
Carlos Abad and Garud Iyengar

Abstract—Demand-side participation is now widely recognized as being extremely critical for satisfying the growing electricity demand in the U.S. The primary mechanism for demand management in the U.S. is demand response (DR) programs that attempt to reduce or shift demand by giving incentives to participating customers via price discounts or rebate payments. Utilities that offer DR programs rely on automated DR devices (ADRs) to automate the response to DR signals. The ADRs are faulty, but the working state of the ADR is not directly observable; one can, however, attempt to infer it from the power consumption during DR events. The utility loses revenue when a malfunctioning ADR does not respond to a DR signal; however, sending a maintenance crew to check and reset the ADR also incurs costs.

In this paper, we show that the problem of maintaining a pool of ADRs using a limited number of maintenance crews can be formulated as a restless bandit problem, and that one can compute a near-optimal policy for this problem using Whittle indices. We show that the Whittle indices can be efficiently computed using a variational Bayes procedure even when the load-shed magnitude is noisy, and when there is a random mismatch between the clocks at the utility and at the meter. The results of our numerical experiments suggest that the Whittle-index-based approximate policy is within 4% of the optimal solution for all reasonably low values of the signal-to-noise ratio in the meter readings.

Index Terms—Automated demand response (ADR), Bayes procedures, dynamic programming.

NOMENCLATURE
Indices

| Symbol | Description |
|--------|-------------|
| $t$    | Index of demand response (DR) events. |
| $k$    | Index of discrete points in belief space $[0, 1]$. |
| $i$    | Index of meter readings during the DR event. |
| $j$    | Index of samples of meter reading vectors. |

Constants and Parameters

| Symbol | Description |
|--------|-------------|
| $\lambda$ | Expected dollar savings for utility when automated DR device (ADR) is working. |
| $c$    | Cost incurred by utility for sending a repair crew. |
| $\theta$ | Customer compensation for participating in a DR event. |

$p$ Prior probability of ADR failure.
$S$ ADR state space.
$\gamma_0$ Nonoperational ADR state.
$\gamma_1$ Operational ADR state.
$A$ Set of actions available to utility before a DR event.
$\alpha_0$ Do-nothing action.
$\alpha_1$ Send-crew-to-reset-ADR action.
$X$ Set of possible meter reading vectors during the DR event.
$\beta$ Discount factor between DR events.
$n$ Discretization size of belief space $[0, 1]$.
$D$ Number of ADRs managed by utility.
$M$ Number of repair crews overseeing ADRs.
$m$ Number of meter readings during the DR event.
$N$ Number of samples of meter reading vectors.
$\nu$ Vector of estimated normal power consumption.
$\sigma$ Standard deviation of load estimation residuals.
$\eta_0$ Load-shed precision (inverse of variance).
$d$ Maximum absolute meter-utility clock mismatch.
$\epsilon$ Load estimation residual.
$\delta$ Actual mismatch between meter and utility clocks.
$\omega$ Low-discrepancy sample in unit hypercube.

Variables

| Symbol | Description |
|--------|-------------|
| $s$    | State of ADR before the DR event. |
| $a$    | Action taken by utility before the DR event. |
| $x$    | Vector of meter readings during the DR event. |
| $b$    | Belief probability that ADR is operational. |
| $\tilde{x}$ | Up-to-date meter reading observation probability. |
| $r$    | Actual load-shed during a DR event. |
| $\delta$ | Actual mismatch between meter and utility clocks. |
| $\omega$ | Low-discrepancy sample in unit hypercube. |

Functions

| Symbol | Description |
|--------|-------------|
| $R_{ax}(a)$ | Utility profit when taking action $a$ in ADR state $s$. |
| $P_{\gamma_0}(a)$ | ADR transition probability. |
| $Q_{ax}(x)$ | Meter reading observation conditional probability. |
| $r_a(b)$ | Expected profit under action $a$ in belief state $b$. |
| $\Gamma_{\alpha}(b)$ | Belief probability transition map. |
| $W_{ab}(x)$ | Up-to-date meter reading observation probability. |
| $V(b)$ | Partially observable Markov decision process value function. |
| $\Phi(\cdot)$ | Standard normal cumulative distribution function. |
| $\varphi(\cdot)$ | Standard normal density function. |

Manuscript received October 18, 2014; revised February 28, 2015 and June 6, 2015; accepted July 23, 2015. Date of publication August 16, 2015; date of current version April 19, 2016. This work was supported in part by the Department of Energy under Grant DE-AR000235, in part by the Office of Naval Research under Grant N000140310514, and in part by the National Science Foundation under Grant DMS-1016571. Paper no. TSG-01039-2014.

The authors are with the Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027 USA (e-mail: ca2446@columbia.edu; garud@ieor.columbia.edu). Digital Object Identifier 10.1109/TSG.2015.2465834
I. INTRODUCTION

DEMAND response (DR) refers to a set of activities where end-use customers change or shift their normal electricity consumption patterns in response to changes in the price of electricity or other incentive payments [1]. DR has been extensively deployed over the past several years to improve electric grid reliability and market efficiency [2], and it is expected to play an even more prominent role with the planned integration of intermittent renewable generation. There are three levels of DR automation: 1) manual DR, where each equipment is manually turned off; 2) semiautomated DR, where an individual triggers a preprogrammed DR strategy via a centralized control system; and 3) fully automated DR, where the DR strategy is initiated by an ADR on receipt of an external communications signal [3]. The utilities offering DR programs prefer to install ADRs since the fully automated system reduces the operating costs of DR programs by increasing DR resource reliability and reducing the amount of effort required from end-use customers [4].

ADRs are faulty [5], [6]. According to some estimates, failure to identify nonfunctioning ADRs can reduce the effectiveness of DR programs by approximately 20%–30% and lead to lost revenues of the order of $1.7 million for a utility with one million customers and 10% participation rate [7]. Consequently, identifying malfunctioning ADRs is of immense economic value for the utilities. In Fig. 1, we schematically display a utility and three ADRs installed at customer locations, where the lightly shaded location denotes a malfunctioning ADR. The current practice is to periodically inspect all ADRs and, if needed, repair them. This is represented in Fig. 1(a) by edges going into all three locations, regardless of the ADR state. Note that two out of three visits could have been avoided if the ADR states were known at the utility. Although newer ADRs are equipped with a two-way communication that allows the utility to remotely observe the ADR state, the vast majority of the deployed ADRs use one-way communication technology and, therefore, the ADR state is not directly observable. We propose a method that is able to infer the ADR state from meter readings, and use the estimates to optimally schedule the ADR maintenance. We illustrate this in Fig. 1(b) by adding feedback edges to the utility. In the ideal situation, the proposed method would be able to accurately infer the ADR states and only schedule visits to malfunctioning ADRs, as represented in Fig. 1(b) by a single edge going to the location with the malfunctioning ADR. We show that our proposed method clearly outperforms any regular maintenance routines such as the reliability-centered maintenance, and deterioration failure models, see [10] for details. These methodologies are typically used for managing traditional electric assets such as generators and transformers. To the best of our knowledge, there is no previous work on probabilistic models for the maintenance of a single ADR, let alone the maintenance scheduling of multiple ADRs being overseen by a small number of repair teams.

We formulate the problem of maintaining a pool of ADRs using a limited number of maintenance crews as a restless bandit (RB) problem [11]. The RB problem is a generalization of the multiarmed bandit (MAB) problem [12]. In the MAB problem, the decision-maker chooses a set of “bandits” to activate based on the current state information, and the state of the chosen bandits evolves according to a known distribution; however, the states of all the inactive bandits remain fixed. Gittins [12] constructed a set of index functions that map the state of a bandit to a real number, and showed that the optimal solution for the MAB problem is to select the bandits in the order determined by the index function. In the RB setting, the states of the inactive bandits also evolve. In the ADR maintenance setting, an active bandit corresponds to an ADR that is being inspected and possibly repaired, and an inactive bandit corresponds to an ADR that is not being inspected; clearly, the state of inactive ADRs continue to evolve according to its failure distribution. Computing optimal policies for the RB problem is hard [13], but a generalization of Gittins’ indices can be used for constructing approximately optimal and very efficiently implementable policies for a number of applications. Whittle [11] established conditions under which one can define generalized Gittins’ indices for the RB problem. We will refer to such RB problems as Whittle indexable. Glazebrook et al. [14] have established

The utility decides whether to send a crew to reset the device or do nothing. This decision would be simple if the true ADR state was directly observable without any errors. In most currently deployed ADRs, the state is not observable; it can only be inferred from the customer’s noisy electricity consumption. Thus, the ADR maintenance problem can be modeled as a Markov decision process (MDP) with partially observable states (POMDP). The POMDP framework incorporates the uncertainty associated with any estimation process into a “belief” probability that the ADR is functioning, and provides a methodology for updating this belief as more information becomes available, i.e., more meter readings are recorded during new DR events. POMDPs are, in general, very hard optimization problems [8], [9]. However, it is often possible to compute the optimal policy for POMDPs with small state, action and observation spaces, or additional structure. We show that the optimal policy in our problem is a single threshold policy where it is optimal to send a maintenance crew whenever the belief probability drops below the threshold value. We show that the optimal threshold can be approximated to any degree of accuracy by solving a single linear program (LP).

Our model for a single ADR maintenance problem falls in the class of random failure models. Currently, existing alternatives to our approach are empirical predictive maintenance routines such as the reliability-centered maintenance, and deterioration failure models, see [10] for details. These methodologies are typically used for managing traditional electric assets such as generators and transformers. To the best of our knowledge, there is no previous work on probabilistic models for the maintenance of a single ADR, let alone the maintenance scheduling of multiple ADRs being overseen by a small number of repair teams.
that machine maintenance models that are either monotone or breakdown/deterioration models are Whittle indexable. The ADR repair scheduling problem is neither a monotone nor a breakdown/deterioration model; therefore, the results in [14] do not extend to this model. We establish that the ADR repair scheduling problem is Whittle indexable, and show that the Whittle indices can be computed by solving a sequence of single ADR maintenance problems.

Finally, we conduct an extensive numerical study where we explore several additional practical issues such as the impact on performance when the utility and meter clock are not synchronous, and the impact of uncertain DR load shed. Our proposed variational Bayes procedure to handle these issues is of independent interest. The results of our numerical experiments suggest that, for reasonable values of the signal-to-noise ratio (SNR) in the meter readings, the Whittle-index policy is within 4% of the optimal policy.

II. ADR MAINTENANCE PROBLEM

In this section, we investigate the maintenance problem for a single ADR. The solution to this problem will be used to solve a relaxation of the multiple ADRs repair scheduling problem in Section III.

For \( t \geq 1 \), let \( s_t \in S \) denote the ADR state just prior to the \( t \)th DR event, let \( a_t \in A \) denote the action taken just prior to the \( t \)th DR event, and let \( x_t \) denote the vector of meter readings recorded during the \( t \)th DR event. During the DR event \( t \), the utility receives a profit \( R_{a_t,s_t} \). The profit matrix \( R \) is given by

\[
R = [R_{a_t}]_{a_t \in A, s_t \in S} = \begin{bmatrix}
\gamma_0 & \gamma_1 \\
-\theta & \lambda - \theta - c \\
\lambda - \theta & \lambda - \theta - c
\end{bmatrix}
\]

(1)

For ease of notation, and without loss of generality, we will assume that \( \theta = 0 \). We assume that an ADR that was functioning during the \( t \)th event, i.e., \( s_t = \gamma_1 \), fails before the \((t+1)\)th DR event, i.e., \( s_{t+1} = \gamma_0 \), with probability \( p \). Thus, under the do-nothing action \( \alpha_0 \), the state transition matrix

\[
P(\alpha_0) = [P_{s_{t+1}}(\alpha_0)]_{s_t \in S} = \begin{bmatrix}
\gamma_0 & \gamma_1 \\
1 & 0 \\
p & 1 - p
\end{bmatrix}
\]

(2)

Since the action \( \alpha_1 \) resets the ADR state to \( \gamma_1 \) just prior to the \( t \)th DR event, it follows that the state transition matrix:

\[
P(\alpha_1) = [P_{s_{t+1}}(\alpha_1)]_{s_t \in S} = \begin{bmatrix}
\gamma_0 & \gamma_1 \\
0 & 1 - p
\end{bmatrix}
\]

(3)

We assume that the true ADR state can only be determined by sending a maintenance crew to inspect it. However, the utility can use the the vector \( x_t \) of metered power consumption to infer the state of the ADR. Let

\[
Q_{\alpha_t}(x_t) = \mathbb{P}(x_t = x_t | a_t = a_t, s_t = s_t)
\]

(4)

denote the conditional probability of observing the vector of meter readings \( x_t = x_t \) during the \( t \)th DR event when action \( a_t = a_t \) and the ADR state \( s_t = s_t \). The tuple \((S, A, X, P, R, Q)\) is a POMDP that completely describes the ADR maintenance problem. Note that we deviate from the standard definition of POMDP (see [15]) wherein the observation \( x_t \) is a function of the current state \( s_t \) and the previous action \( a_{t-1} \).

We do so in order to keep the dynamics in the problem more transparent. Fig. 2 describes the causal relationships between states, actions, reward, and observations in the ADR maintenance problem.

It is well known [16] that a POMDP is equivalent to a belief-state MDP. Let \( h_t = \{a_1, x_1, a_2, x_2, \ldots, a_t-1, x_{t-1}\} \) denote the observed history before the \( t \)th DR event. Then

\[
b_t = \mathbb{P}(s_t = r_t | h_t) \in B = [0, 1]
\]

(5)

denotes the belief probability that the ADR is operational during the \( t \)th DR event. The expected profit \( r(b) \) in belief state \( b \in B \) is given by

\[
r(b) = [r_{\alpha_0}(b), r_{\alpha_1}(b)]^\top = R [1 - b, b]^\top = [\lambda b, \lambda - c]^\top.
\]

(6)

Note that the updated belief state \( b_{t+1} \) after taking action \( a_t \) and observing the meter readings \( x_t \) is \( b_{t+1} = P(s_{t+1} = \gamma_1 | h_t, a_t, x_t) \). Using Bayes’ rule and the causality structure in Fig. 2, one can show that \( b_{t+1} = \Gamma_{\alpha_t}(b_t) \), where the map

\[
\Gamma_{\alpha_t}(b) = \frac{P_{\gamma_1,\gamma_1}(a)Q_{\gamma_1,\alpha}(b) + P_{\gamma_1,\gamma_1}(a)Q_{\gamma_1,\alpha}(1 - b)}{W_{ab}(x)}
\]

(7)

and \( W_{ab}(x) := \mathbb{P}(x | b, a) = Q_{\gamma_1,\alpha}(b) + Q_{\gamma_1,\alpha}(1 - b) \). The belief transition (7) is standard for POMDPs; see Appendix A for details. The tuple \((B, A, \Gamma, r)\) is the belief-state MDP that represents the ADR maintenance problem.
A. Optimal Policy

Our goal is to compute a policy π*: B → A that maximizes the total expected discounted profit

\[ V(\pi)(b_1) = \sum_{t=1}^{\infty} \beta^{t-1} r_{a_t}^\pi(b_t) \]  

where \( \beta \in (0, 1) \) is a discount factor and \( a_t^\pi \) denotes the action taken by policy \( \pi \) in belief state \( b_t \) prior to the \( t \)th DR event. It is well known [17] that there exists a stationary optimal policy \( \pi^* \in \arg\max_{\pi} V^\pi(b_1) \), and that the associated optimal value function \( V^\pi(b) \), satisfies the Bellman equation

\[ V(b) = \max_{a \in A} \left\{ r_a(b) + \beta \mathbb{E}_x[V(\Gamma_{a \times x})] \right\} \]  

where \( \mathbb{E}_x[V(\Gamma_{a \times x})] = \int_{x \in X} V(\Gamma_{a \times x}) \mathbb{P}(x \mid b, a) dx = \int_{x \in X} V(\Gamma_{a \times x}) W_{ab}(x) dx \) denotes the expected future profit. From (9), it follows that action \( a_1 \) is optimal for belief state \( b \) if the function:

\[ f(b) := r_{a_0}(b) - r_{a_1}(b) + \beta \mathbb{E}_x[V(\Gamma_{a_0 \times x})] - \beta \mathbb{E}_x[V(\Gamma_{a_1 \times x})] \leq 0. \]  

We use the following result to show that the optimal policy is a single threshold policy.

Lemma 1: The function \( f \) is convex in \([0, 1]\).

Proof: Note that \( r_{a_0} = \lambda b \) is an affine function of \( b \) and \( r_{a_1}(b) = \lambda - c \) is independent of \( b \). It follows that \( r_{a_1}(b) - r_{a_1}(b) \) is an affine and, therefore, convex function of \( b \). Also, since \( \Gamma_{a_1 \times x} = (1 - p) \) is a convex function of \( b \), in fact, independent of \( b \).

The value function \( V \) of an infinite horizon MDP is convex on \([0, 1]\) [17]. Then, the perspective function \( g(v, u) = u V(v/u) \) is convex on \( \{(v, u) : v/u \in [0, 1], v \in \mathbb{R}_+\} \) [18]. Since \( W_{a_0 \times x}(b) = Q_{\gamma_0 \times a_0}(x)b + Q_{\gamma_0 \times a_0}(x)(1 - b) \) and

\[ \Gamma_{a_0 \times x}(b) = \frac{(1 - p)Q_{\gamma_0 \times a_0}(x)b}{W_{a_0 \times x}(b)} \]  

it follows that:

\[ g_x(b) := V(\Gamma_{a_0 \times x}(b)) W_{a_0 \times x}(b) = g((1 - p)Q_{\gamma_0 \times a_0}(x)b, Q_{\gamma_0 \times a_0}(x)b + Q_{\gamma_0 \times a_0}(x)(1 - b)) \]  

is a convex function of \( b \) for each \( x \in X \). Hence, \( \mathbb{E}_x[V(\Gamma_{a_0 \times x}(b))] = \int_{x \in X} g_x(b) dx \) is a convex function of \( b \).

Theorem 1: The optimal policy \( \pi^* \) for the ADR maintenance problem is a threshold policy, i.e., there exists \( b^* \in \mathbb{R} \) such that

\[ a_1^\pi = \begin{cases} a_1 & \text{if } b_1 \leq b^* \\ a_0 & \text{if } b_1 > b^* \end{cases} \]  

Proof: From Lemma 1, we have that \( f(b) \) is convex. Consequently, the set \( I = \{b : f(b) \leq 0\} \) of belief states for which action \( a_1 \) is optimal is a closed, possibly empty, interval. Suppose \( I \) is empty. Then the optimal policy is of the form (13) with \( b^* < 0 \).

Next, suppose \( I \) is nonempty. To establish the result, it suffices to show that \( 0 \in I \). Suppose not, i.e., \( a_0 \) is strictly optimal for \( b = 0 \). Then

\[ V(0) = r_{a_0}(0) + \beta \mathbb{E}_x[V(\Gamma_{a_0 \times x}(0))] = 0 + \beta V(0) \]  

where we use fact that \( \Gamma_{a_0 \times x}(0) = 0 \). It follows that \( V(0) = 0 \). Moreover

\[ 0 = V(0) > r_{a_1}(0) + \beta \mathbb{E}_x[V(\Gamma_{a_1 \times x}(0))] = \lambda - c + \beta V(1 - p). \]  

Then, for any belief state \( b \)

\[ r_{a_1}(b) + \beta \mathbb{E}_x[V(\Gamma_{a_1 \times x}(b))] = \lambda - c + \beta V(1 - p) < 0. \]  

Since the payoff from action \( a_0 \) is non-negative in any belief state, we have that \( V(b) \geq 0 \) for all \( b \in B \). It follows that \( V(b) = r_{a_0}(b) + \beta \mathbb{E}_x[V(\Gamma_{a_0 \times x}(b))] \), and \( a_0 \) is the unique optimal action for all belief states. Thus, we have established that the interval \( I \) is empty; a contradiction. Hence, \( 0 \in I \), and the optimal policy is of the form (13) with \( b^* \in [0, 1] \).

We conclude this section with the following corollary.

Corollary 1: The level sets \( \{b : f(b) \leq -\mu\} \) of the function \( f \) are all of the form \([0, \xi]\), where we use the convention that \([0, \xi]\) is empty.

Proof: Consider an ADR maintenance problem where the reward associated with the do-nothing action \( a_0 \) is increased by an amount \( \mu \). Then the interval over which it is optimal to take action \( a_1 \) is given by \( \{b : f(b) + \mu \leq 0\} \). From the proof of Theorem 1, it follows that the set \( \{b : f(b) \leq -\mu\} \) is of the form \([0, \xi]\).

We use Corollary 1 in Section III to establish that the ADR repair scheduling problem is Whittle indexable [14] and, therefore, that there exists a well-defined heuristic policy for approximately solving it.

B. Approximating Optimal Threshold

In this section, we describe a numerical scheme to approximate the optimal threshold \( b^* \). We discretize the space \( B = [0, 1] \) into \( n + 1 \) equally spaced points \( B = \{k/n : k = 0, \ldots, n\} \), and round up \( b \) to \( (k - 1)/n, k/n \) to the point \( k/n \). Formally, we consider the MDP with state space \( \hat{B} \), reward function \( r_i \mid \hat{B} \), and state transition \( \hat{\Gamma}_{a \times x}(k) := [\Gamma_{a \times x}(k/n)] \). Let \( \hat{V} = (\hat{V}(0), \ldots, \hat{V}(k), \ldots, \hat{V}(n)) \) denote the value function vector for the \( n + 1 \) states \( k/n \in \hat{B} \). Then

\[ \hat{V}(k) = \max \left\{ r_{a_0}(k/n) + \beta \mathbb{E}_x[\hat{V}(\hat{\Gamma}_{a_0 \times x}(k/n))], r_{a_1}(k/n) + \beta \mathbb{E}_x[\hat{V}(\hat{\Gamma}_{a_1 \times x}(k/n))] \right\}. \]  

It is well known [19] that the vector \( \hat{V} \) can be computed by solving the LP

\[ \min_{\hat{V}} \sum_{k=0}^{n} V(k) \]

subject to

\[ V(k) \geq r_{a_0}(k/n) + \beta \mathbb{E}_x[V(\hat{\Gamma}_{a_0 \times x}(k/n))], \quad \forall k \]

\[ V(k) \geq r_{a_1}(k/n) + \beta V((1 - p)n), \quad \forall k \]  

where we use the fact that \( \hat{\Gamma}_{a_1 \times x}(k) = [(1 - p)n] \). LP (18) consists of \( n \) variables and \( 2n \) constraints and, therefore,
can be solved very fast, provided the conditional expectations \( \mathbb{E}_x[V(\hat{\Gamma}_{a_0x}(k))] \) can be computed efficiently. In Section IV, we show how to efficiently approximate the conditional expectations in (18) using low discrepancy sequences. Given the approximate value function \( \hat{V} \), we approximate threshold \( b^* \) with the largest state \( k/n \) for which the repair action \( a_1 \) is optimal, i.e., \( b^* \approx 1/n \max\{k : \hat{V}(k) = r_{a_1}(k/n) + \beta \hat{V}((1-p)n)\} \).

### III. ADR Repair Scheduling Problem

Suppose there are \( M \) maintenance crews available to maintain \( D(\gg M) \) ADRs. Before each DR event \( t \), the utility must decide which (if any) ADRs to inspect and reset. We formulate the ADR repair scheduling problem as an RB problem [11] where each of the ADR devices is a bandit. Following the notation of bandit problems, we will call an ADR active if it is being inspected, and inactive otherwise.

For RB problems, Whittle [11] proposed a possibly suboptimal policy based on the Lagrangian relaxation for the dynamic program. The Lagrangian relaxation decouples the RB problem into a collection of so-called subsidy-\( \mu \) problems. For each bandit, the subsidy-\( \mu \) problem is an ADR maintenance problem in which the reward under the do-nothing action \( a_0 \) is increased by an amount \( \mu \geq 0 \).

Let \( B_0(\mu) \) denote the set of states for which the do-nothing action \( a_0 \) is optimal for a subsidy level \( \mu \). An RB problem is said to be indexable if \( B_0(\mu) \subseteq B_0(\xi) \) whenever \( \mu \leq \xi \). For indexable RB problems, the Whittle index \( \mu_k \) of bandit \( k \) in state \( b_k \) is defined as the minimum subsidy \( \mu \geq 0 \) which makes the passive action \( a_0 \) optimal at \( b_k \). Let \( K = \{k(\ell) : \ell = 1, \ldots, K\} \) denote the indices of bandits with strictly positive Whittle index arranged in decreasing order of the index, i.e., \( \mu_{\ell-1} \geq \mu_\ell \) whenever \( \ell \leq i \). The Whittle index policy specifies that the set of bandits that are activated at time \( t \) is given by \( K_t = \{k(\ell) : 1 \leq \ell \leq \min\{K, M\}\} \), i.e., at most \( M \) bandits with the largest strictly positive Whittle index are activated. The next theorem ensures that the Whittle indices are well defined for our problem.

**Theorem 2:** The ADR repair scheduling problem is indexable.

**Proof:** The Bellman equation for the subsidy-\( \mu \) problem corresponding to the ADR scheduling problem is given by

\[
V_\mu(b) = \max_{a \in A} \left\{ r_a(b, \mu) + \beta \mathbb{E}_x[V_\mu(\Gamma_{a_0x}(b))] \right\}
\]

where \( r(b, \mu) = r(b) + (\mu, 0)^T \). The set of states for which \( a_0 \) is optimal is of the form \( B_0(\mu) = \{b \in [0, 1] : b > b^*(\mu)\} \) where \( b^*(\mu) \) denotes the optimal threshold corresponding to the subsidy-\( \mu \) problem (19). Hence, in order to establish that the ADR repair scheduling problem is indexable, we need to show that the threshold \( b^*(\mu) \) is nonincreasing in \( \mu \). From Corollary 1, it follows that \( [0, b^*(\mu)] = \{b : f(b) \leq -\mu\} \). Thus, \( b^*(\mu) \) is nonincreasing in \( \mu \).

Since the Whittle index \( \mu^*(b) = \inf\{\mu \geq 0 : b \in B_0(\mu)\} = \inf\{\mu \geq 0 : b > b^*(\mu)\} \), it is clear that \( \bar{\mu} = \inf\{\mu \geq 0 : b^*(\mu) < 0\} \) is an upper bound for \( \mu^*(b) \). Therefore, the Whittle index for any \( b \in B \) can be computed to within an accuracy \( \epsilon \) using a binary search by solving at most \( O(\log_2(\bar{\mu}/\epsilon)) \) LPs of the form (18). Thus, in practice, one needs to solve at most \( O(n \log_2(\bar{\mu}/\epsilon)) \) LPs.

On the other hand, we can also compute Whittle indices to within \( \epsilon \) accuracy by finding the optimal threshold \( b^*(\mu) \) for all \( \mu \in \mathbb{N} = \{k \in [0, \bar{\mu}] \} \). Thus, the overall complexity of computing the Whittle indices is at most \( O(\min\{n \log_2(\bar{\mu}/\epsilon), [\bar{\mu}/\epsilon]\}) \). In our numerical experiments, we calculated a bound on \( \bar{\mu} \) using a doubling strategy, and used the binary search approach to compute the Whittle indices.

### IV. Numerical Implementation

Recall that as long as we can compute the conditional expectation \( \mathbb{E}_x[V(\hat{\Gamma}_{a_0x}(k))] \) efficiently, we can solve LP (18) and, consequently, implement the Whittle index policy. In this section, we discuss how to efficiently compute the required conditional expectation in two situations that arise in the DR context: 1) when the DR load-shed is random; and 2) when the clocks at the utility and at the meter are not synchronized.

We show that these issues can have a significant impact on the performance of the Whittle index policy, and propose a variational Bayes procedure to address them. Finally, we evaluate the performance of our proposed policies with respect to periodic review policies, and with respect to policies that have full information of the state of the ADRs.

#### A. Deterministic Load-Shed and Synchronized Clocks

Let \( x = (x_1, \ldots, x_m) \) denote the sampled power consumption over a DR event of length \( m \) periods. For \( i = 1, \ldots, m \), we assume that the power consumption \( x_i \) is given by

\[
x_i = \begin{cases} y_i - r + \hat{\epsilon}_i & \text{if the ADR is operational} \\
y_i + \hat{\epsilon}_i & \text{otherwise} \end{cases}
\]

where \( y_i \) denotes the estimated power consumption on a non-DR day, \( r \) denotes the load-shed mandated by the utility, and the estimation residual \( \hat{\epsilon}_i \) is assumed to be independent and identically distributed (IID) according to a normal distribution with mean 0 and variance \( \sigma^2 \). As will become clear below, we can work with any specification for the residuals \( \hat{\epsilon}_i \) as long as one is able to simulate samples from the distribution.

We approximate the conditional expectation \( \mathbb{E}_x[V(\hat{\Gamma}_{a_0x}(k))] \) in LP (18) by the finite weighted sum

\[
\mathbb{E}_x[V(\hat{\Gamma}_{a_0x}(k))] \approx \frac{1}{N} \sum_{j=1}^{N} V(\hat{\Gamma}_{a_0x}(k))
\]

where the samples \( \{x^j : j = 1, \ldots, N\} \) are generated IID from the distribution \( W_{a_0, b}(x) \), \( b = k/n \), that is

\[
x^j \sim \begin{cases} N(y, \sigma^2 I) & \text{w.p.} 1-b \\
N(y - r, \sigma^2 I) & \text{w.p.} b \end{cases}
\]

To efficiently sample from the multivariate normal distributions, we use low-discrepancy sequences in the unit \( m \)-dimensional hypercube [20]. Let \( \omega_1, \ldots, \omega_N \in [0, 1]^m \) be the first \( N \) elements of a low-discrepancy sequence. We define \( z^j \in \mathbb{R}^m \) by setting \( z^j_\ell = \Phi^{-1}(\omega^j_\ell) \). Then \( \{z^j : j = 1, \ldots, N\} \) are IID samples from an \( m \)-dimensional standard normal distribution.

To efficiently sample from the multivariate normal distributions, we use low-discrepancy sequences in the unit \( m \)-dimensional hypercube [20]. Let \( \omega_1, \ldots, \omega_N \in [0, 1]^m \) be the first \( N \) elements of a low-discrepancy sequence. We define \( z^j \in \mathbb{R}^m \) by setting \( z^j_\ell = \Phi^{-1}(\omega^j_\ell) \). Then \( \{z^j : j = 1, \ldots, N\} \) are IID samples from an \( m \)-dimensional standard normal distribution. 

To efficiently sample from the multivariate normal distributions, we use low-discrepancy sequences in the unit \( m \)-dimensional hypercube [20]. Let \( \omega_1, \ldots, \omega_N \in [0, 1]^m \) be the first \( N \) elements of a low-discrepancy sequence. We define \( z^j \in \mathbb{R}^m \) by setting \( z^j_\ell = \Phi^{-1}(\omega^j_\ell) \). Then \( \{z^j : j = 1, \ldots, N\} \) are IID samples from an \( m \)-dimensional standard normal distribution. 

To efficiently sample from the multivariate normal distributions, we use low-discrepancy sequences in the unit \( m \)-dimensional hypercube [20]. Let \( \omega_1, \ldots, \omega_N \in [0, 1]^m \) be the first \( N \) elements of a low-discrepancy sequence. We define \( z^j \in \mathbb{R}^m \) by setting \( z^j_\ell = \Phi^{-1}(\omega^j_\ell) \). Then \( \{z^j : j = 1, \ldots, N\} \) are IID samples from an \( m \)-dimensional standard normal distribution.
random variable [21]. We generate meter readings sample \( x_i \) as follows:

\[
x_i = \begin{cases} 
  y + \sigma z_i & \text{w.p. } 1 - b \\
  y - r_1 + \sigma z_i & \text{w.p. } b.
\end{cases}
\]  

(23)

Let \( Q_0(x) \) [resp. \( Q_1(x) \)] denote the probability of observing a vector of meter readings \( x \) under an functional (resp. non-functional) ADR. Then, \( Q_{\gamma \alpha_0}(x) = Q_0(x) \) and \( Q_{\gamma \alpha_1}(x) = Q_{\gamma \alpha_0}(x) = Q_1(x) \), where

\[
Q_0(x) = \prod_{i=1}^{m} \varphi \left( \frac{x_i - y_i}{\sigma} \right), \quad Q_1(x) = \prod_{i=1}^{m} \varphi \left( \frac{x_i - y_i + r_i}{\sigma} \right). \]  

(24)

Given \( Q_0(x') \) and \( Q_1(x') \), we compute \( \hat{g}_{\alpha_0,\nu}(k) = \left[ \Gamma_{\alpha_0,\nu}(b) \right] \) using (7).

B. Random Load-Shed and Unsynchronized Clocks

Here, we assume that the distribution of the load-shed is known—but the exact magnitude is unknown—and that the clock at the utility and in the meter can be mismatched up to \( d \) time units. We use the utility’s clock time as the reference time for all computations, and assume that meter readings are assigned to time instants using the meter clock; therefore, an observation assigned to the time instant \( i \) could, in fact, correspond to an actual instant in the set \( \{ i - d, \ldots, i + d \} \). Suppose a DR event takes place during the period \( \{ d + 1, \ldots, d + m \} \).

Then

\[
x \sim \begin{cases} 
  \mathcal{N}(y - r_1, \sigma I) & \text{for some } r, \delta \in \{ -d, \ldots, d \} \text{ the ADR operational} \\
  \mathcal{N}(y, \sigma^2 I) & \text{the ADR is not operational}
\end{cases}
\]  

(25)

where \( I \in \mathbb{R}^{m+2d} \) denotes a vector with the components \( i \in I_\delta = \{ 1 + d + \delta, \ldots, m + d + \delta \} \) equal to 1, and all other components equal to zero.

Let \( \omega^1, \ldots, \omega^N \) be the first \( N \) elements of a low-discrepancy sequence in the \( (m + 2d) \)-dimensional hypercube and \( \omega^j_i = \Phi^{-1}(\nu_i) \). Then, the \( j \)-th sample from the distribution \( W_{\alpha_0,\nu}(x') \) is given by

\[
x' = \begin{cases} 
  y + \sigma z_i & \text{w.p. } 1 - b \\
  y - r_1 + \sigma z_i & \text{w.p. } b \rho(r, \delta)
\end{cases}
\]  

(26)

where \( \rho \) denotes the joint probability distribution function of the load-shed and the clock mismatch. In this case, the observation probabilities are

\[
Q_0(x) = \prod_{i=1}^{m+2d} \varphi \left( \frac{x_i - y_i}{\sigma} \right), \quad Q_1(x) = \mathbb{E}_x \left[ \prod_{i \in I_k} \varphi \left( \frac{x_i - y_i}{\sigma} \right) \prod_{i \notin I_k} \varphi \left( \frac{x_i - y_i + r_i}{\sigma} \right) \right] \]  

(27)

where the expectation is with respect to the posterior distribution of the mismatch \( \delta \) and the load-shed \( r \). In the next section, we show how to use the variational Bayes procedure to approximate the posterior expectation. Given \( Q_1(x') \) and \( Q_0(x') \), we approximate \( \mathbb{E}_x[V(\Gamma_{\alpha_0,\nu}(k))] \) using (7) and (21).

C. Belief State Update Using Variational Bayes

Consider the following hierarchical Bayesian model:

\[
x = y - r_1 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \quad r \sim \mathcal{N}(\nu_0, 1/\eta_0) \]  

(28)

where \( \nu_0 \) is the expected load-shed during the DR event, and \( \eta_0^{-1/2} \) is its standard deviation. The posterior distribution \( \rho(r, \delta | x, \nu) \) is proportional to \( \mathbb{P}(x | r, \delta, \sigma) \mathbb{P}(r | \nu_0, \eta_0) \mathbb{P}(\delta) \), where

\[
\mathbb{P}(x | r, \delta, \sigma) = \prod_{i \in I_k} \varphi \left( \frac{x_i - y_i + r_i}{\sigma} \right) \prod_{i \notin I_k} \varphi \left( \frac{x_i - y_i}{\sigma} \right). \]  

(29)

Since this posterior distribution is neither in closed form, nor is it easy to sample from, we use the variational inference method [22] and approximate the posterior distribution by the product distribution \( \Delta(\delta) \varrho(r) \). It is easy to establish (see Appendix B for details) that the product distribution that minimizes the Kullback–Leibler distance from the joint posterior distribution is of the form

\[
\Delta(\delta) \varrho(r) \propto \exp \left\{ \frac{\nu}{\sigma^2} 1^T \nu_0 \mathbf{1}_k - \frac{\nu^2 + \eta^{-1}}{2\sigma^2} ||\mathbf{1}_k||_2^2 \right\} \]  

(30)

\[
g(r) = \mathcal{N}(\nu, 1/\eta) \]  

(31)

where \( \nu = \nu_0 + (1/\sigma^2) \sum_{\delta = -d}^{d} \Delta(\delta) \mathbf{1}_k^T \nu_0 \mathbf{1}_k \) and \( \eta = \eta_0^{-1} \nu_0 + (1/\sigma^2) \sum_{\delta = -d}^{d} \Delta(\delta) \mathbf{1}_k^T \nu - \mathbf{x} \). Note the circular dependence of the parameters in the posterior distributions. We use an iterative procedure that alternates between \( \Delta(\delta) \), and the precision \( \eta \) and mean \( \nu \) of the normal distribution for \( r \). We terminate the procedure whenever the relative change in the parameters is small. We approximate the probability \( Q_1(x) \) using the posterior distributions \( \Delta(\delta) \) and \( g(r) \) as follows:

\[
Q_1(x) \approx \sum_{\ell = -L}^{L} \sum_{\delta = -d}^{d} \prod_{i \in I_k} \varphi \left( \frac{x_i - y_i + r_i}{\sigma} \right) \prod_{i \notin I_k} \varphi \left( \frac{x_i - y_i}{\sigma} \right) \Delta(\delta) g(r) \]  

(32)

where \( r_k \) is \( k \) standard deviations \( \eta^{-1/2} \) above the mean \( \nu \). The posterior distribution in the case where either the meter readings are perfectly synchronized or the load-shed magnitude is deterministic can be computed as a special case of this procedure.

D. Problem Parameters and Available Information

We consider the ADR maintenance problem in the following four settings.

Case (a) Clocks synchronized and load-shed deterministic.

Case (b) Clocks possibly mismatched with \( d = 2 \), but load-shed deterministic.

Case (c) Clocks synchronized, but load-shed distributed \( \mathcal{N}(\nu_0, 1/\eta_0) \).

Case (d) Clocks possibly mismatched with \( d = 2 \), and load-shed distributed \( \mathcal{N}(\nu_0, 1/\eta_0) \).
The parameter values for the numerical experiments were set as follows:

1) observations in an hour-long DR event \( m = 10 \);
2) probability of ADR failure \( p = 0.05 \);
3) expected DR savings for utility \( \lambda = 1 \);
4) cost of repair \( c = 3\lambda = 3 \);
5) discount factor \( \beta = 0.9 \);
6) load-shed SNR \( = 20\log_{10}(b_0/\sigma) = (-5, 0, 5)\) dB, i.e., standard deviation \( \sigma \) of the meter noise \( \approx \{1.78, 1, 0.56\} \) times the mean load-shed \( b_0 \);
7) load-shed standard deviation \( b_0^{-1/2} = 0.1\sigma \).

**E. Discretization Size \( n \) and Sample Size \( N \)**

In Table I, we report the optimal threshold \( b^* \), the approximate optimal value function \( V(b) \) at belief state \( b = 1 \), and the elapsed time in seconds to approximate the expectation \( \mathbb{E}_\pi[V(\Gamma_{\alpha_0}(k))] \) and solve LP (18), as a function of the number of samples \( N \) does not have a significant impact on the value function \( V(1) \). We interrupted the computation of the optimal threshold and the optimal value for Case (d) for the largest-sized approximations, i.e., \( N = 500 \) K. The solution time was too large to be of practical use. In Table I, we report this situation with an horizontal line. For the rest of the experiments in this section, we used \( n = 100 \) and \( N = 5 \) K.

**F. Comparison With Periodic Review Policies**

In this section, we report the performance of the proposed threshold policy with respect to the current practice of periodically inspecting the ADRs. Suppose the periodic review interval is \( q \) DR events. Then, the associated value function \( U(q) \) satisfies the recursion

\[
U(q) = -c + \lambda \sum_{j=0}^{q-1} b^j (1-p)^j + \beta^q U(q)
\]

i.e., \( U(q) = 1/(1-\beta^q)(\lambda(1-\beta^q(1-p)^q))/(1-\beta(1-p)-c) \). The value function \( U(q) \) is maximized at \( q^* = 18 \), where \( U(q^*) = 4.10 \). In contrast, the POMDP value function is at least \( V(0) = \{5.14, 5.37, 5.51\} \) for SNR = \(-5, 0, 5\) dB, i.e., a relative improvement over \( U(q^*) \) of at least \( 25\%, 31\%, 34\% \). These results clearly show that the POMDP-based method significantly outperforms any regular maintenance schedule.

**G. ADR Repair Scheduling Problem**

In this section, we report the numerical results for an ADR repair scheduling problem with \( D = 100 \) ADRs and \( M = 5 \) repair crews. We considered two variants of the ADR repair scheduling problem: one where the repair costs for all the ADRs were identical and set equal to \( 3\lambda \), and another where repair cost for each ADR were sampled uniformly from the interval \([-5, 6.5] \lambda \). For \( c > 6.5\lambda \), the utility is better off not repairing the ADR at all.

When the states of the ADRs are fully observable and the repair costs are identical, the associated multidimensional DP can be reduced to a \( 1 \)-D DP with state given by the number of nonworking devices. Thus, one can easily compute the optimal policy using the value iteration algorithm. In this case, we compare the performance of the Whittle index policy for the POMDP with this optimal policy. When the states of the ADRs are fully observable but the repair costs are not identical, it is not possible to compute the optimal policy. In this setting, we compare the Whittle index policy for the POMDP with the Whittle index policy for the full information MDP. Given that the partially observable Whittle index policy can infer the state of an ADR only after observing the meter readings during a DR event, a fairer comparison would be against the optimal policy of the MDP where the state of the ADR is observable only after a DR event. We call this problem, the slow information MDP.

We computed the value function of the Whittle index policies using simulation. We simulated the performance of the policies over a time horizon of \( T = 44 \) DR events with a discount factor \( \beta = 0.9 \). This implies that the simulated \( T \)-horizon value function is within \( 1\% \) of the infinite horizon value function. The results are averaged over 100 runs. The column marked “Whittle” in Table II reports the relative error of the full information Whittle index policy with respect to the optimal full and slow information policy when all the repair costs are identical. The next four columns report the performance of the partial information Whittle index policy for the four different cases listed in Section IV-D. Since the partial information policy does not have access to the state, its performance depends on the load-shed SNR. The row marked

| SNR  | Full | Slow |
|------|------|------|
| 5    | 1.28 | 4.43 |
| 0    | 1.28 | 5.38 |
| -5   | 1.28 | 8.48 |
| [-5, 5] | 1.28 | 5.76 |

**TABLE I**

| \( n \) | \( N \) | Case (a) | Case (b) | Case (c) | Case (d) | \( b^* \) |
|--------|--------|----------|----------|----------|----------|----------|
| 100    | 5K     | 0.160    | 0.150    | 0.170    | 0.150    |          |
| 100    | 500K   | 0.160    | 0.150    | 0.160    |          |          |
| 1K     | 5K     | 0.160    | 0.154    | 0.171    | 0.159    |          |
| 1K     | 500K   | 0.163    | 0.154    | 0.166    |          |          |
| 100    | 5K     | 8.574    | 8.558    | 8.598    | 8.713    |          |
| 100    | 500K   | 8.509    | 8.550    | 8.448    |          |          |
| 1K     | 5K     | 8.792    | 8.477    | 8.423    | 8.648    |          |
| 1K     | 500K   | 8.275    | 8.515    | 8.412    |          |          |
| 100    | 5K     | 33.4     | 1373     | 8205     |          |          |

**TABLE II**

| SNR  | Whittle | Case (a) | Case (b) | Case (c) | Case (d) |
|------|---------|----------|----------|----------|----------|
| 5    | 1.28    | 4.63     | 4.62     | 4.60     | 4.61     |
| 0    | 1.28    | 5.38     | 5.92     | 5.77     | 5.66     |
| -5   | 1.28    | 8.48     | 9.63     | 9.23     | 10.75    |
| [-5, 5] | 1.28 | 5.76 | 6.3 | 6.14 | 6.96 |

| SNR  | Full | Slow |
|------|------|------|
| 5    | 1.66 | 4.57 |
| 0    | 1.66 | 3.3  |
| -5   | 1.66 | 5.5  |
| [-5, 5] | 1.66 | 2.74 | 3.3 | 3.13 | 3.95 |
by SNR [−5,5] reports the performance of the policy when the SNR of each ADR is sampled uniformly from the interval [−5,5]; all other rows report the performance when the SNR for all ADRs was set equal to the value corresponding to that row.

The full information Whittle index policy is close to the optimal policy both in the full information and the slow information case. For very low noise levels, i.e., SNR = 5 dB, the performance of partial information Whittle index policy is no more than 4.63% (resp. 1.57%) worse than that of the full (resp. slow) information optimal policy. On the other hand, for very high noise levels, i.e., SNR = −5 dB, the partial information Whittle index policy could be as bad as 10.75% (resp. 7.89%) suboptimal with respect to the full (resp. slow) information optimal policy. For reasonable noise levels, i.e., SNR = 0 dB for all ADRs, or randomly drawn from [−5,5] dB, the performance of the partial information Whittle index policy is within 6.94% (resp. 3.95%) of the full (resp. slow) information optimal policy. The results in this table suggest that the four cases in decreasing order of the suboptimality of the full information Whittle index policy when the repair costs are not identical. Assuming that the full information Whittle index policy is close to the optimal policy; also in this case, the results in Table III are similar to those observed in the case of identical repair costs.

Table III shows the relative error of the partial information Whittle index policy with respect to the full and slow information Whittle index policies [see Section IV-D for the details on partial information cases (a)–(d)].

| SNR | Case (a) | Case (b) | Case (c) | Case (d) |
|-----|---------|---------|---------|---------|
| 5   | 3.28    | 3.32    | 3.32    | 3.39    |
| 0   | 4.26    | 4.87    | 4.7     | 5.45    |
| -5  | 7.15    | 8.04    | 7.73    | 8.93    |
| [-5,5] | 4.61 | 5.15 | 4.96 | 5.66 |

The full information Whittle index policy is close to the optimal policy; also in this case, the results in Table III are similar to those observed in the case of identical repair costs.

V. CONCLUSION

In this paper, we formulated and solved the ADR repair scheduling problem where the goal is to maintain $D$ ADRs using at most $M$ ($\ll D$) maintenance crews and the ADR state is only partially observable via noisy meter readings. We formulated this problem as an RB problem. We showed that the ADR repair is Whittle indexable, and therefore, one can very efficiently compute a good heuristic policy by suitably decomposing the $D$-ADR repair scheduling problem into $D$ single-ADR maintenance problems. We showed that the optimal solution of the single-ADR maintenance problem is a threshold policy, where it is optimal to send the maintenance crew as soon the belief state drops below a threshold. Using the structure of the single-ADR optimal policy, we showed that the Whittle index as a function of the belief state of an ADR can be computed via a single binary search. We explored the performance of the Whittle index policies when the meter and utility clocks are (resp. not) synchronized and the load-shed is random (resp. deterministic) (see Section IV-D for the details of the four cases). We also developed a hierarchical Bayesian method for computing the joint posterior distribution of the load-shed and clock mismatch. Our numerical experiments suggest that, when the level of noise in the meter readings is on average of the same size of the load shed, the Whittle index policy is at most 4% suboptimal. This problem and its solution was motivated by our research collaboration with AutoGrid, a software provider for managing DR programs.

APPENDIX A

BELIEF-STATE UPDATE

Recall that the updated belief state after taking action $a_t$ and observing the meter readings $x_t$ is $b_{t+1} = \mathbb{P}(s_{t+1} = \gamma_1 \mid h_t, a_t, x_t)$, where $h_t = \{a_1, x_1, a_2, x_2, \ldots, a_{t-1}, x_{t-1}\}$ denotes the observed history before the $t$th DR event. Using the fact that $s_{t+1}$ is conditionally independent of $(h_t, x_t)$ given $(s_t, a_t)$, and that $\mathbb{P}(s_{t+1} = \gamma_1 \mid s_t = s, a_t) = P_{s_\gamma}(a_t)$, it follows that:

$$b_{t+1} = \sum_{s \in S} P_{s_\gamma}(a_t)\mathbb{P}(s_t = s \mid h_t, a_t, x_t). \tag{34}$$

Using Bayes’ rule, we can rewrite

$$\mathbb{P}(s_t = s \mid h_t, a_t, x_t) = \frac{\mathbb{P}(x_t \mid s_t = s, h_t, a_t)}{\mathbb{P}(x_t \mid h_t, a_t)}. \tag{35}$$

Using the fact that $x_t$ is a random function of $s_{t+1}$, and the Markovian property, it follows that:

$$\mathbb{P}(x_t \mid s_t, h_t, a_t) = \sum_{s_{t+1} \in S} \mathbb{P}(x_t \mid s_{t+1}, s_t, h_t, a_t)\mathbb{P}(s_{t+1} \mid s_t, h_t, a_t)$$

$$= \sum_{s_{t+1} \in S} \mathbb{P}(x_t \mid s_{t+1})\mathbb{P}(s_{t+1} \mid s_t, a_t)$$

$$= \mathbb{P}(x_t \mid s_t, a_t) = Q_{s_\gamma}(a_t). \tag{36}$$

Since $s$ is conditionally independent of $a_t$ given $h_t$, we also have that

$$\mathbb{P}(s_t = s \mid h_t, a_t) = \begin{cases} b_t & \text{if } s = \gamma_1 \\ 1 - b_t & \text{if } s = \gamma_0. \end{cases} \tag{37}$$

Thus

$$\mathbb{P}(s_t = \gamma_1 \mid h_t, a_t, x_t) = \frac{Q_{s_\gamma}(a_t)b_t}{W_{a_t,b_t}(x_t)} \tag{38}$$

$$\mathbb{P}(s_t = \gamma_0 \mid h_t, a_t, x_t) = \frac{Q_{s_\gamma}(a_t)(1 - b_t)}{W_{a_t,b_t}(x_t)} \tag{39}$$

where

$$W_{a_t,b_t}(x_t) = \mathbb{P}(x_t \mid h_t, a_t)$$

$$= Q_{s_\gamma}(a_t)b_t + Q_{s_\gamma}(a_t)(1 - b_t). \tag{40}$$

Substituting these expressions in (34), we obtain $b_{t+1} = \Gamma_{a_t,b_t}(b_t)$, as defined in (7).
The product distribution $\Delta(\delta)g(r)$ that minimizes the Kullback–Leibler distance from the joint posterior distribution satisfies (see [22])

$$\log \Delta(\delta) = \mathbb{E}_r[\log \rho(r, \delta, x, \mathcal{H})] + \text{constant} \quad (41)$$

$$\log g(r) = \mathbb{E}_d[\log \rho(r, \delta, x, \mathcal{H})] + \text{constant} \quad (42)$$

where the posterior distribution $\rho(r, \delta, x, \mathcal{H})$ is proportional to $\mathbb{P}(x | r, \delta, \sigma)\mathbb{P}(r | \nu_0, \eta_0)\mathbb{P}(\delta)$. Then

$$\log \Delta(\delta) = \mathbb{E}_r[\log \mathbb{P}(x | r, \delta, \sigma) + \log \mathbb{P}(\delta)] + \text{constant}$$

$$= \mathbb{E}_r[-\frac{1}{2\sigma^2}\|x - y + rI_\delta\|^2 - \log(2d + 1)] + \text{constant}$$

$$= \frac{1}{\sigma^2}\left(\mathbb{E}[r]I_\delta^\top(y - x) - \frac{1}{2}\mathbb{E}[r^2]\|I_\delta\|^2\right) + \text{constant} \quad (43)$$

and

$$\log g(r) = \mathbb{E}_d[\log \mathbb{P}(x | r, \delta, \sigma) + \log \mathbb{P}(r | \nu_0, \eta_0)] + \text{constant}$$

$$= \mathbb{E}_d[-\frac{1}{2\sigma^2}\|x - y + rI_\delta\|^2 - \frac{\eta_0}{2}(r - \nu_0)^2] + \text{constant}$$

$$= -\frac{\eta}{2}(r - \nu)^2 + \text{constant} \quad (44)$$

where

$$\eta = \eta_0 + \frac{1}{\sigma^2}\mathbb{E}_d\|I_\delta\|^2$$

$$\nu = \frac{1}{\eta}\left(\nu_0\eta_0 + \frac{1}{\sigma^2}\mathbb{E}\left[I_\delta^\top(y - x)\right]\right).$$

Hence, $\Delta(\delta)$ [resp. $g(r)$] is the density function of a categorical (resp. normal) random variable. Substituting $\mathbb{E}[r] = \nu$ and $\mathbb{E}[r^2] = \nu^2 + \eta^{-1}$ in (43), $\mathbb{E}_d\|I_\delta\|^2 = \sum_{d} \Delta(\delta)\|I_\delta\|^2$ in (45), and $\mathbb{E}[I_\delta^\top(y - x)] = \sum_{d} \Delta(\delta)I_\delta^\top(y - x)$ in (46), we obtain the expressions given in Section IV-C.

### REFERENCES

[1] National Assessment & Action Plan on Demand Response, Fed. Energy Regul. Comm., Washington, DC, USA, 2013.

[2] J. Wells and D. Haas, Electricity Markets: Consumers Could Benefit From Demand Programs, But Challenges Remain, Darby, PA, USA: DIANE, 2004.

[3] M. A. Piette et al., “Design and operation of an open, interoperable automated demand response infrastructure for commercial buildings,” J. Comput. Inf. Sci. Eng., vol. 9, no. 2, pp. 1–9, 2009.

[4] (2014), Open ADR Alliance: Overview. [Online]. Available: http://www.openadr.org/overview

[5] SGIG Accelerates Grid Modernization in Minnesota, Case Study: Minnesota Power, Office Elect. Del. Energy Rel., U.S. Dept. Energy, Washington, DC, USA, 2012.

[6] Breathing New Life Into Legacy Demand Response (White Paper), Silver Spring Netw., Redwood City, CA, USA, 2013.

[7] The Business Case for DLC Replacement (White Paper), Silver Spring Netw., Redwood City, CA, USA, 2013.

[8] C. H. Papadimitriou and J. N. Tsitsiklis, “The complexity of Markov decision processes,” Math. Oper. Res., vol. 12, no. 3, pp. 441–450, 1987.

[9] O. Madani, S. Hanks, and A. Condon, “On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems,” in Proc. AAAI/IAAI, Menlo Park, CA, USA, 1999, pp. 541–548.

[10] J. Endrenyi et al., “The present status of maintenance strategies and the impact of maintenance on reliability,” IEEE Trans. Power Syst., vol. 16, no. 4, pp. 638–646, Nov. 2001.

[11] P. Whittle, “Restless bandits: Activity allocation in a changing world,” J. Appl. Probab., vol. 25, pp. 287–298, 1988.

[12] J. C. Gittins, “Bandit processes and dynamic allocation indices,” J. Roy. Statist. Soc. B, Methodol., vol. 41, no. 2, pp. 148–177, 1979.

[13] C. H. Papadimitriou and J. N. Tsitsiklis, “The complexity of optimal queuing network control,” Math. Oper. Res., vol. 24, no. 2, pp. 293–305, 1999.

[14] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride, “Some indexable families of restless bandit problems,” Adv. Appl. Probab., vol. 38, no. 3, pp. 643–672, Sep. 2006.

[15] D. Brazdil, “POMDP solution methods,” Dept. Comput. Sci., University of Toronto, Tech. Rep., 2003.

[16] K. J. Åström, “Optimal control of Markov processes with incomplete state information,” J. Math. Anal. Appl., vol. 10, no. 1, pp. 174–205, 1965.

[17] E. J. Sondik, “The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs,” Oper. Res., vol. 26, no. 2, pp. 282–304, 1978.

[18] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, U.K.: Cambridge Univ. Press, 2009, ch. 3.2.

[19] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, vol. 414. Hoboken, NJ, USA: Wiley, 2009, ch. 6.

[20] I. M. Sobel, “On quasi-Monte Carlo integrations,” Math. Comput. Simulat., vol. 47, no. 2, pp. 103–112, 1998.

[21] L. Devroye, “Sample-based non-uniform random variate generation,” in Proc. 18th Conf. Winter Simulat., Washington, DC, USA, 1986, pp. 260–265.

[22] C. M. Bishop, Pattern Recognition and Machine Learning, vol. 1. New York, NY, USA: Springer, 2006, pp. 152–160 and 461–516.

Carlos Abad is currently pursuing the Ph.D. degree with the Department of Industrial Engineering and Operation Research, Columbia University, New York, NY, USA.

His current research interests include convex optimization, compressive sensing, Bayesian inference, decision-making under uncertainty, and their applications to the operation and economics of electrical grids.

Garud Iyengar received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1998.

He joined the Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, USA, in 1998, where he teaches courses in asset allocation, asset pricing, simulation, and optimization. His current research interests include convex, robust, and combinatorial optimization; queuing networks; mathematical and computational finance; and communication and information theory.