2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Learning Optimal Policies in Mean Field Models with Kullback-Leibler Regularization

Ana Bušić[1], Sean Meyn[2] and Neil Cammardella[3]

***Abstract*— The theory and application of mean field games has grown significantly since its origins less than two decades ago. This paper considers a special class in which the game is cooperative, and the cost includes a control penalty defined by Kullback-Leibler divergence, as commonly used in reinforcement learning and other fields. Its use as a control cost or regularizer is often preferred because this leads to an attractive solution. This paper considers a particular control paradigm called Kullback-Leibler Quadratic (KLQ) optimal control, and arrives at the following conclusions: 1. in application to distributed control of electric loads, a new modeling technique is introduced to obtain a simple Markov model for each load (the 'agent' in mean field theory). 2. It is argued that the optimality equations may be solved using Monte-Carlo techniques—a specialized version of stochastic gradient descent (SGD). 3. The use of averaging minimizes the asymptotic covariance in the SGD algorithm; the form of the optimal covariance is identified for the first time.**

I. Introduction

This paper concerns a specific class of algorithms for distributed control, whose origins may be traced to Todorov [28] and Karny [16]. More recent related work concerns applications to economics [15] and power systems [12], [13]. A common theme is the introduction of Kullback-Leibler (KL) divergence (i.e., relative entropy) as a control penalty or regularizer. KL regularization is a theme in reinforcement learning [2], it was a component in the first version of the feedback particle filter [20], and is one part of a successful approach to computational methods in optimal transport [14].

The introduction of KL divergence as a regularizer is motivated in part from the calculus of KL divergence that leads to an elegant characterization of the optimizer in these applications. In the class of problems considered in the present paper it is shown that this calculus leads to Monte-Carlo methods for computation, explicit expressions for the asymptotic covariance (the covariance appearing in the Central Limit Theorem), and a simple approach to minimize the covariance.

We consider a finite-horizon optimal control problem with horizon $K \ge 1$, and state trajectory $X_0^K = \{X_k : 0 \le k \le K\}$ evolving on a finite state space X. It is assumed that there is a nominal pmf (probability mass function) p° that models the "control-free" behavior of X_0^K, and that it is Markovian with transition matrix $P_\circ$:

$$p^\circ(x_0, \dots, x_K) = \nu_0(x_0) \prod_{k=0}^{K-1} P_\circ(x_k, x_{k+1}) \qquad (1)$$

in which ν_0 is the initial distribution, $X_0 \sim \nu_0$. A sequence of real-valued functions $\{\varepsilon_k : 1 \le k \le K\}$ on X^{K+1} is given, and the optimization problem to be solved is

$$J^*(\nu_0) = \min_p \Big\{ D(p \| p^\circ) + \frac{\kappa}{2} \sum_{k=1}^{K} \big[\langle p, \varepsilon_k \rangle \big]^2 \Big\} \qquad (2)$$

where D denotes KL divergence, and $\kappa > 0$. The goal is multi-objective: the optimizer p^* should not be far from p°, and satisfy $\langle p^*, \varepsilon_k \rangle := \mathsf{E}[\varepsilon_k(X_0^k)] \approx 0$ for each k (expectation under p^*).

In the applications of interest in this paper, the pmf p is interpreted as an approximation of the histogram of state trajectories in a large ensemble of cooperating agents—the usual setting of mean field control. As is typical in mean field control, the Kullback-Leibler Quadratic (KLQ) optimization problem (2) does not fall into the class of Markov Decision Processes (MDPs) because the objective is a nonlinear function of p. It does fall into the broader category of mean-field Markov decision theory.

Application to load control The KLQ optimal control problem (2) was motivated by distributed control of electric loads [10], [11], and the numerical results concern the special case of control of a large population of thermostatically controlled loads (TCLs), such as water heaters or refrigerators. The internal temperature in these appliances does not suddenly change when power is turned on or off, so they are natural energy storage devices. This interpretation leads to techniques to create *virtual energy storage* to provide grid services that are supplied today through ancillary services from generation, battery systems, capacitor banks, etc.

In this application the aggregate power consumption from a population of loads is required to approximately track a reference signal. In the KLQ optimal control formulation, the random variable $\varepsilon_k(X_0^K)$ represents tracking error over the kth sampling interval.

It was believed that any realistic Markovian model would require an infinite state space, in which X_k includes temperature and power mode at sampling time k (e.g. [19], [21]). One contribution of the present paper is to obtain a finite state space through a new approach that we call *event triggered sampling*. This avoids inaccuracies introduced by binning as in most prior work, or the complexity of models in continuous time [27], [7].

Funding provided through the French National Research Agency grant ANR-16-CE05-0008, and the National Science Foundation under award EPCN 1935389 is gratefully acknowledged.

[1] Inria and DI ENS, École Normale Supérieure, CNRS, PSL Research University, Paris, France

[2] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, and Inria International Chair, Paris

[3] Applications Engineer at Emera Technologies, LLC, Halifax, Canada

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

Contributions and organization The new modeling technique based on event triggered sampling is described in Sec. II. The contribution here is the formulation of an exact and simple Markov model for a system whose dimension is inherently infinite. The structure of the optimal KLQ solution is described in Sec. III, where we find a source of potential complexity: the functions $\{\varepsilon_k\}$ appearing in (2) must be estimated along with p^*. Solutions are proposed in Sec. IV via a tailored version of stochastic gradient descent, based on the special structure of the optimal KLQ control solution, in which p^* is characterized by a Lagrange multiplier $\lambda^* \in \mathbb{R}^K$.

Let $\{\bar{\lambda}^m\}$ denote the sequence of estimates of λ^* obtained using the proposed algorithm. The main result Prop. 4.1 establishes convergence of the scaled mean square error:

$$\lim_{m\to\infty} m\mathsf{E}\big[\big(\bar{\lambda}^m - \lambda^*\big)\big(\bar{\lambda}^m - \lambda^*\big)^{\intercal}\big] = \Sigma^*_\theta\,, \tag{3}$$

$$\Sigma^*_\theta = \tfrac{\kappa^2}{N} G\Sigma_{\mathcal{E}}G\,, \quad \textit{with } G = [I + \kappa\Sigma_{\mathcal{E}}]^{-1}, \tag{4}$$

where $\Sigma_{\mathcal{E}}$ a covariance matrix associated with $(\varepsilon_1, \dots, \varepsilon_K)$, and $N \ge 1$ a parameter in the algorithm.

Literature review See [9], [26] for recent theory of mean field games. Mean-field techniques to approximate the dynamics of a large population of loads began in the 1980s [17], and saw rapid development over the past decade, such as [19], [21], [27], [1]. Much of this work concerns the creation of real-time regulation services (such as automatic generation control, or AGC), while with low frequency balancing services it is valuable to introduce feed-forward control. For example, load peaks are often highly predictable, and grid assets should be given advance warning to prepare. This is the motivation for finite horizon approaches in [3] and the KLQ formulation that is the focus of this paper.

KLQ optimal control was inspired by the finite-horizon control technique introduced in [13], which was inspired by the earlier work [21]. The technical results in the first two subsections of Sec. III of the present paper are minor extensions of [10, Proposition 3.1], based on the new error process defined in (16a).

It may not come as a surprise that the KLQ solution may be approximated using Monte-Carlo methods. The *Z-learning* algorithm of [28] is a reinforcement learning algorithm designed for this purpose for a similar class of problems, but this approach is not applicable because it is based on the interpretation of λ^* as an eigenvector. No such interpretation is possible here.

The matrix appearing on the right hand side of (3) coincides with the covariance matrix of Polyak and Ruppert [25], [22], [23]. These celebrated papers established convergence of $Z_m = \sqrt{m}[\bar{\lambda}^m - \lambda^*]$ in distribution to a Gaussian $N(0, \Sigma^*_\theta)$ random variable, and show that Σ^*_θ is minimal in a matricial sense. We have not found results establishing convergence of expectations of the form $\mathsf{E}[g(Z_m)]$ for functions g that are unbounded. In particular, the limit (3) is new, based on recent SA theory from [5].

II. Event Triggered Sampling

Sampling theory and control design is developed for TCLs. A standard model is a linear system of the form,

$$\tfrac{d}{dt}\Theta_t = -\alpha(\Theta_t - \Theta^a_t) + \beta m_t + W_t, \tag{5}$$

in which Θ_t denotes the temperature and m_t the power mode at time t: $m_t \in \{0, 1\}$ represents whether the power is on or off; βm_t equals power consumption at time t. The remaining terms are: W_t models disturbances (such as usage), Θ^a_t is the ambient temperature, $\alpha > 0$ models leakage due to imperfect insulation, and β is positive for a TCL providing heating, and negative otherwise.

We introduce here a new modeling technique that results in a model in discrete-time and finite-state space *without any approximation*. The main idea is to avoid uniform time-sampling as in all prior work, and instead sample according to internal events at the load.

Event triggered sampling requires a pre-specified finite set of temperature values denoted S, and a mapping $\mathsf{s}\colon \mathbb{R} \times \{0, 1\} \to \mathsf{S}$. The sampling times are then defined by induction as follows: $\tau_0 = 0$, and for $k \ge 0$, at sampling time τ_k we observe the temperature and power mode $x_k = (\Theta_{\tau_k}, m_{\tau_k})$, and compute the *target temperature* $s_{k+1} = \mathsf{s}(x_k) \in \mathsf{S}$. The next sampling time is then defined by

$$\tau_{k+1} := \min\{t \ge \tau_k : \Theta_t = s_{k+1}\} \tag{6}$$

It is entirely consistent with normal TCL behavior to assume that the power mode is constant on each interval $[\tau_k, \tau_{k+1})$. The change in power mode at time τ_{k+1} will be determined through a randomized policy, designed as a small perturbation of the usual hysteresis control.

However, statistics of the TCL play a role in the formulation of the optimal control problem. For example, a dead-beat control solution is defined so that the following identity holds for each k:

$$\mathsf{E}\Big[\int_{\tau_k}^{\tau_{k+1}} \beta m_t\, dt\Big] = \mathsf{E}\Big[\int_{\tau_k}^{\tau_{k+1}} r_t\, dt\Big] \tag{7}$$

where βm_t is power consumption at time t, and r is the reference signal. The right hand side must be computed or approximated to obtain the functions $\{\varepsilon_k\}$ appearing in (2).

Prop. 2.1 that follows does not require a detailed model. The critical assumptions are summarized here:

(A1) *Decision only at sampling times*: Θ_t is continuous, and m_t is right continuous, with jumps only at sampling times:

$$m_t = m_{\tau_k} \qquad \tau_k \le t < \tau_{k+1}$$

Denote $S_k = \Theta_{\tau_k}$ and $U_k = m_{\tau_k}$. Provided the sampling times are finite valued, the temperature dynamics are deterministic:

$$S_{k+1} = \mathsf{s}(X_k)\,, \qquad k \ge 0 \tag{8}$$

Consequently, *regardless of the statistics of the TCL*, the discrete time process $\{X_k = (S_k, U_k) : k \ge 0\}$ is the state process for a controlled Markov chain. This is a tremendous benefit in terms of computational complexity: If sampling is performed uniformly in time then the state space is infinite.

It is assumed that the power mode is determined by a randomized policy, defined as a sequence of conditional pmfs $\{\phi_k\}$ so that for $k \geq 0$,

$$\mathsf{P}\{U_{k+1} = u' \mid \mathcal{F}_k; X_0^k = x_0^k\} = \phi_{k+1}(u'|x_0^k) \tag{9}$$

where $\mathcal{F}_k := \sigma(X_i, \tau_i : i \leq k)$ and $x_0^k := (x_0, \dots, x_k)$. The policy is called Markov if ϕ_k depends only on the most recent state:

$$\mathsf{P}\{U_{k+1} = u \mid \mathcal{F}_k;\, X_k = x_k\} = \phi_{k+1}(u|x_k) \tag{10}$$

It is called a *stationary Markov policy* if $\phi_{k+1}(u'|x_0^k) = \phi(u'|x_k)$ (the function ϕ does not depend on k).

The dynamics of $\boldsymbol{X}$ are determined by both the policy and the target map s which defines the function

$$T(x, s') = \mathbb{I}\{s' = \mathsf{s}(x)\} \tag{11}$$

The state process is Markovian when the policy is Markov:

Proposition 2.1: Consider the model (5) in which the statistics of the disturbance $\boldsymbol{W}$ are arbitrary. Suppose that ϕ is any stationary Markov policy for which each sampling time is finite-valued with probability one. If in addition (A1) holds, then $\boldsymbol{X}$ is a Markov chain on the state space $\mathsf{X} = \mathsf{S} \times \{0, 1\}$. Its transition matrix is expressed for $x \in \mathsf{X}$ and $x' = (s', u') \in \mathsf{X}$ by

$$P_\phi(x, x') = T(x, s')\phi(u'|x)\,. \tag{12}$$

Consequently, for any function $h\colon \mathsf{X} \to \mathbb{R}$ and any k,

$$\begin{aligned} \mathsf{E}[h(X_{k+1}) \mid \mathcal{F}_k; X_k = x] &= \sum_{x' \in \mathsf{X}} P_\phi(x, x')h(x') \\ &= \sum_{u'=0,1} h(\mathsf{s}(x), u')\phi(u'|x) \end{aligned} \tag{13}$$

For analysis we require further assumptions:

(A2) *Semi-Markov model*: $\tau_{k+1} - \tau_k$ is conditionally independent of $\mathcal{F}_k$ given X_k, with conditional distribution independent of k, with conditional distribution functions denoted for $\delta \geq 0$ and $x \in \mathsf{X}$ by

$$F_\Delta(\delta|x) := \mathsf{P}[\tau_{k+1} - \tau_k \leq \delta | \mathcal{F}_k;\, X_k = x] \tag{14}$$

The inter-sampling times are bounded a.s.: there is $\bar{\Delta}^{\max} < \infty$ such that $F_\Delta(\bar{\Delta}^{\max}|x) = 1$ for each x.

The conditional mean sampling interval is denoted

$$\bar{\Delta}(x) = \mathsf{E}[\tau_{k+1} - \tau_k | \mathcal{F}_k;\, X_k = x] \tag{15}$$

In applications to distributed control, implicit in assumption (A2) is that the population is homogeneous. If each individual is modeled as the TCL ODE (5), this requires that the parameters (α, β) are common across the population, and for N loads we require the stochastic processes $\{\Theta^{a,i}, W_t^i : 1 \leq i \leq N\}$ to be identically distributed. Additional assumptions are required to obtain a mean field limit, such as independence, but this isn't required in this paper since all of our analysis concerns the mean field limit. Relaxation of homogeneity is surely possible by borrowing techniques from analysis of the coupled oscillator model of Kuramoto [30].

The semi-Markov property holds for the SDE model,

$$d\Theta_t = -[\alpha\Theta_t + \beta m_t]dt + d\Theta_t^a + dW_t$$

in which the joint process (Θ_t^a, dW_t) is a deterministic process plus Brownian motion. Such strong statistical assumptions are not necessary.

III. Optimal Control

The goal of KLQ optimal control is a form of *feed-forward control*: we have a reference signal $r\colon \mathbb{R}_+ \to \mathbb{R}$, and our goal is to obtain an open-loop control strategy so that the power approximately tracks the reference signal. This is formalized through the introduction of the conditional mean error sequence defined in (16) below. In power systems language, this may be part of a day-ahead scheduling problem (part of economic dispatch). Alternatively, it may be part of a model predictive control (MPC) architecture, in which case the mean of τ_K may be less than one hour.

A. Kullback-Leibler Quadratic objective and solution

Recall the KLQ optimal control problem (2). The minimum is over all pmfs on X^{K+1}, subject to the constraint that its first marginal is ν_0. The optimizer p^* will then define a randomized policy of the form (9) via Bayes' rule. That is, from the optimal pmf p^* we define ϕ_{k+1}^* by the conditional probability $p^*(u_{k+1}|x_0^k)$, $k \geq 1$. In a distributed control architecture, each agent will apply the same policy based on local observations of x_0^k.

The nominal model defined by the pmf $p^\circ$is Markovian, in which the transition matrix $P_\circ$ appearing in (1) is assumed of the form (12), with stationary Markov policy denoted ϕ°. An approach to constructing the nominal policy is described in the Appendix.

It remains to define $\{\varepsilon_k\}$. For this we choose the conditional mean tracking error over the kth sampling interval:

$$\begin{aligned} \varepsilon_k(x_0^k) &:= \mathcal{U}(x_k) - \mathcal{R}_k(x_0^k) \\ &= \mathsf{E}\Big[\int_{\tau_k}^{\tau_{k+1}} \{\beta m_t - r_t\}\, dt \mid X_0^k = x_0^k\Big] \end{aligned} \tag{16a}$$

$$\begin{aligned} \textit{where}\quad & \mathcal{U}(x_k) = \beta u_k \bar{\Delta}(x_k) \textit{ for } x_k = (s_k, u_k), \\ & \mathcal{R}_k(x_0^k) := \Big[\int_{\tau_k}^{\tau_{k+1}} r_t\, dt \mid X_0^k = x_0^k\Big] \end{aligned} \tag{16b}$$

The expectations in (16) have two interpretations: they hold for an individual load, and also for a mean-field limit when each load uses the same policy.

The KLQ objective is designed to balance two objectives: $|\langle p, \varepsilon_k\rangle|$ should be small for each k (small tracking error), and also $p \approx p^\circ$ (low control cost). A *deadbeat* control solution ignores the second objective, resulting in $\langle p, \varepsilon_k\rangle = 0$ for each k, provided this is feasible.

The following result and Prop. 3.2 that follows are minor extensions of [10, Proposition 3.1]. The results presented here differ because of the more exotic cost structure involving $\{\varepsilon_k\}$, which arises from the new Markovian model.

Proposition 3.1: For each $\kappa > 0$ there is a vector $\lambda^* \in \mathbb{R}^K$ such that the unique optimizer p^* is of the form

$$p^*(x) = p^\circ(x) \exp\Big(\sum_{k=1}^{K} \lambda^*_k \varepsilon_k(x_0^k) - \Gamma^*(x_0)\Big)\,, \tag{17a}$$

in which λ^*, Γ^* and J^* are characterized in the following:

(i) The vector $\lambda^* \in \mathbb{R}^K$ is the solution to

$$\langle p^*, \varepsilon_k\rangle = -\tfrac{1}{\kappa}\lambda^*_k\,, \qquad 1 \le k \le K \tag{17b}$$

(ii) Γ^* is the normalizing constant:

$$\Gamma^*(x_0) = \log \mathsf{E}_\circ\Big[\exp\Big(\sum_{k=1}^{K} \lambda^*_k \varepsilon_k(X_0^k)\Big)\,\big|X_0 = x_0\Big] \tag{17c}$$

(iii) The value function (2) may be expressed

$$J^*(\nu_0) = -\langle \nu_0, \Gamma^*\rangle - \tfrac{1}{2\kappa}\|\lambda^*\|^2 \tag{17d}$$

The proposition raises many questions. First, (17d) suggests that $J^*(\nu_0)$ depends linearly on ν_0. This is *not the case* since λ^* depends on ν_0, as seen in the fixed point equation (17b). This fixed point equation has a solution, since it is the stationary point equation for a convex optimization problem. This is explained below (18), followed by representations of p^* and ϕ^*_{k+1} in several special cases.

B. Largrangian relaxations

The optimization problem (2) is alternatively expressed as follows, with auxiliary variable $\gamma \in \mathbb{R}^K$:

$$J^*(\nu_0) = \min_p \{D(p\|p^\circ) + \frac{\kappa}{2}\|\gamma\|^2\} \quad \text{s.t. } \gamma_k = \langle\, p, \varepsilon_k\,\rangle\,,$$

for $1 \le k \le K$. This is regarded as the *primal*. Letting λ_k denote the Lagrange multiplier for the kth constraint, we arrive at the following Lagrangian $\mathcal{L}$ and dual function φ^*:

$$\begin{aligned} \mathcal{L}(p,\gamma,\lambda) = D(p\|p^\circ) &+ \frac{\kappa}{2}\|\gamma\|^2 \\ &+ \sum_{k=1}^{K} \lambda_k[\gamma_k - \langle\, p, \varepsilon_k\,\rangle] \end{aligned} \tag{18a}$$

$$\varphi^*(\lambda) = \min_{p,\gamma} \mathcal{L}(p,\gamma,\lambda)\,, \qquad \lambda \in \mathbb{R}^K\,, \tag{18b}$$

where the minimum in (18b) is over all pmfs p with first marginal ν_0, and all $\gamma \in \mathbb{R}^K$.

Prop. 3.2 that follows implies that (17b) is the first-order condition for optimality of the dual. The result is a minor extension of Prop. 3.1 and Lemma 3.3 of [10].

Proposition 3.2: The dual function φ^* is concave and coercive, with partial derivatives

$$\frac{\partial}{\partial\lambda_k}\varphi^*(\lambda) = -\frac{1}{\kappa}\lambda_k - \langle\, p^\lambda, \varepsilon_k\,\rangle \tag{19a}$$

$$\frac{\partial}{\partial\lambda_j}\frac{\partial}{\partial\lambda_k}\varphi^*(\lambda) = -\frac{1}{\kappa}\mathbb{I}\{k=j\} - \Sigma^\lambda_\varepsilon(i,j) \tag{19b}$$

in which p^λ is a pmf of the form (17a) with λ^* replaced by λ, and $\Sigma^\lambda_\varepsilon$ is the covariance matrix of $(\varepsilon_1,\dots,\varepsilon_K)$ under p^λ.

C. Markovian realizations

The pmf p^* is not Markovian in general, but we obtain a Markovian realization in several special cases.

Deterministic model In the deterministic TCL model we have $\tau_{k+1} - \tau_k = \bar\Delta(X_k)$ for each k. In this special case we obtain a Markovian solution by expanding the state process:

Proposition 3.3: If the TCL model is deterministic, then

$$\mathcal{R}_k(X_0^k) = \mathcal{R}^d_k(\Phi_k)$$

for functions $\mathcal{R}^d_k\colon \mathsf{X}\times\mathbb{R}_+ \to \mathbb{R}$, in which $\Phi_k = (X_k, \tau_k)$. In this case $\{\Phi_k\}$ is Markovian under p^*, and the optimal policy can be expressed in the form

$$\mathsf{P}\{U^*_{k+1} = u \mid \mathcal{F}_k\} = \phi^*_{k+1}(u|X^*_k, \tau_k)\,, \quad k \ge 0 \tag{20}$$

The representation of the optimal policy simplifies further when the error function $\varepsilon_k(x_0^k)$ depends only on the current state, of the form

$$\mathcal{E}_k(x_k) = \mathcal{U}(x_k) - R_k(x_k)\,, \qquad 1 \le k \le K\,, \tag{21}$$

where $\mathcal{U}$ is defined in (16b).

Proposition 3.4: Suppose that $\varepsilon_k(x_0^k) = \mathcal{E}_k(x_k)$ for each k. Then, the optimal pmf (17a) is Markovian,

$$\begin{aligned} p^*(x_0,\dots,x_K) &= \nu_0(x_0)\prod_{k=0}^{K-1}\check P_k(x_k,x_{k+1}) \\ \textit{with}\quad \check P_{k-1}(x,x') &= \frac{h_k(x')}{h_{k-1}(x)}P_\circ(x,x')e^{\lambda^*_k\mathcal{E}_k(x')} \end{aligned} \tag{22a}$$

in which $h_K \equiv 1$, and $\{h_k : k < K\}$ are defined recursively:

$$h_{k-1}(x) = \sum_{x'} P_\circ(x,x')e^{\lambda^*_k\mathcal{E}_k(x')}h_k(x')\,, \qquad x \in \mathsf{X}\,.$$

The policy is Markovian: with $x' = (\mathsf{s}(x), u')$,

$$\phi^*_{k+1}(u'|x) = e^{\lambda^*_{k+1}\mathcal{E}_{k+1}(x')}\frac{h_{k+1}(x')}{h_k(x)}\phi^\circ(u'|x) \tag{22b}$$

See [10] for the proof of a similar result, subject to the assumption that $\phi^\circ(u'|x)$ depends on x only through $\mathsf{s}(x)$, resulting in a slightly simpler representation.

Constant reference signal. The assumptions of Prop. 3.4 hold when the reference signal is independent of time, even for the non-deterministic TCL model:

Proposition 3.5: If $r_t = r_0$ for all $t \ge 0$, then the assumptions of Prop. 3.4 hold with $R_k(x_k) = r_0\bar\Delta(x_k)$ and $\bar\Delta$ defined in (15). ∎

Sampling design. Returning to the deterministic TCL model, the assumptions of Prop. 3.4 hold provided S is chosen so that $\bar\Delta(X_k) = \tau_{k+1} - \tau_k$ does not depend upon X_k (see (15)). This conclusion is generalized in the following.

Proposition 3.6: Suppose that $\{\widetilde\Delta_{k+1} := \tau_{k+1} - \tau_k : k \ge 0\}$ are independent and identically distributed, with distribution independent of X_0. Then, the assumptions of Prop. 3.4 hold with $R_k(x_k) = \mathsf{E}\big[\int_{\tau_k}^{\tau_{k+1}} r_t\, dt\big]$.

The assumptions in any of Propositions 3.3, 3.5, or 3.6 will never hold exactly. They do suggest approximations of an optimal policy in certain regimes.

IV. Monte-Carlo Methods

The fixed point equation (17b) motivates practical numerical techniques to obtain the Lagrange multiplier λ^* that determines p^* via (17a).

The proof of [10, Proposition 3.1] can be extended to establish the form of the pmf solving (18b):

$$p^{\lambda}(x) = p^{\circ}(x) \exp\Bigl(\sum \lambda_k \varepsilon_k(x_0^k) - \Gamma^{\lambda}(x_0)\Bigr) \tag{23}$$

where again $\Gamma^{\lambda}(x_0)$ is a normalized constant for each $x_0 \in \mathsf{X}$.

We present the algorithm for a simplified objective function, in which the error $\varepsilon_k(x_0^k)$ is of the form (21), ensuring that p^{λ} is Markovian for any λ. The term $R_k(x_k)$ is chosen to approximate

$$\mathcal{R}_k^p(x_k) := \mathsf{E}\Bigl[\int_{\tau_k}^{\tau_{k+1}} r_t \, dt \mid X_k = x_k\Bigr] \tag{24}$$

Outside of very special cases, such as the setting of Prop. 3.6, this conditional expectation depends on the pmf p. In Sec. IV-A we present a stochastic gradient algorithm that can be applied when the functions $\{R_k\}$ are given. For example, $R_k(x_k)$ taken equal to (24) under p°, which is reasonable when the reference signal is not large. An algorithm to estimate both $\{\mathcal{R}_k^p\}$ and λ^* is presented in Sec. IV-C.

It is straightforward to extend the algorithm of Sec. IV-A and Prop. 4.1 to the general case in which this constraint on $\varepsilon_k(x_0^k)$ is relaxed. The algorithm is more complex since computation of the pmf (23) is more complex in this case.

A. Stochastic gradient ascent

The goal is to obtain a recursive algorithm generating estimates $\{\lambda^n : n \ge 0\}$ that converges to λ^* a.s. from each initial $\lambda^0 \in \mathbb{R}^K$. To simplify notation we use p^n to denote p^{λ} with $\lambda = \lambda^n$. The pmf p^n is Markovian, obtained as in Prop. 3.4: there is a collection of functions $\{h_k^n\}$ on X such that for each $x = (s, u) \in \mathsf{X}$ and $x' = (\mathsf{s}(x), u')$,

$$\phi_{k+1}^n(u'|x) = \frac{h_{k+1}^n(x')}{h_k^n(x)} e^{\lambda_{k+1}^n \mathcal{E}_{k+1}(x')} \phi^{\circ}(u'|x) \,.$$

The stochastic gradient ascent (SGA) algorithm described here is designed to approximate gradient ascent,

$$\begin{aligned} \lambda_k^{n+1} &= \lambda_k^n + \alpha\kappa \frac{\partial}{\partial \lambda_k} \varphi^*(\lambda^n) \\ \kappa \frac{\partial}{\partial \lambda_k} \varphi^*(\lambda^n) &= -\lambda_k^n - \kappa \langle p^n, \mathcal{E}_k \rangle \,, \quad 1 \le k \le K \end{aligned} \tag{25}$$

where the second equation follows from (19a) (with ε_k replaced by $\mathcal{E}_k$), and $\alpha > 0$ is the step-size.

The SGA recursion takes the form,

$$\lambda_k^{n+1} = \lambda_k^n + \alpha_{n+1} \widetilde{\nabla}_k^{n+1} \,, \quad 1 \le k \le K \tag{26}$$

with $\{\alpha_{n+1}\}$ a non-negative step-size sequence, and $\widetilde{\nabla}_k^{n+1}$ is an unbiased estimate of the scaled gradient $\kappa \nabla \varphi^*(\lambda^n)$. Even subject to the simplification obtained by replacing $\{\varepsilon_k\}$ by $\{\mathcal{E}_k\}$ in the KLQ objective, constructing p^n is complex if $K \times n_{\mathsf{S}}$ is large. For this reason, for each n we draw many samples from the distribution p^n to approximate this gradient.

Given λ^n, the random vector $\widetilde{\nabla}_k^{n+1}$ is constructed via the following steps:

1. Generate data. Obtain N independent trajectories, begining with the initialization, $X_0^{n+1,i} \sim \nu_0$, $1 \le i \le N$.

Subsequent states are drawn independently and sequentially: given $\{x^i := X_{k-1}^{n+1,i} : 1 \le i \le N\}$, obtain $S_k^{n+1,i} = \mathsf{s}(x^i)$ and $U_k^{n+1,i} = 1$ with probability $\phi_{k+1}^n(\,\cdot\,|x^i)$, which gives $X_k^{n+1,i} = (S_k^{n+1,i}, U_k^{n+1,i})$ for each i.

The approach in Sec. IV-C also requires the sampling times, $\tau_{k+1}^{n+1,i} = \tau_k^{n+1,i} + \widetilde{\Delta}_{k+1}^{n+1,i}$. When using a simulator based on the semi-Markov model, then $\widetilde{\Delta}_{k+1}^{n+1,i} = \tau_{k+1}^{n+1,i} - \tau_k^{n+1,i}$ is drawn from the CDF in (14), using $x = X_k^{n+1,i}$. These random variables could also be obtained through data collected from appliances, or a high-fidelity simulator.

2. Gradient approximation.

$$\widetilde{\nabla}_k^{n+1} = -\lambda_k^n - \kappa \frac{1}{N} \sum_{i=1}^N \mathcal{E}_k(X_k^{n+1,i}) \,, \quad 1 \le k \le K \,. \tag{27}$$

The K-dimensional random vector $\widetilde{\nabla}^{n+1}$ has mean $\nabla \varphi^*(\lambda^n)$, so that the update (26) is the desired stochastic approximation algorithm based on (25).

The recursion (26) is convergent under standard assumptions on the step-size sequence [6]. In the following we consider the special case $\alpha_n = n^{-\varrho}$ for $\varrho > 0$ and apply Polyak-Ruppert averaging: fix $m_0 \ge 0$ and define recursively,

$$\bar{\lambda}^{m+1} = \bar{\lambda}^m + \delta_{m+1}\{-\bar{\lambda}^m + \lambda^{m+1}\} \,, \quad m \ge m_0 \,, \tag{28}$$

with $\delta_{m+1} = {}^{1}\!/\!{}_{(m-m_0+1)}$, and $\bar{\lambda}^{m_0} = 0$.

B. Rates of convergence

It is well known that the CLT holds for $\{\bar{\lambda}^m\}$ with minimal asymptotic covariance, but we are not aware of results establishing convergence of the second moments. The mean-square convergence rate of the estimates $\{\bar{\lambda}^m : m > m_0\}$ is established here by applying recent techniques from the stochastic approximation (SA) literature.

It will be convenient to begin with abstract notation, in which the estimates $\{\lambda^n : n \ge 0\}$ are expressed as the output of an SA recursion in two forms,

$$\begin{aligned} \lambda^{n+1} &= \lambda^n + \alpha_{n+1} f(\lambda^n, Z_{n+1}) \\ &= \lambda^n + \alpha_{n+1} \bigl[\bar{f}(\lambda^n) + D^{n+1}\bigr] \end{aligned} \tag{29}$$

In the first, the sequence $\{Z_{n+1}\}$ is i.i.d. (reflecting the independent sampling to obtain $\{X_k^{n+1,i}\}$), and in the second the sequence $\{D^{n+1}\}$ is a vector-valued martingale difference sequence. We have $\bar{f}(\lambda) = \mathsf{E}[f(\lambda, Z_{n+1})]$ for any λ (the expectation is independent of n).

Let $\Sigma_D = \mathrm{Cov}\bigl(f(\lambda^*, Z_{n+1})\bigr)$, and $A = \partial \bar{f}(\lambda^*)$. The matrix $G = -A^{-1}$ is the matrix gain in Ruppert's stochastic Newton Raphson algorithm [24], and the Polyak-Ruppert covariance matrix is $\Sigma_\theta^* = G \Sigma_D G^{\intercal}$, subject to the assumption that A is Hurwitz [4], [25], [22], [23].

One conclusion of Prop. 4.1 is that the matrices A and Σ_D may be expressed in terms of $\Sigma_{\mathcal{E}}$, the covariance of the random vector $(\mathcal{E}_1(X_1); \mathcal{E}_2(X_2); \cdots ; \mathcal{E}_K(X_K))$ under the optimal pmf p^*.

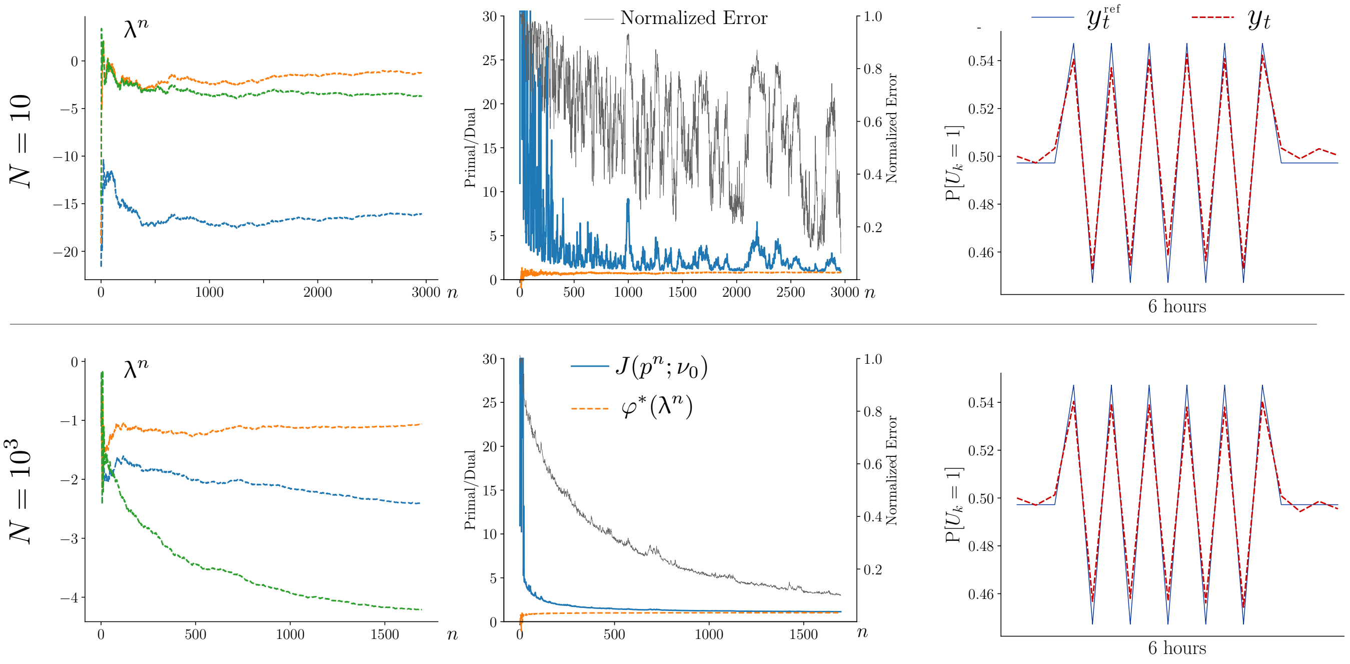

Fig. 1: SGA without averaging, for two values of N in (30). Column 1: evolution of $\{\lambda_i^n : n \ge 0, i = 1, 2, 3\}$ (values of i are blue, orange, green respectively). Column 2: evolution of the primal and dual functions, and the normalized error. Column 3: tracking performance using the final value of λ^n.

Proposition 4.1: Using step-size $\alpha_n = n^{-\varrho}$, with $\varrho \in ({}^1\!/_2, 1)$, the estimates $\{\lambda^{n+1}\}$ converge to λ^* as $n \to \infty$. The PR-estimates converge with optimal mean-square convergence rate: for any fixed $m_0 \ge 1$ the limit (3) holds, with Σ_θ^* given in (4) identical to the Polyak-Ruppert covariance matrix.

Proof The update equation (26) may be expressed as the SA recursion (29) in which $\bar{f}(\lambda) = \nabla\varphi^*(\lambda)$ and $\{D^{n+1}\}$ is a K-dimensional martingale difference sequence. This follows from independent sampling to obtain (27), which results in

$$\widetilde{\nabla}_k^{n+1} = \kappa \frac{\partial}{\partial \lambda_k} \varphi^*(\lambda^n) + D_k^{n+1}$$

$$\textit{with} \quad D_k^{n+1} = \frac{\kappa}{N} \sum_{i=1}^N \widetilde{\mathcal{E}}_k(X_k^{n+1,i})$$

in which $\widetilde{\mathcal{E}}_k(x) = \mathcal{E}_k(x) - \mathsf{E}[\mathcal{E}_k(X_k)]$, with expectation taken under p^n. Prop. A.1 combined with Prop. 3.2 completes the proof, on recognizing that $A = \kappa\nabla^2\varphi^*(\lambda^*) = -I - \kappa\Sigma_{\mathcal{E}}$.

The expression for $\{D_k^{n+1}\}$ gives the steady-state covariance formula $\Sigma_D = \frac{\kappa^2}{N}\Sigma_{\mathcal{E}}$. ■

C. Concurrent estimation of mean reference signal

Suppose that the expectation (24) is not dependent on the policy. While this is unlikely to hold exactly, as discussed earlier it is a reasonable approximation in some settings.

Two changes are required in the SGA algorithm. First, we require the additional data,

$$\widehat{R}_k^{n+1,i} = \int r_t \, dt$$

where the integral is over $[\tau_k^{n+1,i}, \tau_{k+1}^{n+1,i}]$. The gradient approximation is then modified as follows:

$$\widetilde{\nabla}_k^{n+1} = -\lambda_k^n - \kappa \frac{1}{N} \sum_{i=1}^N \bigl(\mathcal{U}_k(X_k^{n+1,i}) - \widehat{R}_k^{n+1,i} \bigr) \tag{30}$$

With this modification the recursion (26) can be applied to approximate λ^*.

However, an additional step is required to draw samples in the first step of SGA, since we must estimate the functions $\{\mathcal{E}_k : 1 \le k \le K\}$. Given the definition $\mathcal{E}_k(x_k) := \mathcal{U}(x_k) - R_k(x_k)$, it remains to estimate the second term. Introduce the unbiased estimates $\{\mathcal{R}_k^{n+1}\}$ of (24) via

$$R_k^{n+1}(x) = \frac{1}{N_x^{n+1}} \sum_{i=1}^N \mathbb{I}\{X_k^{n+1,i} = x\} \widehat{R}_k^{n+1,i}$$

$$N_x^{n+1} = \sum_{i=1}^N \mathbb{I}\{X_k^{n+1,i} = x\}$$

To ensure consistency these estimates must be averaged. Obtain $\{\mathcal{E}_k^n : 1 \le k \le K\}$ recursively via,

$$\mathcal{E}_k^{n+1}(x) = \mathcal{E}_k^n(x) + \beta_{n+1}\{-\mathcal{E}_k^n(x) + \mathcal{U}(x) - R_k^{n+1}(x)\}$$

We then define p^{n+1} using this estimate:

$$p^{n+1}(x) = p^\circ(x) \exp\Bigl(\sum \lambda_k^{n+1} \mathcal{E}_k^{n+1}(x_k) - \Gamma^{n+1}(x_0) \Bigr)$$

To ensure that $\mathcal{E}_k^{n+1}(x) \approx \mathcal{R}_k^p(x_k)$ in (24), with $p = p^n$, requires a two-time scale algorithm in which ${}^{\beta_n}\!/_{\alpha_n} \to \infty$ as $n \to \infty$; e.g., $\alpha_n = n^{-\varrho}$ and $\beta_n = n^{-\varrho'}$, $\frac{1}{2} < \varrho' < \varrho < 1$.

D. Example

The SGA algorithm was applied to control a fleet of homogeneous refrigerators modeled according to (5), such that their aggregate power consumption approximately tracks a reference signal while maintaining all temperatures within a finite interval $[\theta^{\min}, \theta^{\max}]$. This motivates the constraint $\mathsf{S} \subset [\theta^{\min}, \theta^{\max}]$, containing both $\theta^{\max}$ and $\theta^{\min}$.

The set is expressed $\mathsf{S} = \mathsf{S}_+ \cup \mathsf{S}_-$, in which the union need not be disjoint, and the mapping s is defined as follows: for each $\theta \in \mathsf{S}$,

$$\mathsf{s}(x) = \begin{cases} \arg\max\{\theta_+ \in \mathsf{S}_+ : \theta_+ < \theta\} & x = (\theta, 1) \\ \arg\min\{\theta_- \in \mathsf{S}_- : \theta_- > \theta\} & x = (\theta, 0) \end{cases}$$

We are not assuming that $\Theta_t \in [\theta^{\min}, \theta^{\max}]$ *for all* t, but by design this constraint is satisfied for $t = \tau_k$ when $k \geq 1$. It follows by construction that $m_t = 1$ whenever $\Theta_t \geq \theta^{\max}$, and $m_t = 0$ whenever $\Theta_t \leq \theta^{\min}$.

The numerical results that follow are based on models considered in [11], and the TCL model was a typical refrigerator model from [18]. The set S was obtained based on consideration of the deterministic ODE obtained from (5) with $\boldsymbol{W} \equiv 0$. The values in S were selected so that $\bar{\Delta}(x)$ is approximately independent of $x \in \mathsf{X}$. It was found that the value $|\mathsf{S}| = 36$ was small enough to ensure feasibility of tracking for the reference signals considered.

The nominal model defined by ϕ° was designed to approximate deterministic hysteresis control—see [11] for details its construction.

We display normalized data: $y^{\text{ref}}_t = r_t/\beta$ and $y_t = \mathsf{E}[m_t]$ (the probability of a refrigerator being on, estimated via Monte-Carlo). The plots shown in Fig. 1 demonstrate the results of two numerical experiments, identical except for the choice of N. The left column contains plots displaying the evolution of $\{\lambda^n_i : n \geq 0\}$ for selected values of i. The middle column contains plots displaying the evolution of the primal and dual functions, and the normalized error, defined as their difference divided by the value of the primal. Notice how these trajectories are much more volatile when N is small, as expected.

The right column contains plots displaying the tracking performance using the final value of λ^n. The value of κ was chosen to achieve satisfactory tracking of this feasible reference signal.

V. Conclusions

The SGA algorithm proposed in this paper is simple and easily analyzed, as seen by the explicit expression for the asymptotic covariance in (4). This simplicity is a product of the simple model obtained from event triggered sampling—without (8), representations of the KLQ solution are far more complex [8], [11].

Observe that a SGQ algorithm does not require estimation of $\varepsilon_k(x^k_0)$, defined in (16a). The formula (19a) and the smoothing property of conditional expectations gives,

$$\frac{\partial}{\partial \lambda_k}\varphi^*(\lambda) = -\frac{1}{\kappa}\lambda_k - \mathsf{E}\Big[\mathcal{U}(X_k) - \int_{\tau_k}^{\tau_{k+1}} \{\beta m_t - r_t\}\, dt\Big]$$

where the expectation is under p^λ. The simplifications in Sec. IV were imposed to simplify both constructing p^λ and sampling from this pmf.

Topics of current interest include: combining PR averaging with other acceleration techniques; error bounds for the approach posed in Sec. IV-C, and conducting numerical experiments to gain insight; the construction of useful finite-n error bounds for SGA (tractable since the noise is martingale difference in (29)), complementing the asymptotic theory.

Appendix I
Moment Bounds for Stochastic Approximation

The PR averaged estimate $\bar{\lambda}^m$ obtained from (28) is in fact the average of $\{\lambda^n\}$. This fact is used in our analysis of (29), along with the scaled average of the disturbance:

$$\bar{\lambda}^m = \tfrac{1}{m-m_0}\sum_{n=m_0+1}^{m} \lambda^n\,, \quad W_m := \tfrac{1}{\sqrt{m-m_0}}\sum_{n=m_0+1}^{m} D^{n+1} \qquad (31)$$

Under the assumptions of Prop. A.1, the distribution of W_m is approximately Gaussian $N(0, \Sigma_D)$ when $m \gg m_0$.

Proposition A.1: Consider the general K-dimensional SA algorithm (29), with step-size $\alpha_n = n^{-\varrho}$, $\varrho \in ({}^1\!/\!_2, 1)$, in which $\boldsymbol{Z}$ is i.i.d. on a finite state space.

Suppose moreover that $f(\,\cdot\,, z)$ is globally Lipschitz continuous for each z, so that $\bar{f}$ is also Lipschitz; it is continuously differentiable in a neighborhood of the unique root λ^*, and that $A = \partial \bar{f}\,(\lambda^*)$ is Hurwitz. Finally, assume that the ODE $\frac{d}{dt}x = \bar{f}(x)$ is exponentially asymptotically stable.

Then there exists a finite constant B such that $\mathsf{E}[\|\lambda^n - \lambda^*\|^4] \leq B\alpha_n^2$ for each n, and the limit (4) holds for the averaged estimates $\{\bar{\lambda}^m\}$.

Proof Exponentially asymptotically stability implies that an "ODE@∞" has the same property [29]. This is one key assumption in [5, Thm. 3.6] to obtain the bound $\mathsf{E}[\|\lambda^n - \lambda^*\|^4] \leq B\alpha_n^2$ for a fixed constant B.

It remains to establish (4). For this, write

$$\bar{f}(\lambda^n) = A[\lambda^n - \lambda^*] + \mathcal{E}(\lambda_n)$$

in which the error term satisfies $\mathsf{E}[\|\mathcal{E}(\lambda^n)\|^2] \leq B_f \alpha_n^2$ under the given assumptions.

Denote $\widetilde{\lambda}^n := \lambda^n - \lambda^*$. Subtracting λ^* from each side of (29), dividing each side by α_{n+1}, and rearranging terms gives

$$\tfrac{1}{\alpha_{n+1}}\widetilde{\lambda}^{n+1} - \tfrac{1}{\alpha_n}\widetilde{\lambda}^n = A\widetilde{\lambda}^n + D^{n+1} + \mathcal{E}(\lambda_n) + \gamma_n\widetilde{\lambda}^n$$

with $\gamma_n = {}^1\!/\!_{\alpha_{n+1}} - {}^1\!/\!_{\alpha_n} \leq \rho n^{\rho-1}$. Averaging each side and applying (31),

$$\begin{aligned}\tfrac{1}{\alpha_m}\widetilde{\lambda}^m - \tfrac{1}{\alpha_{m_0+1}}\widetilde{\lambda}^{m_0+1} &= A(\bar{\lambda}^m - \lambda^*) + \tfrac{1}{\sqrt{m-m_0}}W_m \\ &\quad + \tfrac{1}{m-m_0}\sum_{n=m_0+1}^{m}\big(\mathcal{E}(\lambda_n) + \gamma_n\widetilde{\lambda}^n\big)\end{aligned}$$

Written in the more suggestive form,

$$\sqrt{m-m_0}\,(\bar{\lambda}^m - \lambda^*) = -A^{-1}W_m + \mathcal{E}'_m$$

it follows from the previous bounds that $\mathsf{E}[\|\mathcal{E}'_m\|^2] \to 0$ as $m \to \infty$, which implies (4). ∎

References

[1] M. Almassalkhi, J. Frolik, and P. Hines. Packetized energy management: asynchronous and anonymous coordination of thermostatically controlled loads. In *Proc. of the American Control Conf.*, pages 1431–1437, 2017.

[2] J. Bas Serrano, S. Curi, A. Krause, and G. Neu. Logistic Q-learning. In A. Banerjee and K. Fukumizu, editors, *Proc. of The Intl. Conference on Artificial Intelligence and Statistics*, volume 130, pages 3610–3618, 13–15 Apr 2021.

[3] E. Benenati, M. Colombino, and E. Dall'Anese. A tractable formulation for multi-period linearized optimal power flow in presence of thermostatically controlled loads. In *IEEE Conference on Decision and Control*, pages 4189–4194. IEEE, 2019.

[4] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, Berlin Heidelberg, 2012.

[5] V. Borkar, S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. The ODE method for asymptotic statistics in stochastic approximation and reinforcement learning. *arXiv e-prints:2110.14427*, pages 1–50, 2021.

[6] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, Delhi, India, 2nd edition, 2021.

[7] A. Bušić and S. Meyn. Distributed control of thermostatically controlled loads: Kullback-Leibler optimal control in continuous time. In *Proc. of the Conf. on Dec. and Control*, pages 7258–7265, Dec 2019.

[8] A. Bušić and S. Meyn. Ordinary Differential Equation Methods for Markov Decision Processes and Application to Kullback–Leibler Control Cost. *SIAM J. Control Optim.*, 56(1):343–366, 2018.

[9] P. E. Caines. Mean field games. In J. Baillieul and T. Samad, editors, *Encyclopedia of Systems and Control*, pages 1197–1202. Springer London, London, 2021.

[10] N. Cammardella, A. Bušić, Y. Ji, and S. Meyn. Kullback-Leibler-Quadratic optimal control of flexible power demand. In *Proc. of the Conf. on Dec. and Control*, pages 4195–4201, Dec. 2019.

[11] N. Cammardella, A. Bušić, and S. Meyn. Kullback-Leibler-quadratic optimal control. *SIAM Journal on Control and Optimization*, page arXiv:2004.01798, April 2023.

[12] Y. Chen, M. U. Hashmi, J. Mathias, A. Bušić, and S. Meyn. Distributed control design for balancing the grid using flexible loads. In S. Meyn, T. Samad, I. Hiskens, and J. Stoustrup, editors, *Energy Markets and Responsive Grids: Modeling, Control, and Optimization*, pages 383–411. Springer, New York, NY, 2018.

[13] M. Chertkov and V. Y. Chernyak. Ensemble control of cycling energy loads: Markov Decision Approach. In *IMA volume on the control of energy markets and grids*. Springer, 2018.

[14] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[15] P. Guan, M. Raginsky, and R. Willett. Online Markov decision processes with Kullback-Leibler control cost. *IEEE Trans. Automat. Control*, 59(6):1423–1438, June 2014.

[16] M. Kárný. Towards fully probabilistic control design. *Automatica*, 32(12):1719 –1722, 1996.

[17] R. Malhame and C.-Y. Chong. Electric load model synthesis by diffusion approximation of a high-order hybrid-state stochastic system. *IEEE Transactions on Automatic Control*, 30(9):854–860, Sep. 1985.

[18] J. Mathieu. *Modeling, Analysis, and Control of Demand Response Resources*. PhD thesis, University of California at Berkeley, 2012.

[19] J. Mathieu, S. Koch, and D. Callaway. State estimation and control of electric loads to manage real-time energy imbalance. *IEEE Trans. Power Systems*, 28(1):430–440, 2013.

[20] P. Mehta and S. Meyn. A feedback particle filter-based approach to optimal control with partial observations. In *Proc. of the Conf. on Dec. and Control*, pages 3121–3127, Dec 2013.

[21] S. Meyn, P. Barooah, A. Bušić, Y. Chen, and J. Ehren. Ancillary service to the grid using intelligent deferrable loads. *IEEE Trans. Automat. Control*, 60(11):2847–2862, Nov 2015.

[22] B. T. Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika (in Russian). translated in Automat. Remote Control, 51 (1991)*, pages 98–107, 1990.

[23] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

[24] D. Ruppert. A Newton-Raphson version of the multivariate Robbins-Monro procedure. *The Annals of Statistics*, 13(1):236–245, 1985.

[25] D. Ruppert. Efficient estimators from a slowly convergent Robbins-Monro processes. Technical Report Tech. Rept. No. 781, Cornell University, School of Operations Research and Industrial Engineering, Ithaca, NY, 1988.

[26] A. Taghvaei and P. G. Mehta. A survey of feedback particle filter and related controlled interacting particle systems. *arXiv preprint arXiv:2301.00935*, 2023.

[27] S. H. Tindemans, V. Trovato, and G. Strbac. Decentralized control of thermostatic loads for flexible demand response. *IEEE Transactions on Control Systems Technology*, 23(5):1685–1700, Sept 2015.

[28] E. Todorov. Linearly-solvable Markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Proc. Advances in Neural Information Processing Systems*, pages 1369–1376, Cambridge, MA, 2007.

[29] M. Vidyasagar. Convergence of stochastic approximation via martingale and converse Lyapunov methods. *Mathematics of Control, Signals, and Systems*, pages 1–24, 2023.

[30] H. Yin, P. Mehta, S. Meyn, and U. Shanbhag. Learning in mean-field games. *IEEE Trans. Automat. Control*, 59(3):629–644, March 2014.