SIAM J. CONTROL OPTIM.
Vol. 61, No. 5, pp. 3234–3258
© 2023 Society for Industrial and Applied Mathematics

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

KULLBACK–LEIBLER-QUADRATIC OPTIMAL CONTROL*

NEIL CAMMARDELLA†, ANA BUŠIĆ‡, AND SEAN P. MEYN§

Abstract. This paper presents approaches to mean-field control, motivated by distributed control of multiagent systems. Control solutions are based on a convex optimization problem, whose domain is a convex set of probability mass functions (pmfs). The main contributions follow: (1) Kullback–Leibler-quadratic (KLQ) optimal control is a special case in which the objective function is composed of a control cost in the form of Kullback–Leibler divergence between a candidate pmf and the nominal, plus a quadratic cost on the sequence of marginals. Theory in this paper extends prior work on deterministic control systems, establishing that the optimal solution is an exponential tilting of the nominal pmf. Transform techniques are introduced to reduce complexity of the KLQ solution, motivated by the need to consider time horizons that are much longer than the intersampling times required for reliable control. (2) Infinite-horizon KLQ leads to a state feedback control solution with attractive properties. It can be expressed as state feedback, in which the state is the sequence of marginal pmfs, or an open loop solution is obtained that is more easily computed. (3) Numerical experiments are surveyed in an application of distributed control of residential loads to provide grid services, similar to utility-scale battery storage. The results show that KLQ optimal control enables the aggregate power consumption of a collection of flexible loads to track a time-varying reference signal, while simultaneously ensuring each individual load satisfies its own quality of service constraints.

Key words. mean-field games, distributed control, demand dispatch

MSC codes. 90C40, 93E20, 90C46, 93E35, 60J20

DOI. 10.1137/21M1433654

1. Introduction.

The goal of this paper is to obtain control solutions for mean-field models. The optimization problems considered are generalizations of standard Markov decision process (MDP) objectives, in both finite-horizon and average-cost settings.

1.1. Mean-field control.

The mean-field control problem is an approach to distributed control of a collection of $\mathcal{N}$ homogeneous "agents," with $\mathcal{N} \gg 1$, modeled as discrete-time stochastic systems, with state processes at time k denoted $\{X_k^i : 1 \le i \le \mathcal{N}\}$. To avoid a long detour on notation it is assumed that the common state space X is finite. For a single value k and time horizon $K \ge 1$, the empirical distributions are denoted

*Received by the editors July 14, 2021; accepted for publication (in revised form) July 11, 2023; published electronically October 25, 2023.
https://doi.org/10.1137/21M1433654
Funding: The authors received support from National Science Foundation grant EPCN 1935389 and from the France 2030 Plan, PEPR TASE, AI-NRGY project ANR-22-PETA-0044.
†Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (ncammardella@ufl.edu).
‡Inria and DI ENS, École Normale Supérieure, CNRS, PSL Research University, Paris, France (ana.busic@inria.fr, https://www.di.ens.fr/∼busic/).
§Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA, and Inria International Chair, Paris, France (meyn@ece.ufl.edu, http://www.meyn.ece.ufl.edu/).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$p^{\mathcal{N}}(\vec{x}) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \mathbb{I}\{(X_0^i, \ldots, X_K^i) = \vec{x}\}, \qquad \vec{x} \in \mathsf{X}^{K+1}, \tag{1.1a}$$

$$\nu_k^{\mathcal{N}}(x) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \mathbb{I}\{X_k^i = x\}, \qquad x \in \mathsf{X}, \tag{1.1b}$$

where $\vec{x} = (x_0, \ldots, x_K)$ denotes an arbitrary element of X^{K+1}. The set of probability mass functions (pmfs) on X^{K+1} is denoted by $\mathcal{S}(\mathsf{X}^{K+1})$ for $K \geq 1$ and $\mathcal{S}(\mathsf{X})$ for $K = 0$.

The integer $\mathcal{N}$ is regarded as a parameter in mean-field theory, and assumptions imply that there is convergence as $\mathcal{N} \to \infty$,

$$\lim_{\mathcal{N}\to\infty} p^{\mathcal{N}}(\vec{x}) = p(\vec{x}), \qquad \lim_{\mathcal{N}\to\infty} \nu_k^{\mathcal{N}}(x_k) = \nu_k(x),$$

where $\nu_k \in \mathcal{S}(\mathsf{X})$ is the kth marginal of $p \in \mathcal{S}(\mathsf{X}^{K+1})$ for $0 \leq k \leq K$.

In this paper this limit is achieved by assuming homogeneity of the statistics of each agent: for each i the state evolution is consistent with p:

$$\mathsf{P}\{X_{k+1}^i = x_{k+1} \mid (X_0^i, \ldots, X_k^i) = \vec{x}_0^k\} = p(x_{k+1} \mid \vec{x}_0^k), \tag{1.2}$$

where the conditional pmfs are obtained from the Bayes rule.

The paper concerns design of p to balance two objectives, based on a reference signal $\{r_k\}$, and function $\mathcal{Y}\colon \mathsf{X} \to \mathbb{R}$:

(i) $\nu_k \sim \nu_k^0$, where $\{\nu_k^0\}$ models *nominal behavior.*

(ii) $\langle \nu_k, \mathcal{Y} \rangle := \sum_{x\in\mathsf{X}} \nu_k(x)\mathcal{Y}(x) \approx r_k$.

The agents considered in section 4 represent a population of residential water heaters, and $\mathcal{Y}\colon \mathsf{X} \to \mathbb{R}_+$ is chosen so that $\langle \nu_k^{\mathcal{N}}, \mathcal{Y} \rangle$ is the average power consumption over the population of loads.

Two approaches to design are developed in this paper.

Feedforward control. A sequence $\{\mathcal{C}_k : 1 \leq k \leq K\}$ of real-valued cost functions on the marginals is specified, and p^* is obtained as the solution to

$$J^\star(\nu_0^0) = \min_p \sum_{k=1}^{K} \mathcal{C}_k(\nu_k), \tag{1.3}$$

where the minimum is over all pmfs with first marginal ν_0^0. The two goals motivate the objective function

$$\mathcal{C}_k(\nu) = \mathcal{D}(\nu, \nu_k^0) + \frac{\kappa}{2}\big[\langle \nu, \mathcal{Y} \rangle - r_k\big]^2, \qquad \nu \in \mathcal{S}(\mathsf{X}), \tag{1.4}$$

in which $\kappa > 0$ is a penalty parameter, and $\mathcal{D}$ penalizes deviation from nominal behavior. The finite-horizon optimal control problem is thus

$$J^\star(\nu_0^0) = \min_p \sum_{k=1}^{K} \left[\mathcal{D}(\nu_k, \nu_k^0) + \frac{\kappa}{2}\big[\langle \nu_k, \mathcal{Y} \rangle - r_k\big]^2 \right]. \tag{1.5}$$

It is envisioned that this finite horizon optimal control problem will be a component of a model predictive control (MPC) strategy, with time horizons for computation updates dictated by performance requirements and model accuracy.

Feedback control. If the nominal model is Markovian, then the evolution of the marginals follows the dynamics of a controlled nonlinear state space model,

$$\nu_{k+1} = f_k(\nu_k, \phi_k), \qquad k \geq 0, \quad \nu_0^0 \text{ given}, \tag{1.6}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where $\{\phi_k\}$ is the input sequence, evolving on an abstract set Φ. A feedback policy takes the form $\phi_k = \mathcal{K}_k(\nu_k)$.

Design choices for $\mathcal{K}_k$ are proposed based on an infinite-horizon solution of (1.5). Justification requires further assumptions, including time-homogeneous dynamics for (1.6), which holds if the nominal model is a time-homogeneous Markov chain.

1.2. MDPs and mean-field control. The Markovian assumption for the nominal model is based on the standard controlled Markov chain model used in MDPs.

The model considered here is specified by a state space denoted S and input space U, and we denote $\mathsf{X} := \mathsf{S} \times \mathsf{U}$ (assumed finite). The joint state-input process is denoted $\boldsymbol{X} = \{X_k = (S_k, U_k) : k \ge 0\}$. In finite-horizon optimal control the model includes a sequence of controlled transition matrices $\{T_k : k \ge 0\}$ and cost functions $\{c_k : k \ge 0\}$ with $c_k \colon \mathsf{X} \to \mathbb{R}$ for each k.

The dynamics of $\boldsymbol{x} = (\boldsymbol{S}, \boldsymbol{U}) = \{S_k, U_k : k \ge 0\}$ are determined by the transition matrices as follows. It is assumed that $\boldsymbol{X}$ is adapted to a filtration $\{\mathcal{F}_k : k \ge 0\}$ (so that X_k is $\mathcal{F}_k$-measurable for each k), and

$$\mathsf{P}\{S_{k+1} = s' \,|\, \mathcal{F}_k;\ S_k = s\,,\ U_k = u\} = T_k(x, s')\,, \qquad x = (s,u) \in \mathsf{X}, s' \in \mathsf{S}. \tag{1.7}$$

The set of functions from S to the simplex $\mathcal{S}(\mathsf{U})$ is denoted Φ, and we let ϕ denote a generic element of Φ. The decision rule defining the input sequence is assumed to be Markovian: with $\phi_k \in \Phi$ for each k,

$$\mathsf{P}\{U_k = u \,|\, \mathcal{F}_{k-1};\ S_k = s\} = \phi_k(u \,|\, s)\,, \qquad x = (s,u) \in \mathsf{X}. \tag{1.8}$$

The finite-horizon optimal control problem of MDP theory is a special case of (1.3), in which $\mathcal{C}_k$ linear for each k; in this case $\mathcal{C}_k(\nu_k) = \langle \nu_k, c_k \rangle = \sum_{x \in \mathsf{X}} \nu_k(x) c_k(x)$ for each k, and the sum on the right-hand side of (1.3) may be expressed

$$\sum_{k=1}^{K} \langle \nu_k, c_k \rangle = \sum_{k=1}^{K} \mathsf{E}[c_k(X_k)]\,, \qquad X_k \sim \nu_k\,,$$

where $\boldsymbol{X}$ evolves according to the controlled Markovian dynamics. This interpretation is the first step in the linear programming (LP) approach to MDPs introduced by Manne [5, 34]. The second step is to recognize that the dynamics can be expressed as a sequence of linear constraints on the marginals,

$$\sum_{u'} \nu_k(s', u') = \sum_{s,u} \nu_{k-1}(s,u) T_{k-1}(x, s')\,, \quad s' \in \mathsf{S}\,,\ 1 \le k \le K\,, \quad \nu_0^0 \text{ given}. \tag{1.9}$$

Another special case of (1.3) is variance-penalized optimal control, for which $\mathcal{C}_k(\nu_k) = \langle \nu_k, c \rangle + \kappa[\langle \nu_k, c^2 \rangle - \langle \nu_k, c \rangle^2]$, with $\kappa > 0$ a penalty parameter. The solution to the optimization problem (1.3) can be expressed using a randomized state feedback policy of the form (1.8) [2, 41, 36].

1.3. Kullback–Leibler-quadratic control. In this approach to feedforward control we choose a Markovian model of the form (1.7), (1.8) to define nominal behavior: for a collection $\{\phi_k^0\} \subset \Phi$,

$$p^0(\vec{x}) = \nu_0^0(x_0) P_0^0(x_0, x_1) P_1^0(x_1, x_2) \cdots P_{K-1}^0(x_{K-1}, x_K), \tag{1.10a}$$

$$P_k^0(x, x') = T_k(x, s') \phi_{k+1}^0(u' \,|\, s')\,, \qquad x, x' \in \mathsf{X}. \tag{1.10b}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Any other $\{\phi_k\} \subset \Phi$ defines a Markov chain $\boldsymbol{X}$ with transition matrices,

$$P_k(x,x') := \mathsf{P}\{X_{k+1} = x' \mid X_k = x\} = T_k(x,s')\phi_{k+1}(u' \mid s')\,. \tag{1.11}$$

The marginals evolve according to linear dynamics, similar to (1.9),

$$\nu_k = \nu_{k-1}P_{k-1}\,, \qquad 1 \le k \le K, \tag{1.12}$$

in which ν_k is interpreted as an n-dimensional row vector with $n = |\mathsf{X}|$.

We obtain a convex program by optimizing over $\{\nu_k\}$, similar to the LP approach of [34]. Scalar variables $\{\gamma_k\}$ are introduced to simplify the objective, in anticipation of a Lagrangian decomposition:

$$J^\star(\nu_0^0) := \min_{\nu,\gamma} \left[\sum_{k=1}^K \mathcal{D}(\nu_k, \nu_k^0) + \frac{\kappa}{2} \sum_{k=1}^K \gamma_k^2 \right] \tag{1.13a}$$

$$\text{s.t.} \quad \gamma_k = \langle \nu_k, \mathcal{Y} \rangle - r_k\,, \tag{1.13b}$$

$$\sum_{u'} \nu_k(s',u') = \sum_{s,u} \nu_{k-1}(s,u)T_{k-1}(x,s')\,, \qquad 1 \le k \le K. \tag{1.13c}$$

The *relative entropy rate* is adopted as the cost of deviation:

$$\mathcal{D}(\nu_k, \nu_k^0) := \sum_{s,u} \nu_k(s,u) \log\left(\frac{\phi_k(u \mid s)}{\phi_k^0(u \mid s)}\right). \tag{1.14}$$

The terminology is justified through the following steps. First, we have seen that any randomized policy gives rise to a pmf $p \in \mathcal{S}(\mathsf{X}^{K+1})$ that is Markovian: $p(\vec{x}) = \nu_0^0(x_0)P_0(x_0,x_1)P_1(x_1,x_2)\cdots P_{K-1}(x_{K-1},x_K)$. The *relative entropy* (also known as Kullback–Leibler divergence) is the mean log-likelihood:

$$D(p\|p^0) = \sum L(\vec{x})\, p(\vec{x}), \tag{1.15}$$

where $L = \log(p/p^0)$ is an extended-real-valued function on X^{K+1}. The expression for P_k in (1.11) and the analogous formula for P_k^0 using ϕ_{k+1}^0 gives

$$L(\vec{x}) = \log\left(\frac{p(\vec{x})}{p^0(\vec{x})}\right) = \sum_{k=0}^{K-1} \log\left(\frac{P_k(x_k,x_{k+1})}{P_k^0(x_k,x_{k+1})}\right) \quad = \sum_{k=1}^{K} \log\left(\frac{\phi_k(u_k \mid s_k)}{\phi_k^0(u_k \mid s_k)}\right). \tag{1.16}$$

Consequently, $D(p\|p^0) = \sum_{k=1}^K \mathcal{D}(\nu_k, \nu_k^0)$.

The simple proof of Proposition 1.1 may be found in [13].

PROPOSITION 1.1. *With $\mathcal{D}$ chosen as the relative entropy rate* (1.14)*, the optimization problem* (1.13) *is convex in $\{\nu_k, \gamma_k : 1 \le k \le K\}$. Furthermore, the linear constraints in* (1.13c) *are equivalent to* (1.12)*.*

1.4. Motivation from linear systems theory. The approach to feedback control proposed in section 3 begins with consideration of the infinite-horizon KLQ problem. This is tractable only subject to additional assumptions.

It is assumed that the nominal model is a time-homogeneous Markov chain and that the reference signal is *constant,* $r_k \equiv r$, $k \ge 0$. On optimizing for each $r \in \mathbb{R}$ we obtain a continuous family of optimizers, $\{\phi_k^\star(u \mid s; r) : (s,u) \in \mathsf{X}, k \ge 0\,,\ r \in \mathbb{R}\}$. A potentially useful policy for tracking is then

$$\phi_k(u \mid s) = \phi_k^\star(u \mid s; r_k)\,, \qquad (s,u) \in \mathsf{X},\ \ k \ge 0. \tag{1.17}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Motivation for this approach may be found in the theory of optimal control for linear systems.

Consider the linear system with n-dimensional state $\boldsymbol{X}$, m-dimensional input $\boldsymbol{U}$, and scalar output $\boldsymbol{Y}$, evolving as

$$X_{k+1} = AX_k + BU_k + N_{k+1}\,, \quad Y_k = C^T X_k + W_{k+1}, \tag{1.18}$$

in which $\{N_{k+1}, W_{k+1}\}$ are independent and identically distributed (i.i.d.), mutually independent, with zero mean and finite covariances. The cost is quadratic, $c(x,u;r) = (y-r)^2 + u^T R u$ with $R > 0$.

The goal is to solve the average-cost optimal control problem. The solution is obtained via state-augmentation: define $X_k^r = [X_k; r_k]$, where $r_{k+1} = r_k = r$ defines the dynamics. The solution is linear state feedback,

$$U_k = -K^\star X_k + G^\star r\,, \qquad k \ge 0\,, \tag{1.19}$$

where $[K^\star; G^\star]$ is the optimal gain. The optimal gain does *not* depend on r or the distribution of N_k or W_k.

The special case in which the disturbances are *zero* is most closely related to the nonlinear control problem considered in section 3. Consider the objective

$$J_K^\star(x) = \min_{\boldsymbol{U}} \sum_{n=0}^{K} c(X_k, U_k) = \min_u \big\{ c(x,u;r) + J_{K-1}^\star(Ax+Bu) \big\}\,, \qquad X_0 = x \in \mathbb{R}^n\,.$$

It is not useful to let $K \to \infty$ without modification, since the cost $c(x,u;r)$ is never zero. This is why the relative value functions $h_K(x) = J_K^\star(x) - J_K^\star(0)$ are introduced, which solve the Bellman equation in modified form,

$$\eta_K + h_K(x) = \min_u \big\{ c(x,u;r) + h_{K-1}(Ax+Bu) \big\}\,, \qquad X_0 = x \in \mathbb{R}^n\,,$$

with $\eta_K = J_K^\star(0) - J_{K-1}^\star(0)$. As $K \to \infty$, the pair (η_K, h_K) converge to a solution to the average-cost optimality equation (ACOE),

$$\eta^\star + h^\star(x) = \min_u \big\{ c(x,u;r) + h^\star(Ax+Bu) \big\}\,, \qquad x \in \mathbb{R}^n\,,$$

whose minimizer is precisely (1.19). The proof is standard, though usually presented in the purely stochastic setting. It is especially simple in this LQR setting since each of the functions $\{h_N\}$ is quadratic [41, 36].

When $\boldsymbol{r}$ is time varying, it is standard practice to apply the "hack"

$$U_k = -K^\star X_k + G^\star r_k\,, \qquad k \ge 0\,. \tag{1.20}$$

The most compelling motivation is found in the deterministic, continuous time setting: under mild conditions, the return difference equation tells us that the closed loop dynamics from reference input to output are *passive* [3]. Passivity is lost for discrete time models but can be expected to hold approximately when the discrete time model is obtained from sampling a continuous time system.

1.5. Main results. The contributions of this paper fall into three categories:

(1) Feedforward control. Consideration of the dual of the convex optimization problem (1.13) leads to many insights. The main conclusions summarized here are a special case of Theorem 2.1.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

THEOREM 1.2. [KLQ solution]. *Consider the convex program* (1.13). *An optimizer* $\{\phi_k^\star : 1 \le k \le K\}$ *exists, is unique, and is of the form*

$$\phi_k^\star(u \mid s) = \phi_k^0(u \mid s) \exp\left(\sum_{s'} T_k(x,s') g_{k+1}^\star(s') + \lambda_k^\star \mathcal{Y}(s,u) - g_k^\star(s) \right), \tag{1.21a}$$

$$\text{where} \quad g_k^\star(s) = \log\left(\sum_u \phi_k^0(u \mid s) \exp\left(\sum_{s'} T_k(x,s') g_{k+1}^\star(s') + \lambda_k^\star \mathcal{Y}(s,u) \right) \right), \tag{1.21b}$$

and $\{\lambda_k^\star : 1 \le k \le K\}$, $\{g_k^\star(s) : 1 \le k \le K\}$ *are the Lagrange multipliers for the constraints* (1.13b) *and* (1.13c), *respectively, and* $g_{K+1} \equiv 0$.

Proposition 2.2 motivates a two-step approach in which $\lambda^\star$ is obtained as the solution to a convex program that maximizes the dual function $\varphi^\star$, and then $g^\star$ are computed through the nonlinear recursion (1.21b). Hence the larger computational challenge is computing $\lambda^\star$. Expressions for the derivatives of $\varphi^\star$ involve means and variances of $\mathcal{Y}(X_k)$, which invites the application of Monte Carlo techniques when the state space is large or even uncountable—see [13] for details.

(2) Feedback. Section 3 concerns control design following steps analogous to the approach used in linear systems theory to obtain the feedback control strategy (1.20). Justification of the ACOE requires that we turn to a time-homogeneous model, meaning that $T_k = T$ and $\phi_k^0 = \phi^0$, independent of k.

Even with $r_k \equiv r$ fixed, the solution to (1.4) is not time homogeneous, but on letting $K \to \infty$ the policies converge to a solution of an ACOE. This is equivalently expressed as the solution to a deterministic optimal control problem,

$$\text{System: } \nu_{k+1} = f(\nu_k, \phi_k), \quad \text{Cost: } c(\nu, \phi; r) = D_\infty(\hat{\nu}, \phi) + \frac{\kappa}{2}[\langle \nu, \mathcal{Y} \rangle - r]^2, \tag{1.22}$$

where the marginals $\{\nu_k\}$ are viewed as a state process, evolving on the simplex $\mathcal{S}(\mathsf{X})$, and $\phi_k \in \Phi$ is regarded as an input. The system equation is of the form (1.12), but simplified because of the time-homogeneity assumptions imposed here, giving

$$f(\nu, \phi)\big|_{x'=(s',u')} = \sum_{x \in \mathsf{X}} \nu(x) T(x,s') \phi(u' \mid s').$$

Hence f is bilinear in the pair (ν, ϕ). Identification and justification of the term D_∞ in (1.22) require further notation and analysis.

Consider the infinite-horizon objective,

$$\eta^\star(r) = \min \limsup_{K \to \infty} \frac{1}{K} \sum_{k=1}^K \left[\mathcal{D}(\nu_k, \nu_k^0) + \frac{\kappa}{2} \big[\langle \nu_k, \mathcal{Y} \rangle - r \big]^2 \right] \tag{1.23}$$

in which the minimum is over all $\{\phi_k\} \subset \Phi$. The following notational conventions are required to describe the structure of its solution:

(i) Any $\phi \in \Phi$ defines a transition matrix P_ϕ, and any pmf π that is invariant for P_ϕ admits the decomposition

$$\pi(s,u) = \phi(u \mid s) \hat{\nu}(s), \tag{1.24}$$

where $\hat{\nu}$ is the steady-state pmf for $\boldsymbol{S}$ under this policy.

(ii) With ϕ and $\hat{\nu}$ as above, the steady-state relative entropy rate is denoted

$$D_\infty(\hat{\nu}, \phi) := \sum_{s,u} \phi(u \mid s) \hat{\nu}(s) \log\left(\frac{\phi(u \mid s)}{\phi^0(u \mid s)} \right). \tag{1.25}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

THEOREM 1.3. [infinite-horizon KLQ solution]. *Suppose that the nominal transition matrix P^0 has unique invariant pmf π^0, and fix any $\kappa > 0$ and $r \in \mathbb{R}$. Then, there is a solution to* (1.23) *in which $\phi_k^\star = \phi^\star$ for each k, obtained from the optimization problem*

$$\arg\min_{\pi,\phi} \left\{ D_\infty(\hat{\nu}, \phi) + \frac{\kappa}{2}[\langle \pi, \mathcal{Y} \rangle - r]^2 : \pi P_\phi = \pi \right\}. \tag{1.26}$$

This optimization problem is convex with unique solution $\{\pi^\star, \phi^\star\}$.

The convex program (1.26) reduces to the "IPD" convex program of [37, 8, 22] as $\kappa \to \infty$ (see discussion surrounding (1.29) in the literature review). The two convex programs are differentiated by the introduction of a quadratic cost on the marginals, so the policy $\phi^\star$ obtained from (1.26) is henceforth called the *IPD-Q solution.* Much of section 3 is devoted to obtaining approximations of this solution, as well as computational methods to obtain the exact solution.

(a) HJB solution and LQR approximation. Viewed as a deterministic optimal control problem, with system and cost given in (1.22), another solution to (1.23) is obtained as state feedback $\phi_k^\star = \mathcal{K}^\star(\nu_k^\star, r)$ for some mapping $\mathcal{K}^\star \colon \mathcal{S}(\mathsf{X}) \times \mathbb{R} \to \Phi$. The IPD-Q solution is obtained via $\phi^\star = \mathcal{K}^\star(\nu^\star; r)$ with $\nu^\star$ the steady-state marginal of $\boldsymbol{S}$ under $P_{\phi^\star}$.

Because computation of $\mathcal{K}^\star$ is complex if $|\mathsf{X}|$ is large, and in anticipation of finer analysis of the performance of this policy, much of section 3.1 is devoted to "small signal" approximations.

Let $\{x^i = (s^i, u^i) : 1 \le i \le n\}$ be an enumeration of the state space X with $n = |\mathsf{X}|$. As a corollary to Propositions 3.2 and 3.3, coefficients $\{K_{i,j}^\star, G_i^\star\}$ are constructed for which $\phi_k^\star(u^i \mid s^i, r) = \phi_k(u^i \mid s^i, r) + O(r^2)$ for each i with

$$\phi_k(u^i \mid s^i, r) := \phi^0(u^i \mid s^i) \exp\left(\frac{1}{\phi^0(u^i \mid s^i)} \left(-\sum_j K_{i,j}^\star \widetilde{\nu}_k(x^j) + G_i^\star r \right) - \Gamma(s^i, r) \right), \tag{1.27}$$

$\widetilde{\nu}_k(x^i) := \nu_k(x^i) - \pi^0(x^i)$ for each i, and Γ is a normalizing constant, defined so that $\phi_k(\cdot \mid s^i, r)$ is a pmf on U for each s^i, r.

(b) Lagrangian relaxation. A Lagrangian relaxation leads to a characterization of the IPD-Q solution in terms of a standard ACOE. Similar to (1.13b), we introduce the variable $\gamma = \langle \nu_k, \mathcal{Y} \rangle - r$, and let $\lambda^\star \in \mathbb{R}$ denote the Lagrange multiplier associated with this scalar constraint; it is identified in (3.4) as $\lambda^\star = \kappa[r - \langle \pi^\star, \mathcal{Y} \rangle]$.

The relative value function $h^\star$ that solves the ACOE provides a representation of the IPD-Q solution in (3.6):

$$\phi^\star(u \mid s) = \phi^0(u \mid s) \exp\big(\bar{h}^\star(s,u) + \lambda^\star \mathcal{Y}(s,u) - \Gamma^\star(s) \big),$$

with $\bar{h}^\star(x) = \sum_{u'} T(x, s') h^\star(s')$ and $\Gamma^\star(s)$ a normalizing factor.

(c) ODE solution and small signal approximation. Rather than compute $\lambda^\star$ for each r, it is argued that it is simpler to let λ be the independent variable. The family of relative value functions $\{h^\lambda : \lambda \in \mathbb{R}\}$ solves an ordinary differential equation, whose vector field is identified in (3.10). In addition to offering a tool for exact computation, this leads to approximation of the IPD-Q solution.

These conclusions lead to several approaches to feedback control for tracking a time varying reference signal. Remember, in the following three options, the family $\{\phi_k : k \ge 0\}$ is proposed for local decision making in a mean-field control architecture:

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

1. The feedback solution (1.17) using the collection $\{\phi_k^\star(\cdot \mid \cdot ; r) : k \ge 0\,,\ r \in \mathbb{R}\}$.
2. The open-loop strategy $\phi_k(\cdot \mid \cdot) = \phi^\star(\cdot \mid \cdot ; r_k)$, with $\{\phi^\star(\cdot \mid \cdot ; r) : r \in \mathbb{R}\}$ the IPD-Q solutions.
3. In option 2 above, it is assumed that r_k is made available to each agent, at each time k, as an external control signal. A refinement is obtained by designing a control signal $\{\zeta_k : k \ge 0\}$ based on filtering measurements, such as error feedback,

$$\zeta_k = \sum_{i=0}^{k} g_{k-i} e_i\,, \quad k \ge 0\,, \qquad e_i = r_i - \langle \nu_i, \mathcal{Y} \rangle. \tag{1.28}$$

The randomized decision rule for each agent is then $\phi_k(\cdot \mid \cdot) = \phi^\star(\cdot \mid \cdot ; \zeta_k)$ with $\{\phi^\star(\cdot \mid \cdot ; \zeta) : \zeta \in \mathbb{R}\}$ the IPD-Q solutions. The linearized dynamics described in Proposition 3.5 can aid in the design of the filter in (1.28).

(3) Application to demand dispatch. The original motivation for the research surveyed here is application to distributed control of power systems. The term *demand dispatch* was introduced in the conceptual article [7] to describe the possibility of distributed intelligence in electric loads, designed so that the population would help provide supply-demand balance in the power grid.

The numerical results surveyed in section 4 illustrate the application of KLQ to control a large population of residential loads. As expected, tracking error can be made arbitrarily small with large $\kappa > 0$, provided the reference signal is feasible.

It is found in numerical experiments that the histograms defining the state of the mean-field model rapidly "forget" their initial conditions—see the full arXiv version [13] for details.

1.6. Literature review.

Mean-field control. The optimization problem (1.3) is inspired by mean-field game theory [31, 28, 29, 10, 26] (see [16, 17, 10, 42] for recent surveys).

Mean-field control differs from mean-field game theory only because of greater control at the microscopic layer: we do not assume that an individual in the population is free to optimize based on its local objective function, so we avoid the fragility of Nash equilibria. This description is similar to *ensemble control* in physics (see [32] for history), and many in the power systems area opt for this term rather than mean-field control (see [23, 22] and their references).

Demand dispatch. The goal of demand dispatch is to modify the behavior of loads so that their aggregate power consumption tracks a reference signal $\{r_k\}$ that is synthesized by a *balancing authority* (BA). Randomized control techniques have been proposed in [35, 43, 37, 1, 23, 4] based on various control architectures.

The following control strategy is common to the approaches described in [37, 22]. It is assumed that a family of transition matrices $\{P_\zeta : \zeta \in \mathbb{R}\}$ is available at each load. A sequence $\{\zeta_0, \zeta_1, \dots\}$ is broadcast from the BA, based on measurements of the grid, and at time k the ith load transitions according to this law:

$$\mathsf{P}\{X_{k+1}^i = x' \mid X_k^i = x,\ \zeta_k = \zeta\} = P_\zeta(x, x').$$

The feedback solution (1.28) was proposed in [37] and tested in this and later research using $e_i = r_i - \langle \nu_i^{\mathcal{N}}, \mathcal{Y} \rangle$ [22].

IPD. The paper [37] reinterprets the control solution of [44] as a technique to create the family $\{P_\zeta\}$ through the solution to the nonlinear program:

$$P_\zeta := \arg\max \left\{ \zeta \langle \pi, \mathcal{Y} \rangle - \mathcal{R}(P \| P^0) \right\}, \qquad \zeta \in \mathbb{R}\,, \tag{1.29}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where $\mathcal{R}$ denotes the rate function of Donsker and Varadhan [25, 30],

$$\mathcal{R}(P\|P^0) := \sum_{x,x'} \pi(x)P(x,x') \log\left(\frac{P(x,x')}{P^0(x,x')}\right), \tag{1.30}$$

in which π is the invariant pmf for P. The maximum in (1.29) is over all (π, P) subject to the invariance constraint $\pi P = \pi$ [37, 8]. The convex program (1.29) is called the *individual perspective design* (IPD) in [8].

Hence IPD-Q may be interpreted as a new approach to designing $\{P_\zeta\}$.

The finite-horizon version of (1.29) is also considered in [37, 8], similar to the KLQ formulation:

$$p^\zeta := \arg\max_p \left\{ \zeta \mathsf{E}_p \left[\sum_{k=1}^K \mathcal{Y}(x_k) \right] - D(p\|p^0) \right\}. \tag{1.31}$$

Provided the entries of $T_k(x,s)$ take on only binary values, the finite-horizon IPD solution is obtained as a tilting of the nominal model:

$$p^\zeta(\vec{x}) = p^0(\vec{x}) \exp\left(\zeta \sum_{k=1}^K \mathcal{Y}(x_k) - \Lambda(\zeta) \right) \quad \textit{with } \Lambda(\zeta) \textit{ a normalizing constant.} \tag{1.32}$$

KLQ and optimal transport. Extensions of the KLQ objective will likely provide useful relaxations of the classical *optimal transport* problem, in which the goal is to steer p^0 to a given target pmf $p^\star$ [39, 21]. Rather than match the target pmf, we might match M generalized moments, minimizing $D(p\|p^0)$ subject to $\langle p, \mathcal{G}_i \rangle = \langle p^\star, \mathcal{G}_i \rangle$ for each i, with $\mathcal{G}_i \colon \mathsf{X}^{K+1} \to \mathbb{R}$.

A special case is the tracking problem,

$$\min_p \big\{ D(p\|p^0) \quad \text{subject to } \mathsf{E}_p\big[\mathcal{Y}(X_k)\big] = r_k\,, \quad 1 \le k \le K \big\}. \tag{1.33}$$

This optimization problem is proposed in [23, section 5], along with the explicit solution

$$p^\star(\vec{x}) = p^0(\vec{x}) \exp\left(\sum_{k=1}^K \beta_k \mathcal{Y}(x_k) - \Lambda(\beta) \right) \tag{1.34}$$

in which $\beta \in \mathbb{R}^K$ are Lagrange multipliers corresponding to the K constraints, and $\Lambda(\beta)$ a normalizing constant.

The convex program formulation (1.13) has many advantages. First, (1.13) is always feasible, while feasibility of (1.33) requires conditions on p^0 and $\{r_k\}$. Theorem 1.2 requires no assumptions on the model or reference signal. Flexibility in choice of κ allows for *learning* the characteristics of an "expensive" reference signal. It is anticipated that the penalty parameter κ can be used to make trade-offs between tracking performance and robustness to modeling error: robustness and sensitivity analysis will be a topic of future research.

Finally, as assumed to obtain the representation (1.32), the formula (1.34) is meaningful only when $T_k(x,s)$ take on only binary values. A goal of the research surveyed in this paper is to remove this restriction.

The similarity between (1.32) and (1.34) is not accidental but follows from an alternative interpretation of the IPD design (1.31). For a scalar $r_0 \in \mathbb{R}$, consider the constrained optimization problem

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\max_p \{-D(p\|p^0)\} \quad \text{subject to } \mathsf{E}_p\left[\sum_{k=1}^K \mathcal{Y}(x_k)\right] = Kr_0\,, \quad 1 \le k \le K. \tag{1.35}$$

The dual function $\varphi^\star\colon \mathbb{R} \to \mathbb{R}$ is defined by

$$\varphi^\star(\lambda) = \max_p \left\{ \lambda \mathsf{E}_p\left[\sum_{k=1}^K \mathcal{Y}(x_k)\right] - D(p\|p^0) \right\} - \lambda K r_0,$$

where $\lambda \in \mathbb{R}$ is a Lagrange multiplier. It is evident that the optimizer $p^{*\lambda}$ is an IPD solution for each λ. Consequently, for each ζ, the IPD solution (1.31) also solves (1.35) for some scalar $r_0(\zeta)$.

Contributions. Most of the contributions were surveyed in section 1.5. The main contribution of this paper is the discovery of hidden convexity in the nonlinear program (1.13), which leads to structure for the optimal solution in Theorem 1.2. Properties of the dual surveyed in Theorem 2.1 lead to computational techniques for this new class of optimal control formulations; see Proposition 2.2 and its corollary. The application of these techniques to the infinite-horizon setting in section 3 is novel, and the main results surveyed there are new.

Portions of the results reported here were summarized in the conference article [15]. In this preliminary work, the transition matrix T_k was assumed *deterministic,* so that all randomness arose from the randomized policy. All of the results in this paper allow for general Markovian dynamics.

Extensions to resource allocation are summarized in [14]. More on these topics may be found in the first author's Ph.D. dissertation [12].

Organization. The remainder of this paper is organized as follows. Section 2 describes a relaxation technique motivated by the desire to reduce computational complexity, along with a full analysis of the convex program (1.13) and its dual. Section 3 contains extensions to the infinite-horizon setting. Results from numerical experiments are collected together in section 4. Conclusions and directions for future research are contained in section 5.

2. Kullback–Leibler-quadratic optimal control.

2.1. Subspace relaxation. A relaxation of the convex program (1.13) is described here. Motivation is most clear from consideration of distributed control of a collection of residential water heaters. These loads are valuable as sources of virtual energy storage since they in fact are energy storage devices (in the form of heat rather than electricity) and are also highly flexible. Flexibility comes in part from their extremely nonsymmetric behavior: a typical unit may be on for just five minutes and off continuously for more than six hours. The intersampling time at the load should be far less than five minutes to obtain a reliable model for control.

On the other hand, it is valuable for the time horizon to be on the order of several hours. For example, peak-shaving is more effective when water heaters have advance warning to preheat the water tanks. To obtain a useful control solution will thus require a very large value of K in (1.13). To reduce complexity, an approach is proposed here based on lossy compression of $\{r_k\}$ using transform techniques.

The transformations are based on a collection of functions $\{w_n : 1 \le n \le N\}$, with $w_n\colon \{0, 1, \dots, K\} \to \mathbb{R}$ for each n, and $N \ll K$. The transformed signal is the N-dimensional vector $\hat{r}$ with $\hat{r}_n = \sum_k w_n(k) r_k$ for each n, and the transformed function on X^{K+1} is denoted $\widehat{\mathcal{Y}}_n(\vec{x}) = \sum_{k=1}^K w_n(k)\mathcal{Y}(x_k)$ for $1 \le n \le N$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The goal is to achieve the approximation $\langle p, \widehat{\mathcal{Y}}_n \rangle \approx \hat{r}_n$ for each n, while maintaining $p \approx p^0$. For example, a Fourier series can be used, with frequency $\omega > 0$, and N is necessarily odd: $w_n(k) \in \{1, \sin(\omega mk), \cos(\omega mk) : 1 \le m \le (N-1)/2\}$. The *degenerate* family is defined using $N = K$ and

$$w_n(k) = \mathbb{I}\{n = k\}\,, \qquad 1 \le n, k \le K. \tag{2.1}$$

The optimal control problem with *subspace relaxation* is defined as

$$J^\star(\nu_0^0) := \min_{\nu,\gamma} \sum_{k=1}^K \mathcal{D}(\nu_k, \nu_k^0) + \frac{\kappa}{2} \sum_{n=1}^N \gamma_n^2 \tag{2.2a}$$

$$\text{s.t.} \quad \gamma_n = \langle p, \widehat{\mathcal{Y}}_n \rangle - \hat{r}_n\,, \quad 1 \le n \le N, \tag{2.2b}$$

$$\sum_{u'} \nu_k(s', u') = \sum_{s,u} \nu_{k-1}(s, u) T_{k-1}(x, s')\,, \qquad 1 \le k \le K\,,\ s' \in \mathsf{S}. \tag{2.2c}$$

This reduces to (1.13) in the degenerate case (2.1).

The theory that follows is based in part on a relaxation of the dynamical constraints (2.2c), through the introduction of a Lagrange multiplier for each k. This is precisely the first step in the construction of the Hamiltonian in the minimum principle approach to optimal control [33].

2.2. Duality. Structure for the solution of (2.2) will be obtained by consideration of a dual, in which $\lambda \in \mathbb{R}^N$ and $g \in \mathbb{R}^{K \times J^\star}$ denote the vectors of Lagrange multipliers for the first and second sets of constraints, respectively. The matrix g is interpreted as a sequence of functions $g_k \colon \mathsf{S} \to \mathbb{R}$ that are entirely analogous to the costate variables in the minimum principle (the Lagrange multipliers for the dynamical constraints) [33].

The Lagrangian is denoted

$$\begin{aligned} \mathcal{L}(\nu, \gamma, \lambda, g) = &\sum_{k=1}^K \mathcal{D}(\nu_k, \nu_k^0) + \frac{\kappa}{2} \sum_{n=1}^N \gamma_n^2 + \sum_{n=1}^N \lambda_n \left(\gamma_n + \sum_{k=1}^K w_n(k) \big[r_k - \langle \nu_k, \mathcal{Y} \rangle \big] \right) \\ &+ \sum_{k=1}^K \sum_{s'} \left(\sum_{u'} \nu_k(s', u') - \sum_{s,u} \nu_{k-1}(s, u) T_{k-1}(x, s') \right) g_k(s') \end{aligned} \tag{2.3}$$

and the dual function is defined to be its minimum, $\varphi^\star(\lambda, g) := \min_{\nu,\gamma} \mathcal{L}(\nu, \gamma, \lambda, g)$.

The *dual* of the optimization problem (2.2) is defined as the maximum of the dual function $\varphi^\star$ over λ and g (see [33] for a complete and accessible treatment of this theory). We will see that there is no duality gap, so that for a quadruple $(\nu^\star, \gamma^\star, \lambda^\star, g^\star)$,

$$J^\star(\nu_0^0) = \mathcal{L}(\nu^\star, \gamma^\star, \lambda^\star, g^\star) = \varphi^\star(\lambda^\star, g^\star)\,.$$

In the following subsections a representation of the dual function is obtained that is suitable for optimization, which results in a valuable representation for the optimal policy. Properties of the dual function are contained in Theorem 2.1 and Proposition 2.2 that follow. The statement of these results requires additional notation: define a function $\mathcal{T}_k^\lambda \colon \mathbb{R}^{|\mathsf{S}|} \to \mathbb{R}^{|\mathsf{S}|}$, for $f \colon \mathsf{S} \to \mathbb{R}$ and $\lambda \in \mathbb{R}^N$, via

$$\mathcal{T}_k^\lambda(f; s) = \log \left(\sum_u \phi_k^0(u \,|\, s) \exp \left(\sum_{s'} T_k(x, s') f(s') + \check{\lambda}_k \mathcal{Y}(s, u) \right) \right), \quad s \in \mathsf{S}\,, \tag{2.4}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where $\check{\lambda}_k = \sum_{n=1}^N \lambda_n w_n(k)$. The maximum of the dual function over g is denoted

$$\varphi^\star(\lambda) := \max_g \phi^\star(\lambda, g) = \varphi^\star(\lambda, g^\lambda),$$

where g^λ is a maximizer, $g^\lambda \in \arg\max_g \phi^\star(\lambda, g)$. It is shown in Proposition 2.2 that the vector valued function g^λ satisfies the recursion

$$g_k^\lambda = \mathcal{T}_k^\lambda(g_{k+1}^\lambda), \quad 1 \le k \le K, \quad \text{where} \quad g_{K+1}^\lambda \equiv 0. \tag{2.5}$$

This forms part of the proof of Theorem 2.1, with complete details postponed to Appendix B.

THEOREM 2.1. *There exists a maximizer $\{\lambda_n^\star, g_k^\star : 1 \le n \le N, 1 \le k \le K\}$ for $\varphi^\star$, and there is no duality gap: $\varphi^\star(\lambda^\star, g^\star) = J^\star(\nu_0^0)$. The optimal policy is obtained from $\{g_k^\star\}$ via*

$$\phi_k^\star(u \mid s) = \phi_k^0(u \mid s) \exp\left(\sum_{s'} T_k(x, s') g_{k+1}^\star(s') + \check{\lambda}_k^\star \mathcal{Y}(s, u) - g_k^\star(s) \right), \tag{2.6}$$

where $g_k^\star(s) = \mathcal{T}_k^\lambda(g_{k+1}^\star; s)$ for $1 \le k \le K$, and $g_{K+1}^\star \equiv 0$,

and $\{\check{\lambda}_k^\star\}$ are obtained from $\{\lambda_n^\star\}$ via (2.4).

The proof of the following is also contained in Appendix B. Denote for each k,

$$G_k^\lambda(x) = \sum_s T_{k-1}(x, s) g_k^\lambda(s). \tag{2.7}$$

PROPOSITION 2.2. *The following hold for the dual of (2.2): for each $\lambda \in \mathbb{R}^N$,*

(i) *a maximizer g^λ is given by (2.5);*

(ii) *the maximum of the dual function over g is the concave function*

$$\varphi^\star(\lambda) = \lambda^T \hat{r} - \frac{1}{2\kappa} \|\lambda\|^2 - \langle \nu_0^0, G_1^\lambda \rangle; \tag{2.8}$$

(iii) *the function (2.8) is continuously differentiable, and*

$$\frac{\partial}{\partial \lambda_n} \varphi^\star(\lambda) = \hat{r}_n - \frac{1}{\kappa} \lambda_n - \sum_{k=1}^K w_n(k) \langle \nu_k^\lambda, \mathcal{Y} \rangle, \qquad 1 \le n \le N, \tag{2.9}$$

where $\{\nu_k^\lambda\}$ is the sequence of marginals obtained from the randomized policy defined in (2.6), substituting $\{g_k^\star\}$ by $\{g_k^\lambda\}$ defined in (i).

To conclude this section, we provide representations of the log-likelihood ratio, $L(\vec{x})$, relative entropy $D(p^\lambda \| p^0)$, and primal objective function for the pmf $p^\lambda \in \mathcal{S}(\mathsf{X}^{K+1})$ obtained from the randomized policy defined in (2.6), substituting $\{g_k^\star\}$ by $\{g_k^\lambda\}$ defined in Proposition 2.2, part (i). We defer to [13] for the proof of the following.

COROLLARY 2.3. *The following hold for all $\{\check{\lambda}_k, g_k^\lambda : 1 \le k \le K\}$:*

(i) *The log-likelihood ratio can be expressed as*

$$L(\vec{x}) = \sum_{k=1}^K \{\Delta_k(x_{k-1}, s_k) + \check{\lambda}_k \mathcal{Y}(x_k)\} - G_1^\lambda(x_0), \tag{2.10}$$

where for each k (recalling $x_k = (s_k, u_k)$),

$$\Delta_k(x_{k-1}, s_k) = G_k^\lambda(x_{k-1}) - g_k^\lambda(s_k). \tag{2.11}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

(ii) *The relative entropy is given by*

$$D(p^\lambda \| p^0) = \sum_{k=1}^K \check{\lambda}_k \langle \nu_k^\lambda, \mathcal{Y} \rangle - \langle \nu_0^0, G_1^\lambda \rangle. \tag{2.12}$$

(iii) *The value of the primal is given by*

$$J(p^\lambda, \nu_0^0) := D(p^\lambda \| p^0) + \frac{\kappa}{2} \sum_{n=1}^N \gamma_n^2 \tag{2.13a}$$

$$= -\langle \nu_0^0, G_1^\lambda \rangle + \sum_{k=1}^K \check{\lambda}_k \langle \nu_k^\lambda, \mathcal{Y} \rangle + \frac{\kappa}{2} \sum_{n=1}^N \gamma_n^2 \tag{2.13b}$$

with $\gamma_n = \langle p^\lambda, \widehat{\mathcal{Y}}_n \rangle - \hat{r}_n$.

The stochastic process $\{\Delta_k(X_{k-1}, S_k)\}$ is a martingale difference sequence; it vanishes when nature is deterministic, reducing to the solution obtained in [15].

3. Feedback formulations. We now turn to the IPD-Q convex program (1.26). It is assumed throughout this section that $T_k = T$ and $\phi_k^0 = \phi^0$, independent of k.

The relationship between IPD-Q and (1.23) will be clear after justification of the term $D_\infty(\hat{\nu}, \phi)$ defined in (1.25). Consider any $\phi \in \Phi$, which gives rise to a Markov chain with transition matrix P_ϕ. The relative entropy (1.15) was previously expressed as a sum over $\vec{x} \in \mathsf{X}^{K+1}$ in (1.15). The notation $D(p\|p^0) = D^K(p\|p^0)$ is required in the following, since K is a variable in (1.23).

PROPOSITION 3.1. *Suppose that p is obtained using the policy ϕ, and initial pmf ν_0^0 common with p^0. Suppose moreover that P_ϕ has a unique invariant pmf π. Then,*

$$D_\infty(\hat{\nu}, \phi) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^K \mathcal{D}(\nu_k, \nu_k^0) = \lim_{K\to\infty} \frac{1}{K} D^K(p\|p^0) = \mathcal{R}(P_\phi \| P^0),$$

where $\mathcal{R}$ denotes the rate function (1.30) *using $P = P_\phi$:*

$$\mathcal{R}(P_\phi \| P^0) = \sum_{x,x'} \pi(x) P_\phi(x,x') \log\left(\frac{P_\phi(x,x')}{P^0(x,x')}\right).$$

Proof. The proof of the first identity begins with

$$\frac{1}{K} \sum_{k=1}^K \mathcal{D}(\nu_k, \nu_k^0) = \frac{1}{K} \sum_{k=1}^K \mathsf{E}[F(X_k))]$$

with $F(x) = \log[\phi_k(u \,|\, s)/\phi_k^0(u \,|\, s)]$ for $x = (s,u) \in \mathsf{X}$. The average converges to $D_\infty(\hat{\nu}, \phi)$ as $K \to \infty$ since the invariant pmf π is unique. ∎

The distinct approaches to optimal control pursued in this section follow the distinct approaches to optimal control in general, via the HJB equations and optimal control via the minimum principle:

(i) In section 3.1 IPD-Q is interpreted as a solution to an HJB equation, which results in a solution in state feedback form, $\phi_k^\star = \mathcal{K}^\star(\nu_k^\star, r)$, for some mapping $\mathcal{K}^\star \colon \mathcal{S}(\mathsf{X}) \times \mathbb{R} \to \Phi$. The solution to IPD-Q is $\phi^\star(u \,|\, s) = \mathcal{K}^\star(\nu^r, r)$, in which ν^r is the steady-state marginal for $\boldsymbol{S}$ under the IPD-Q policy. Computation

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

of $\mathcal{K}^\star$ may be difficult if the state space is large. An LQR approximation is proposed, justified for small $|r|$, and the approximation (1.27) may also be found at the close of section 3.1.

(ii) The approach taken in section 3.2 is in essence the infinite-K limit of the approach taken in section 2.2 which, as noted following (2.2), is the minimum principle approach. It is well known that this approach provides only an open-loop solution.

3.1. HJB approach. The solution to the optimal control problem (1.22) may be characterized using techniques from deterministic optimal control theory.

The ACOE holds for deterministic systems, precisely as reviewed in section 1.4 for the linear quadratic problem:

$$\eta^\star + \mathcal{H}^\star(\nu) = \min_{\phi} \big\{ c(\nu, \phi; r) + \mathcal{H}^\star(f(\nu, \phi)) \big\}, \qquad \nu \in \mathcal{S}(\mathsf{X}), \tag{3.1}$$

with $c(\nu, \phi; r)$ defined in (1.22), $\mathcal{H}^\star \colon \mathcal{S}(\mathsf{X}) \to \mathbb{R}$ the relative value function, and $\eta^\star$ the optimal average cost. The minimizer $\phi^\star$ defines $\phi^\star_k = \mathcal{K}^\star(\nu^\star_k, r)$.

We are not aware of solution techniques for this instance of the ACOE, beyond the standard value iteration algorithm or other generic approaches.

The relative value function $\mathcal{H}^\star$ and feedback law $\mathcal{K}^\star$ can be approximated through a small signal linearization of the dynamics, and a quadratic approximation of the cost. We begin with an approximation for the latter.

The proof of Proposition 3.2 follows from the definition (1.25) and a Taylor's series approximation of the logarithm. For any ϕ, denote by $\widetilde{\phi}(u \mid s) := \phi(u \mid s) - \phi^0(u \mid s)$ the deviation.

PROPOSITION 3.2. *Suppose that* $\langle \pi^0, \mathcal{Y} \rangle = 0$. *Then, the cost function* (1.22) *is nearly quadratic in deviations,*

$$c(\nu, \phi; r) = \|\widetilde{\phi}\|_R^2 + \frac{\kappa}{2}(y - r)^2 + O(\|\widetilde{\phi}\|_R^3)$$

in which $y = \sum_x [\nu(x) - \pi^0(x)] \mathcal{Y}(x)$, *and* $\|\widetilde{\phi}\|_R^2 = \sum_{s,u} \frac{\hat{\nu}(s)}{\phi^0(u|s)} \widetilde{\phi}(u \mid s)^2$.

Approximation of the dynamics by a linear system is justified when $|r|$ is small, and $\nu_0^0 \approx \pi^0$, the invariant pmf for P^0. The corresponding stationary pmf for $\boldsymbol{S}$ is denoted $\hat{\nu}^0$ (recall (1.24)).

Let $\mathsf{X} = \{x^i : 1 \le i \le n\}$ with $n = |\mathsf{X}|$. The LQR approximation has state denoted $\widetilde{\mathcal{X}}_k$ and input $\widetilde{\mathcal{U}}_k$ at time k, with $\widetilde{\mathcal{X}}^i_k$ an approximation of $\widetilde{\nu}_k(x^i) := \nu_k(x^i) - \pi^0(x^i)$, and $\widetilde{\mathcal{U}}^i_k$ an approximation of $\widetilde{\phi}^i_k(u^i \mid s^i) := \phi^i_k(u^i \mid s^i) - \phi^0(u^i \mid s^i)$. The definition of the linearization is a system model of the form (1.18),

$$\widetilde{\mathcal{X}}_{k+1} = A\widetilde{\mathcal{X}}_k + B\widetilde{\mathcal{U}}_k, \qquad \widetilde{\mathcal{Y}}_k = C^T \widetilde{\mathcal{X}}_k,$$

in which $\widetilde{\mathcal{Y}}_k$ is an approximation of $\langle \widetilde{\nu}_k, \mathcal{Y} \rangle$. Expressions for the $n \times n$ matrices A and B, and the n-dimensional column vector C, are provided in the following.

PROPOSITION 3.3. *The small signal approximation holds with*

$$A_{i,j} = P^0(x^j, x^i), \quad B_{i,j} = \mathbb{I}\{i = j\} \hat{\nu}^0(s^j), \quad C_i = \mathcal{Y}(x^i), \qquad 1 \le i, j \le n.$$

Proof. The expression for C is by definition of $\widetilde{\mathcal{Y}}_k$. The other matrices are obtained through the standard first-order Taylor series approximations:

$$A_{i,j} := \frac{\partial}{\partial \nu^j} f_i(\nu, \phi) \Big|_{\nu = \pi^0, \phi = \phi^0} = P^0(x^j, x^i)$$

with $\nu^j = \nu(x^j)$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

The input $\widetilde{\mathcal{U}}_k$ is an n-dimensional column vector, so that B is an $n \times n$ matrix. It is obtained from the Taylor series approximation,

$$B_{i,j} := \frac{\partial}{\partial \phi^j} f_i(\nu, \phi)\Big|_{\nu=\pi^0, \phi=\Phi^0} = \mathbb{I}\{i=j\} \sum_{x \in \mathsf{X}} \pi^0(x) T(x, s^j),$$

where $\phi^j = \phi(u^j \mid s^j)$. By invariance of π^0 it follows that B is diagonal, with ith diagonal entry equal to $\hat{\nu}^0(s^i)$. ∎

Propositions 3.2 and 3.3 imply that for small r, the solution to the nonlinear optimal control problem is approximated by the average-cost LQR solution using

$$c(\widetilde{\mathcal{X}}, \widetilde{\mathcal{U}}; r) = \|\widetilde{\mathcal{U}}\|_R^2 + \frac{\kappa}{2}(\widetilde{\mathcal{Y}} - r)^2, \quad \widetilde{\mathcal{Y}} = C^T \widetilde{\mathcal{X}},$$

giving $\widetilde{\mathcal{U}}_k = -K^\star \widetilde{\mathcal{X}}_k + G^\star r$, with gain matrices $K^\star$ $(n \times n)$ and $G^\star$ $(n \times 1)$.

This leads to the policy approximation. Write $\widetilde{\mathcal{U}}_k = \mathcal{U}_k - \mathcal{U}_k^0$ with $\mathcal{U}_k^0$ the vector representation of the nominal policy. The ith entry of the input is expressed

$$\begin{aligned} \mathcal{U}_k^i &= \phi^0(u^i \mid s^i) + [-K^\star \widetilde{\mathcal{X}}_k + G^\star r]_i \\ &= \phi^0(u^i \mid s^i) \left[1 + \frac{1}{\phi^0(u^i \mid s^i)} \left(-[K^\star \widetilde{\mathcal{X}}_k]_i + G_i^\star r \right) \right]. \end{aligned}$$

This implies the small signal approximation (1.27). It is conjectured that (1.27) is within $O(r^2)$ of optimal (in terms of the objective in (1.26)).

3.2. Minimum principle approach. As previously observed, the optimization problem (1.26) falls outside of traditional MDP theory:

(i) The control cost is absent and is replaced by a cost on the randomized policy.

(ii) A quadratic cost on π appears, rather than linear as anticipated in the LP formulations of MDPs.

An MDP model is constructed here through a series of steps, with the first step addressing (i). For this it is natural to view the input as an element of the simplex $\mathcal{S}(\mathsf{U})$. This is not the same setting as section 3.1: in this subsection, the notation $\phi(\cdot \mid s)$ is interpreted as static state feedback from state s to input $\phi(\cdot \mid s) \in \mathcal{S}(\mathsf{U})$.

To remove the quadratic cost on π requires a Lagrangian relaxation, similar to what was used in section 2. For $\lambda \in \mathbb{R}$ denote

$$[\pi^\lambda, \phi^\lambda, \gamma^\lambda] = \arg\min_{\pi, \phi, \gamma} \left\{ \mathcal{D}(\phi) + \frac{\kappa}{2}\gamma^2 + \lambda[\gamma - \langle \pi, \mathcal{Y} \rangle + r] : \pi P_\phi = \pi \right\}. \tag{3.2}$$

For each λ this is viewed as a standard average-cost optimal control problem with state process $\boldsymbol{S}$. The controlled transition matrix and cost function are defined by

$$\begin{aligned} T_\mu(s, s') &:= \sum_u \mu(u) T((u, s), s'), && s, s' \in \mathsf{S}, \quad \mu \in \mathcal{S}(\mathsf{U}), \\ c(s, \mu) &:= \sum_u \mu(u) \left[\log\left(\frac{\mu(u)}{\phi^0(u \mid s)} \right) - \lambda \mathcal{Y}(s, u) \right], && s \in \mathsf{S}, \quad \mu \in \mathcal{S}(\mathsf{U}). \end{aligned}$$

Under any policy $\phi \in \Phi$ the resulting process $\boldsymbol{S}$ is Markovian. With a slight abuse of notation, its transition matrix is denoted

$$T_\phi(s, s') := \sum_u \phi(u \mid s) T((u, s), s')$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

and the cost as a function of s under this policy is denoted

$$c_\phi(s) = \sum_u \phi(u \mid s) c(s, \phi(u \mid s)) = \sum_u \phi(u \mid s) \left[\log\left(\frac{\phi(u \mid s)}{\phi^0(u \mid s)} \right) - \lambda \mathcal{Y}(s,u) \right], \quad s \in \mathsf{S}.$$

The solution to (3.2) gives $\gamma^\lambda = -\lambda/\kappa$ and

$$[\pi^\lambda, \phi^\lambda] = \arg\min_{\hat\nu, \phi} \left\{ \sum_s \hat\nu(s) c_\phi(s) : \hat\nu T_\phi = \hat\nu \right\}. \tag{3.3}$$

This is a standard MDP formulation, in which the optimization over feedback laws ϕ is explicit.

The Lagrange multiplier λ is treated as the independent parameter rather than r. This is justified through the correspondence $\gamma^\lambda = -\lambda/\kappa$, and the following definition imposes complementary slackness:

$$r^\lambda = -\gamma + \langle \pi^\lambda, \mathcal{Y} \rangle = \langle \pi^\lambda, \mathcal{Y} \rangle + \lambda/\kappa. \tag{3.4}$$

As λ ranges from $-\infty$ to $+\infty$, so do the values of r^λ because $\langle \pi^\lambda, \mathcal{Y} \rangle$ is bounded and continuous in λ.

Continuity of $\langle \pi^\lambda, \mathcal{Y} \rangle$ and other conclusions are obtained from prior research (in particular [9]), because the optimization problem (3.3) is identical to the IPD optimization problem (1.29), in which ζ is replaced by λ.

To match the setting of [9], denote the one-step reward as the negative of cost, $\varrho(s, \phi) = -c(s, \phi)$. Based on the foregoing, the solution to (3.3) is characterized by the average reward optimality equation

$$\xi^\lambda + h^\lambda(s) = \max_\phi \left\{ \varrho(s, \phi) + \sum_{s'} T_\phi(s, s') h^\lambda(s') \right\}. \tag{3.5}$$

The maximizer provides a representation for the optimal policy similar to (1.21a):

$$\phi^\lambda(u \mid s) = \phi^0(u \mid s) \exp\big(\bar h^\lambda(s,u) + \lambda \mathcal{Y}(s,u) - \Gamma^\lambda(s) \big), \tag{3.6}$$

with $\bar h^\lambda(x) = \sum_{u'} T(x, s') h^\lambda(s')$ and $\Gamma^\lambda(s)$ the normalizing factor,

$$\Gamma^\lambda(s) = \log \sum_u \phi^0(u \mid s) \exp\big(\bar h^\lambda(s,u) + \lambda \mathcal{Y}(s,u) \big). \tag{3.7}$$

ODE solution. The reader is referred to [9] for full details on this solution technique to compute the solution to (3.5). The main ideas are recalled here, in part because they are required in a small signal approximation.

It is shown in this prior work that the relative value functions can be constructed so that they are continuously differentiable in λ. Letting $H^\lambda = \frac{d}{d\lambda} h^\lambda$, we obtain

$$\overline{\mathcal{Y}}^\lambda + H^\lambda(s) = \mathcal{Y}_s + \sum_{s'} T_\lambda(s, s') H^\lambda(s'), \quad s \in \mathsf{S},\ \lambda \in \mathbb{R},$$

in which $T_\lambda = T_{\phi^\lambda}$,

$$\mathcal{Y}_s := \frac{d}{d\lambda} \varrho(s, \phi) = \sum_u \phi(u \mid s) \mathcal{Y}(s,u) \quad \text{and} \quad \overline{\mathcal{Y}}^\lambda = \frac{d}{d\lambda} \xi^\lambda. \tag{3.8}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

This fixed point equation is known as Poisson's equation, whose solution is often expressed $H^\lambda = Z_\lambda \mathcal{Y}$ with Z_λ known as the fundamental matrix (obtained as a simple matrix inverse). Also obtained is

$$\overline{\mathcal{Y}}^\lambda = \sum_x \pi^\lambda(x)\mathcal{Y}(x), \tag{3.9}$$

where $\pi^\lambda(s,u) = \hat{\nu}^\lambda(s)\phi^\lambda(u \mid s)$, with $\hat{\nu}^\lambda$ the unique invariant pmf for T_λ.

This defines the ODE solution for the family of relative value functions

$$\frac{d}{d\lambda} h_\lambda = Z_\lambda \mathcal{Y} \tag{3.10}$$

with boundary condition $h^\lambda \equiv 0$ when $\lambda = 0$. The right-hand side depends on h_λ through Z_λ, but the dependency is smooth.

Small signal approximation. The small signal approximation here is defined in a setting similar to Proposition 3.3: it is assumed that the reference signal is small in magnitude, and that $r \equiv 0$ achieves zero cost in (1.26). This holds if $\sum \pi^0(x)\mathcal{Y}(x) = 0$, which will be assumed henceforth.

A slight change in notation is required here, as compared to section 3.1: $\widetilde{X}_k$ and $\widetilde{U}_k$ are n-dimensional column vectors that denote the exact deviation, $\widetilde{X}^i_k := \widetilde{\nu}_k(x^i)$ and $\widetilde{U}^i_k = \widetilde{\phi}_k(u^i \mid s^i)$ for each i and k. The approximation requires the following notation:

(i) $\varsigma^2_\lambda = \frac{d^2}{d\lambda^2}\xi^\lambda$ for $\lambda \in \mathbb{R}$.

(ii) $\bar{H}^\lambda = \frac{d}{d\lambda}\bar{h}^\lambda$, $\bar{H}^\lambda_s = \sum_u \phi^0(u \mid s)\bar{H}^\lambda(s,u)$, $\mathcal{Y}_s = \sum_u \phi^0(u \mid s)\mathcal{Y}(s,u)$.

(iii) $\Lambda(x) = \bar{H}^0(x) + \mathcal{Y}(x) - (\bar{H}^0_s + \mathcal{Y}_s)$, $\Lambda_{\mathsf{n}}(x) = \big(\varsigma_0^2 + 1/\kappa\big)^{-1}\Lambda(x)$.

Approximation of the state dynamics begins with an approximation of the input. The proof of Lemma 3.4 is postponed to Appendix C.

LEMMA 3.4. *The small-r approximation holds for the solution to IPD-Q:*

$$\widetilde{\phi}(u \mid s, r) = \phi^0(u \mid s)\exp\big(\Lambda_{\mathsf{n}}(x)r\big) + O(r^2). \tag{3.11}$$

The following linear systems approximation follows easily from Lemma 3.4. We defer to [13] for details of the proof.

PROPOSITION 3.5. *Suppose that the input $\phi_k(u \mid s) = \phi^\star(u \mid s, r_k)$ is applied to the nonlinear system* (1.22)*. The closed loop dynamics then admit the approximation*

$$\widetilde{X}_{k+1} = A\widetilde{X}_k + BG^\star r_k + \varepsilon_k + O(r_k^2) \tag{3.12}$$

in which A and B are defined in Proposition 3.3*, $G^\star$ is the column vector with entries $G^\star_i = \phi^0(u^i \mid s^i)\Lambda_{\mathsf{n}}(x^i)$, and ε_k is quadratic in the deviation $(\widetilde{\nu}_k, \widetilde{\phi}_k)$:*

$$\varepsilon^i_k = \sum_j \widetilde{\nu}_k(x^j)T(x^j, s^i)\widetilde{\phi}_k(u^i \mid s^i)\,, \qquad x^i = (s^i, u^i) \in \mathsf{X}.$$

4. Applications to demand dispatch. An application of the control framework described in the previous sections is *demand dispatch,* an evolving science for automatically controlling flexible loads to help maintain supply-demand balance in the power grid. The goal of demand dispatch is to modify the behavior of flexible loads such that the aggregate power consumption tracks a reference signal that is broadcast by a BA.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Keep in mind that in the numerical examples here we focus entirely on the mean-field model. We know from prior work that evolution of the empirical distributions does closely track this idealization: for reasonably large $\mathcal{N}$, following the notation (1.1b), the approximation $\nu_k^{\mathcal{N}} \approx \nu_k$ holds and the covariance of the error grows slowly with k (error is reduced with feedback [19, 20]). Although the control architecture in this prior work is very different, it should not surprise the reader that the law of large numbers and associated central limit theorem hold in the setting of this paper.

Also, the numerical results here focus entirely on the solutions surveyed in section 2. As explained in section 3, the IPD-Q solution for real-time feedback reduces to something similar to what has been extensively explored in prior work [19, 20].

Although these techniques can be applied to any flexible load, the experiments in this section demonstrate distributed control of a population of residential water heaters or refrigerators. An MDP model is constructed in which the state is the standard used to capture hysteresis control, $S_k = (\theta_k, U_{k-1})$, in which $\theta_k \in \mathbb{R}$ is the temperature, and $U_k \in \{0, 1\}$ denotes power mode for each k. Remember the physical system operates in continuous time, and k represents the kth sampling time. This means that U_{k-1} represents the power mode during the sampling interval ending at the kth sampling time.

4.1. Designing the nominal model. Construction of the nominal model with transition matrices $\{P_k^0\}$ of the form (1.10b) requires specification of dynamics of nature and the nominal policy. In the case of water heaters, the sequence of transition matrices $\{T_k\}$ for nature were based on input-output data obtained from Oak Ridge National Laboratories [22]. For refrigerators, T was taken independent of k, constructed based on simulations of the standard linear TCL model,

$$\theta_{k+1} = \theta_k + \alpha[\theta^a - \theta_k] - \beta U_k + D_{k+1}\,, \tag{4.1}$$

in which $\alpha, \beta > 0$, θ^a denotes the (time-invariant) ambient air temperature, and the disturbance process $\boldsymbol{D}$ captures modeling error and usage.

In all cases the nominal policy was chosen time-homogeneous: $\phi_k^0 \equiv \phi^0$ is a fixed randomized policy, designed to approximate deterministic hysteresis control. We describe the construction for water heaters, following [37, 22]. We defer to [13] for details on the construction of the nominal policy.

4.2. Tracking. In practical applications the aggregate power is of interest, which is approximated by $\varrho \mathcal{N} y_k$ at time k, where ϱ is the rated power of a single load. Hence the total population size $\mathcal{N}$ must be taken into account in any tracking problem. In plots that follow, we choose to focus on the "normalized" response, defined as follows:

$$y_k^{\text{ref}} = r_k/\varrho\,, \qquad \hat{y}_k^{\text{ref}} = \hat{r}_k/\varrho\,, \qquad y_k = \langle \nu_k, \mathcal{Y} \rangle / \varrho.$$

In this context, y_k can be interpreted as the probability of a load being on.

The two sets of plots in Figure 1 are distinguished by the reference signal. In each case the reference signal is a square wave. In (a) the signal is feasible, and in (b) it violates the energy limits of the collection of water heaters [27]. In Figure 1(a) it is seen that tracking is nearly perfect for sufficiently large κ. Tracking of the larger reference signal would require temperature deviations to exceed the deadband of the water heater. Instead, we observe in Figure 1(b) a graceful truncation of the reference signal.

The next set of experiments was designed to assess sensitivity of KLQ optimal control to modeling error. Specifically, what are the consequences of ignoring the randomness of nature?

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

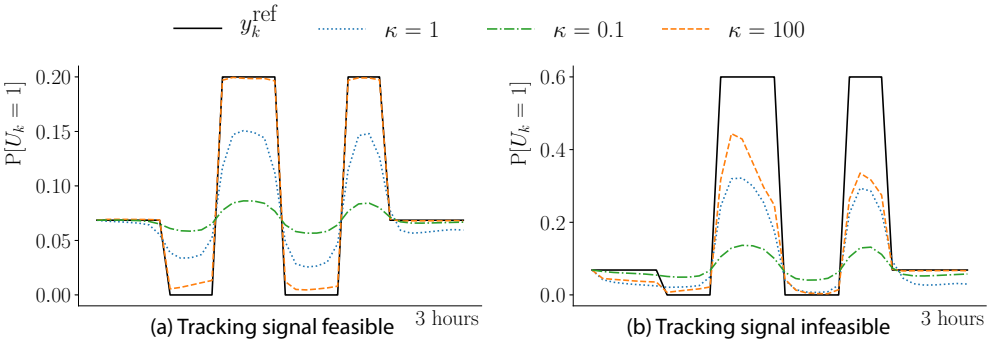

FIG. 1. *Tracking error:* (a) *reference signal is feasible;* (b) *reference signal is infeasible.*

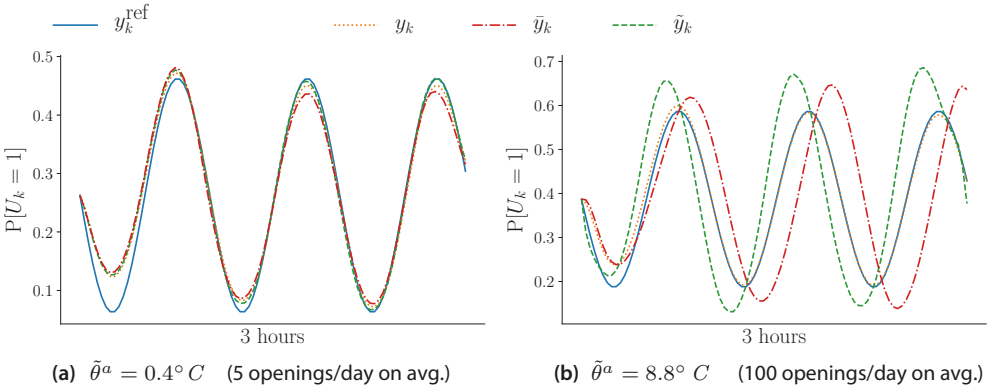

FIG. 2. *Sensitivity experiments support the use of MPC with a deterministic approximation for randomness from nature.*

A particular choice of statistics for (4.1) was chosen in order to mimic the effect of a refrigerator door opening at random times throughout the day: $\boldsymbol{D}$ is i.i.d., with

$$D_{k+1} = \begin{cases} \bar{d} & \text{with probability} \quad \varepsilon, \\ 0 & \text{with probability} \quad 1-\varepsilon, \end{cases}$$

where ε determines the average amount of door openings per day, and $\bar{d}$ was chosen so that the temperature inside the refrigerator increases when the door is open even when the power mode is on. A deterministic approximation of (4.1) was constructed for comparison, in which D_{k+1} is replaced by its mean:

$$\theta_{k+1} = \theta_k + \alpha[\bar{\theta}^a - \theta_k] - \beta U_k \tag{4.2}$$

with $\bar{\theta}^a = \theta^a + \tilde{\theta}^a$ with $\tilde{\theta}^a = \bar{d}\varepsilon/\alpha$.

Optimal policies were calculated for each of three models: the stochastic model (4.1), its deterministic approximation (4.2), and the cruder deterministic approximation obtained on setting $D_{k+1} \equiv 0$ in (4.1) (equivalently, (4.2) with $\bar{\theta}^a = \theta^a$). Each policy was then tested on the stochastic model (4.1).

Figure 2 displays the results from these experiments, where in each plot

- y_k^{ref} is the reference signal,
- y_k is the policy that is optimal for the stochastic model,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

- $\tilde{y}_k$ is the policy that is optimal for (4.2),
- $\bar{y}_k$ is the policy that is optimal for (4.2) using $\bar{\theta}^a = \theta^a$.

The accurate tracking $y_k \approx y_k^{\text{ref}}$ is expected because this reference signal is feasible, and $\kappa > 0$ was chosen to be large.

It is seen in Figure 2(a) that all four trajectories are nearly identical for the smaller disturbance. The deviation is far greater in (b), for which the disturbance is greater. However, y_k and $\bar{y}_k$ are nearly identical for about the first 30 minutes. This suggests that a deterministic approximation, combined with MPC, may be used in place of the stochastic model.

4.3. Information architectures. The choice of information architecture is an interesting topic for future research. Here are three possibilities:

(i) *Smart BA*: The BA uses the reference signal $\{r_k\}$ and its estimate of ν_0^0 to compute $\lambda^\star$ and broadcast it to the loads.

(ii) *Smart load*: The BA broadcasts $\{r_k\}$ to the loads. Each load computes $\lambda^\star$ based on its internal model and $\nu_0^0 = \delta_{x_0}$, with $x_0 \in \mathsf{X}$ its current state.

(iii) *Genius load*: The BA broadcasts $\{r_k\}$ to the loads. Each load computes $\lambda^\star$ based on its internal model and its estimate of ν_0^0.

Each approach has its strengths and weaknesses. Approaches (i) and (iii) require knowledge of the initial marginal pmf of the population, ν_0^0. If a perfect estimate is assumed, then the total cost in cases (i) and (iii) is equal to $J^\star(\nu_0^0)$. But, how can a load estimate the marginal pmf of the population? Numerical results from [13] suggest coupling of the marginals from distinct initial conditions. If enough time has passed since the latest MPC iteration, the pmfs $\{\nu_k\}$ computed locally can be used to approximate the marginal pmf of the population (perhaps smoothed using the techniques of [19, 20]).

In contrast, the total cost for case (ii) is the sum, $\sum_{i=1}^d \nu_0^0(x^i) J^\star(\delta_{x^i})$, since each load optimizes according to its own initial state, x^i. Even when the aggregate can easily track $\{r_k\}$, the cost $J^\star(\delta_{x^i})$ may be very large for individuals that are at odds with the reference signal. For example, an increase in power consumption could be requested while a water heater is near its upper temperature limit and must turn off. So, it is possible that approach (ii) will impose greater stress on the loads as compared to the other two options, or will lead to reduced capacity.

5. Conclusions. The paper provides a complete theory for KLQ and infinite-horizon counterparts, without the restriction to deterministic dynamics imposed in [15, 23]. Plans for future research include the following:

(i) Monte Carlo approaches for both KLQ and IPD-Q. The approximation (1.27) invites actor critic methods for approximating the best coefficients $\{K_{i,j}^\star, G_i^\star\}$ based on training data with nonconstant reference signal, rather than approximation.

(ii) Evaluate robustness and sensitivity to other types of modeling error.

(iii) Investigate alternative transform techniques.

(iv) Consider other cost functions, such as the Wasserstein distance. Preliminary results are summarized in [24].

(v) Investigate the relationship between optimality and coupling of the pmfs, and the implications to control design.

(vi) Careful design of a terminal cost function may result in better performance for smaller time horizons [18].

(vii) How is the relative value function $\mathcal{H}$ appearing in (3.1) related to h^λ appearing in (3.5) (with $\lambda = \lambda_r$)?

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Appendix A. Convexity. The proofs of Theorem 2.1 and Proposition 2.2 make use of the following four lemmas. The first is based on a well-known result regarding relative entropy. For any function $h\colon \mathsf{X}^{K+1} \to \mathbb{R}$ denote $\Lambda^0(h) := \sup_p \{\langle p, h\rangle - D(p\|p^0)\}$.

LEMMA A.1 (convex dual of relative entropy). *For each $p^0 \in \mathcal{S}(\mathsf{X}^{K+1})$ and function $h\colon \mathsf{X}^{K+1} \to \mathbb{R}$, the (possibly infinite) value of $\Lambda^0(h)$ coincides with the log moment generating function: $\Lambda^0(h) = \log\langle p^0, e^h\rangle$. Moreover, provided $\Lambda^0(h) < \infty$, the supremum defining this quantity is uniquely attained with $p^\star = p^0 \exp(h - \Lambda^0(h))$. That is, the log-likelihood $L^\star = \log(dp^\star/dp^0)$ is given by $L^\star(\vec{x}) = h(\vec{x}) - \Lambda^0(h)$.*

LEMMA A.2. *The dual function can be expressed*

$$\varphi^\star(\lambda, g) = \lambda^T \hat{r} - \frac{1}{2\kappa}\|\lambda\|^2 - \langle \nu_0^0, G_1^\lambda\rangle + \sum_{k=1}^K \min_s \Big[g_k(s) - \mathcal{T}_k^\lambda(g_{k+1}; s)\Big]. \tag{A.1}$$

Proof. First, make the substitution $\nu_k(s,u) = \hat{\nu}_k(s)\phi_k(u \mid s)$, so that the Lagrangian (2.3) can be written

$$\begin{aligned}\mathcal{L}(\nu,\gamma,\lambda,g) &= \sum_{n=1}^N \Big(\frac{\kappa}{2}\gamma_n^2 + \lambda_n\gamma_n + \lambda_n\hat{r}_n\Big) - \sum_{s,u}\nu_0^0(s,u)\sum_{s'}T_0(x,s')g_1(s') \\ &+ \sum_{k=1}^K \sum_s \hat{\nu}_k(s)\sum_u \phi_k(u\mid s)\left(L_k(s,u) - \sum_{s'}T_k(x,s')g_{k+1}(s') - \check{\lambda}_k\mathcal{Y}(s,u)\right) \\ &+ \sum_{k=1}^K\sum_s \hat{\nu}_k(s)g_k(s)\end{aligned} \tag{A.2}$$

with $g_{K+1} \equiv 0$, and $L_k(s,u) = \log\frac{\phi_k(u|s)}{\phi_k^0(u|s)}$. This amounts to a Lagrangian decomposition since the minimization of the Lagrangian is equivalent to solving K separate convex programs to obtain each of the minimizers $\{\nu_k^{\lambda,g} : \nu_k^{\lambda,g}(s,u) = \hat{\nu}_k^{\lambda,g}(s)\phi_k^{\lambda,g}(u \mid s), (s,u) \in \mathsf{X}, 1 \le k \le K\}$. That is, $\arg\min_\phi \mathcal{L} =$

$$\left\{\operatorname*{arg\,min}_{\phi_k : 1\le k\le K} \sum_u \phi_k(u\mid s)\Big[L_k(s,u) - \sum_{s'}T_k(x,s')g_{k+1}(s') - \check{\lambda}_k\mathcal{Y}(s,u)\Big]\right\}. \tag{A.3}$$

Lemma A.1 implies that the minimizer is obtained with $\Lambda_k(s) = \mathcal{T}_k^\lambda(g_{k+1}; s)$ and

$$\phi_k^{\lambda,g}(u\mid s) = \phi_k^0(u\mid s)\exp\left(\sum_{s'}T_k(x,s')g_{k+1}(s') + \check{\lambda}_k\mathcal{Y}(s,u) - \Lambda_k(s)\right). \tag{A.4}$$

Lemma A.1 also gives the value

$$\min_{\phi_k}\sum_u \phi_k(u\mid s)\left[L_k(s,u) - \sum_{s'}T_k(x,s')g_{k+1}(s') - \check{\lambda}_k\mathcal{Y}(s,u)\right] = -\mathcal{T}_k^\lambda(g_{k+1}; s)$$

resulting in

$$\begin{aligned}\min_\nu \mathcal{L}(\nu,\gamma,\lambda,g) &= \sum_{n=1}^N\Big(\frac{\kappa}{2}\gamma_n^2 + \lambda_n\gamma_n + \lambda_n\hat{r}_n\Big) - \sum_{s,u}\nu_0^0(s,u)\sum_{s'}T_0(x,s')g_1(s') \\ &+ \sum_{k=1}^K \min_{\hat{\nu}_k}\langle \hat{\nu}_k, g_k - \mathcal{T}_k^\lambda(g_{k+1})\rangle.\end{aligned} \tag{A.5}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Next, observe that the minimizer $\hat{\nu}_k^{\lambda,g}$ is obtained when the support of each $\hat{\nu}_k$ satisfies

$$\operatorname{supp}\big(\hat{\nu}_k(s)\big) \subseteq \arg\min_s \Big[g_k(s) - \mathcal{T}_k^\lambda(g_{k+1};s)\Big]$$
$$\text{so that} \quad \min_s \Big[g_k(s) - \mathcal{T}_k^\lambda(g_{k+1};s)\Big] = \langle\, \hat{\nu}_k^{\lambda,g}, g_k - \mathcal{T}_k^\lambda(g_{k+1})\,\rangle.$$

We also have $\gamma_n^\lambda = -\frac{1}{\kappa}\lambda_n$ Substituting the minimizers $\{\nu_k^{\lambda,g}, \gamma_n^\lambda\}$ into (A.5), and applying (2.7), results in (A.1). ∎

Appendix B. Duality.

LEMMA B.1. *The maximum of the dual function over g is*

$$\varphi^\star(\lambda) := \max_g \varphi^\star(\lambda, g) = \lambda^T \hat{r} - \frac{1}{2\kappa}\|\lambda\|^2 - \langle\, \nu_0^0, G_1^\lambda\,\rangle \tag{B.1}$$

with $G_1^\lambda(x) = \sum_{s'} T_0(x,s') g_1^\lambda(s')$. A maximizer g^λ is given by the recursive formula:

$$g_k^\lambda = \mathcal{T}_k^\lambda(g_{k+1}^\lambda)\,,\ 1 \le k \le K\,,\ \textit{where } g_{K+1}^\lambda \equiv 0. \tag{B.2}$$

Proof. Adding a constant to any of the $(g_1, g_2, \dots, g_K)$ does not change the value of $\mathcal{L}$ (2.3) or $\varphi^\star$ (2.8), so without loss of generality we assume

$$\min_s \Big[g_k(s) - \mathcal{T}_k^\lambda(g_{k+1};s)\Big] = 0 \text{ for each } k, \tag{B.3}$$

and consequently

$$g_k \ge \mathcal{T}_k^\lambda(g_{k+1}) \text{ for each } k\,. \tag{B.4}$$

Thus, in view of (A.1),

$$\varphi^\star(\lambda) = \lambda^T \hat{r} - \frac{1}{2\kappa}\|\lambda\|^2 - \min_{g_1} \sum_{s,u} \nu_0^0(s,u) \sum_{s'} T_0(x,s') g_1(s')\,, \tag{B.5}$$

where the minimum is subject to the constraint (B.4). Next, observe that $\mathcal{T}_k^\lambda$ is a monotone operator, so that for each $k \le K$,

$$g_k \ge \mathcal{T}_k^\lambda \circ \mathcal{T}_{k+1}^\lambda \circ \cdots \circ \mathcal{T}_K^\lambda(g_{K+1}) \doteq g_k^\lambda\,, \quad \text{where } g_{K+1} \equiv 0.$$

Based on the expression (B.5), we now show that the maximum $\arg\max_g \Phi^\star(\lambda, g)$ is obtained by choosing each g_k to reach this lower bound, giving (B.2). Indeed, g_1^λ achieves the minimum in (B.5), since $g_1^\lambda \le g_1$ for any g_1 for which (B.4) holds. This result along with (B.3) yields (B.1). ∎

For an inductive proof of the following see [13].

LEMMA B.2. *The maximizers $\{g_k^\lambda\}$ have at most linear growth in $\|\lambda\|$:*

$$|g_k^\lambda(s)| \le C_k \|\lambda\|\,, \quad 1 \le k \le K, \tag{B.6}$$

where $C_k = \|\mathcal{Y}\|_\infty \sum_{i=k}^K \|w(i)\|$ and $w(i)$ is the vector $\{w_1(i), w_2(i), \dots, w_N(i)\}$.

Proof of Theorem 2.1. We prove the existence of a maximizer $\lambda^\star$ by showing that $\Phi^\star(\lambda)$ is an anticoercive function, i.e., $\Phi^\star(\lambda) \to -\infty$ as $\|\lambda\| \to \infty$. By Lemma B.2, there exists $C_1 < \infty$ such that

$$\varphi^\star(\lambda) = \lambda^T \hat{r} - \frac{1}{2\kappa}\|\lambda\|^2 - \sum_{s,u} \nu_0^0(s,u) \sum_{s'} T_0(x,s') g_1^\lambda(s')$$
$$\le \|\lambda\|\|\hat{r}\| - \frac{1}{2\kappa}\|\lambda\|^2 + \max_{s'} |g_1^\lambda(s')| \le \|\lambda\|\|\hat{r}\| - \frac{1}{2\kappa}\|\lambda\|^2 + C_1\|\lambda\|.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Since $\phi^\star(\lambda)$ is upper-bounded by an anticoercive function, $\phi^\star(\lambda)$ itself is an anticoercive function. Thus a maximizer $\lambda^\star$ exists, and $(\lambda^\star, g^\star) = (\lambda^\star, g^{\lambda^\star})$ by (B.2).

The primal is a convex program, as established in Proposition 1.1. To show that there is no duality gap it is sufficient that Slater's condition holds [6, section 5.3.2]. This condition holds: the relative interior of the constraint set for the primal is nonempty since it contains $\{\nu_k^0\}$. Optimality of (2.6) is established by substituting $g_{k+1}^\star$ into (A.4) and by making the substitution $g_k^\star = \mathcal{T}_k^\lambda(g_{k+1}^\star)$ implied by (B.2). ∎

Proof of Proposition 2.2. This proof has three parts:

(i) Equation (2.5) is proven by Lemma B.1.

(ii) Equation (2.8) is proven by Lemma B.1.

(iii) The representation of the derivative in part (iii) is standard (e.g., section 5.6 of [6], or [13]). ∎

Appendix C. IPD-Q.

Proof of Lemma 3.4. An application of the implicit function theorem tells us that $\{r^\lambda, \phi^\lambda : \lambda \in \mathbb{R}\}$ are smooth as functions of λ, whose derivatives may be expressed

$$\frac{d}{d\lambda} r^\lambda = \frac{d}{d\lambda} \langle \pi^\lambda, \mathcal{Y} \rangle + 1/\kappa = \varsigma_\lambda^2 + 1/\kappa,$$
$$\frac{d}{d\lambda} \log\big(\phi^\lambda(u \mid s)\big) = \bar{H}^\lambda(x) + \mathcal{Y}(x) - \frac{d}{d\lambda}\Gamma^\lambda(s).$$

The first identities follow from (3.4) and then (3.8). The formula for the derivative of $\log(\phi^\lambda)$ is immediate from (3.6).

The proof of (3.11) requires approximations for r^λ and ϕ^λ in a neighborhood of zero. The first approximation is given by $r^\lambda = \big(\varsigma_0^2 + 1/\kappa\big)\lambda + O(\lambda^2)$. The definition (3.7) implies that

$$\frac{d}{d\lambda}\Gamma^\lambda(s)\big|_{\lambda=0} = \bar{H}_s^0 + \mathcal{Y}_s,$$
$$\text{which gives} \quad \log\big(\phi^\lambda(u \mid s)\big) = \log\big(\phi^0(u \mid s)\big) + \Lambda(x)\lambda + O(\lambda^2).$$

An inversion is applied to express λ as a function of r, giving

$$\phi(u \mid s, r) = \phi^0(u \mid s) \exp\big(\Lambda(x)\lambda_r\big) + O(r^2)$$

with $\lambda_r = \big(\varsigma_0^2 + 1/\kappa\big)^{-1} r + O(r^2)$. Hence (3.11) follows from a first-order Taylor series approximation of the exponential. ∎

REFERENCES

[1] M. Almassalkhi, J. Frolik, and P. Hines, *Packetized energy management: Asynchronous and anonymous coordination of thermostatically controlled loads*, in Proceedings of the American Control Conference, 2017, pp. 1431–1437.

[2] E. Altman, *Constrained Markov Decision Processes*, Chapman & Hall/CRC, Boca Raton, FL, 1999.

[3] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[4] 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 Proceedings of the IEEE Conference on Decision and Control, 2019, pp. 4189–4194.

[5] V. S. Borkar, *Convex analytic methods in Markov decision processes*, in Handbook of Markov Decision Processes, Internat. Ser. Oper. Res. Management Sci. 40, Kluwer, Boston, 2002, pp. 347–375.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, 2004.
[7] A. Brooks, E. Lu, D. Reicher, C. Spirakis, and B. Weihl, *Demand dispatch*, IEEE Power Energy Mag., 8 (2010), pp. 20–29.
[8] A. Bušić and S. Meyn, *Distributed randomized control for demand dispatch*, in Proceedings of the Conference on Decision and Control, 2016, pp. 6964–6971.
[9] 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 (2018), pp. 343–366.
[10] P. E. Caines, *Mean field games*, in Encyclopedia of Systems and Control, J. Baillieul and T. Samad, eds., Springer, London, 2021, pp. 1197–1202.
[11] A. Bušić, S. Meyn, and N. Cammardella, *Learning optimal policies in mean-field models with Kullback–Leibler regularization*, in Proceedings of the IEEE Conference on Decision and Control, 2023.
[12] N. Cammardella, *Creating Virtual Energy Storage through Optimal Allocation and Control of Flexible Power Consumption*, Ph.D. thesis, University of Florida, Gainesville, 2021.
[13] N. Cammardella, A. Bušić, and S. Meyn, *Kullback-Leibler-Quadratic Optimal Control*, https://arxiv.org/abs/2004.01798, 2020.
[14] N. Cammardella, A. Bušić, and S. Meyn, *Simultaneous allocation and control of distributed energy resources via Kullback-Leibler-quadratic optimal control*, in Proceedings of the American Control Conference, 2020, pp. 514–520.
[15] N. Cammardella, A. Bušić, Y. Ji, and S. Meyn, *Kullback-Leibler-quadratic optimal control of flexible power demand*, in Proceedings of the Conference on Decision and Control, 2019, pp. 4195–4201.
[16] R. Carmona and F. Delarue, *Probabilistic Theory of Mean Field Games with Applications* I*: Mean Field FBSDEs, Control, and Games, Probability Theory and Stochastic Modelling*, Springer, New York, 2018.
[17] R. Carmona and F. Delarue, *Probabilistic Theory of Mean Field Games with Applications* II*: Mean Field Games with Common Noise and Master Equations, Probability Theory and Stochastic Modelling*, Springer, New York, 2018.
[18] R.-R. Chen and S. P. Meyn, *Value iteration and optimization of multiclass queueing networks*, Queueing Syst., 32 (1999), pp. 65–97.
[19] Y. Chen, *Markovian Demand Dispatch Design for Virtual Energy Storage to Support Renewable Energy Integration*, Ph.D. thesis, University of Florida, Gainesville, 2016.
[20] Y. Chen, A. Bušić, and S. Meyn, *State estimation for the individual and the population in mean-field control with application to demand dispatch*, IEEE Trans. Automat. Control, 62 (2017), pp. 1138–1149.
[21] Y. Chen, T. T. Georgiou, and M. Pavon, *Optimal transport in systems and control*, Annu. Rev. Control Robot. Auton. Syst., 4 (2020), pp. 89–113.
[22] Y. Chen, M. U. Hashmi, J. Mathias, A. Bušić, and S. Meyn, *Distributed control design for balancing the grid using flexible loads*, in Energy Markets and Responsive Grids: Modeling, Control, and Optimization, S. Meyn, T. Samad, I. Hiskens, and J. Stoustrup, eds., Springer, New York, 2018, pp. 383–411.
[23] M. Chertkov and V. Y. Chernyak, *Ensemble control of cycling energy loads: Markov decision approach*, in Energy Markets and Responsive Grids: Modeling, Control, and Optimization, IMA Vol. Control Energy Markets Grids 162, Springer, New York, 2018.
[24] T. L. Corre and S. M. Ana Bušić, *Feature Projection for Optimal Transport*, https://arxiv.org/abs/2208.01958, 2023.
[25] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed., Springer, New York, 1998.
[26] O. Guéant, J.-M. Lasry, and P.-L. Lions, *Mean Field Games and Applications*, Springer, Berlin, 2011, pp. 205–266.
[27] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, *Aggregate flexibility of thermostatically controlled loads*, IEEE Trans. Power Syst., 30 (2015), pp. 189–198.
[28] M. Huang, P. E. Caines, and R. P. Malhame, *Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ε-Nash equilibria*, IEEE Trans. Automat. Control, 52 (2007), pp. 1560–1571.
[29] M. Huang, R. P. Malhame, and P. E. Caines, *Large population stochastic dynamic games: Closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle*, Commun. Inf. Syst., 6 (2006), pp. 221–251.
[30] I. Kontoyiannis and S. P. Meyn, *Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes*, Electron. J. Probab., 10 (2005), pp. 61–123.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[31] J. M. Lasry and P. L. Lions, *Mean field games*, Jpn. J. Math., 2 (2007), pp. 229–260.
[32] J.-S. Li, *Ensemble control of finite-dimensional time-varying linear systems*, IEEE Trans. Automat. Control, 56 (2010), pp. 345–357.
[33] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969, reprinted 1997.
[34] A. S. Manne, *Linear programming and sequential decisions*, Management Sci., 6 (1960), pp. 259–267.
[35] J. Mathieu, S. Koch, and D. Callaway, *State estimation and control of electric loads to manage real-time energy imbalance*, IEEE Trans. Power Syst., 28 (2013), pp. 430–440.
[36] S. Meyn, *Control Systems and Reinforcement Learning*, Cambridge University Press, Cambridge, UK, 2022.
[37] 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 (2015), pp. 2847–2862.
[38] N. Parikh and S. Boyd, *Proximal Algorithms, Foundations and Trends in Optimization*, Now Publishers, 2013.
[39] G. Peyré and M. Cuturi, *Computational Optimal Transport*, https://arxiv.org/abs/1803.00567, 2020.
[40] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, Springer, New York, 2010.
[41] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, New York, 2014.
[42] A. Taghvaei and P. G. Mehta, *A survey of feedback particle filter and related controlled interacting particle systems (CIPS)*, Annu. Rev. Control, 55 (2023), pp. 356–378, https://doi.org/10.1016/j.arcontrol.2023.03.006.
[43] S. H. Tindemans, V. Trovato, and G. Strbac, *Decentralized control of thermostatic loads for flexible demand response*, IEEE Trans. Control Syst. Technol., 23 (2015), pp. 1685–1700.
[44] E. Todorov, *Linearly-solvable Markov decision problems*, in Proceedings of Advances in Neural Information Processing Systems, B. Schölkopf, J. Platt, and T. Hoffman, eds., Cambridge, MA, 2007, pp. 1369–1376.

Downloaded 11/30/23 to 128.227.227.177 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.