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

Convex Q-Learning in Continuous Time with Application to Dispatch of Distributed Energy Resources

Fan Lu, Joel Mathias, Sean Meyn*, and Karanjit Kalsi

***Abstract*— Convex Q-learning is a recent approach to reinforcement learning, motivated by the possibility of a firmer theory for convergence, and the possibility of making use of greater a priori knowledge regarding policy or value function structure. This paper explores algorithm design in the continuous time domain, with a finite-horizon optimal control objective. The main contributions are**

(i) The new *Q-ODE*: a model-free characterization of the Hamilton-Jacobi-Bellman equation.

(ii) A formulation of Convex Q-learning that avoids approximations appearing in prior work. The Bellman error used in the algorithm is defined by filtered measurements, which is necessary in the presence of measurement noise.

(iii) Convex Q-learning with linear function approximation is a convex program. It is shown that the constraint region is bounded, subject to an exploration condition on the training input.

(iv) The theory is illustrated in application to resource allocation for distributed energy resources, for which the theory is ideally suited.

I. Introduction

This paper concerns control techniques for the nonlinear state space model

$$\tfrac{d}{dt}x_t = F(x_t, u_t, t)\,, \qquad x_0 \in \mathbb{R}^n\,, \tag{1}$$

in which the state process x and input process u evolve on n and m dimensional Euclidean space, respectively. The goal is to approximate the solution to the finite time-horizon optimal control problem: with time horizon $\mathcal{T} > 0$, cost function $c\colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}_+$, and terminal cost $J_0\colon \mathbb{R}^n \to \mathbb{R}_+$, the objective to be minimized is

$$J(x) = \int_0^{\mathcal{T}} c(x_t, u_t, t)\, dt + J_0(x_{\mathcal{T}})\,, \quad x_0 = x\,, \tag{2}$$

whose infimum over all continuous inputs is denoted $J^*(x)$. It is known that the infimum is a minimum under general conditions, and the optimal input is state feedback, $u_t^* = \phi^*(x_t^*, t)$ for $0 \le t \le \mathcal{T}$.

*S. Meyn is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA (e-mail: meyn@ece.ufl.edu).

F. Lu is with the Department of Applied Mathmatics and Statistics, University of California, Santa Cruz, CA 95064, USA (e-mail: flu16@ucsc.edu).

J. Mathias is with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ, 85281, USA (e-mail: Joel.Mathias@asu.edu).

K. Kalsi is with the Pacific Northwest National Laboratory, Richland, WA 99354, USA (e-mail: karanjit.kalsi@pnnl.gov).

SM acknowledges support from ARO award W911NF2010055, National Science Foundation award EPCN 1935389, and from an Inria International Chair, Paris, France.

In this paper, techniques to approximate the optimal policy are proposed based on recent approaches to reinforcement learning (RL), inspired by Manne's linear programming approach to optimal control. One approach is known as *convex Q-learning* [13], [20], [14], and a dual approach is known as *logistic Q-learning* [1], [22] (developed so far only for stochastic control). A starting point is the well-known sample path bound implied by the HJB equation:

$$c(x_t, u_t, t) + \tfrac{d}{dt}J^*(x_t, t) \ge 0\,, \qquad 0 \le t \le \mathcal{T}\,. \tag{3}$$

where $J^*(x_t, t)$ is the *cost to go* (see Section II). This holds for any input-state sequence, and is tight, in the sense that the lower bound is achieved for any t for which $u_t = \phi^*(x_t, t)$.

The inequality (3) could be used in the formulation of a Q-learning algorithm based on linear programming techniques, following [13], [16], [20], [14].

One challenge is that a direct extension of this prior work for models in discrete time is not practical, as it would involve use of derivatives of measurements. Filtering techniques are introduced to address this challenge. *Control design is formulated in continuous time precisely so that these practical challenges are most clearly evident.*

Contributions **(i)** The inequality (3) is refined to define the *Q-ODE* (10), a model-free characterization of the HJB equation that lends itself to RL algorithm design.

(ii) The Q-ODE and the new bound presented in Prop. 2.1 lead to the new formulation of convex Q-learning.

(iii) Convex Q-learning is a convex program when the function class is linearly parameterized. It is shown in Prop. 3.2 that the constraint region for this convex program is bounded, subject to an exploration condition on the input used for training.

(iv) The algorithms described in Section III may be trained using a simulator, or based on data collected from experiments on the physical control system. In practice however we often have a model, and we may obtain a better control solution by making use of this information. Section IV proposes a marriage of convex Q-learning with MPC, and shows that it provides efficient solutions to complex control problems found in power systems applications.

Results from numerical experiments in Section IV bring many insights: a) convex Q-learning can be adapted to impose structure on the value function—for example, we might know that the value function is convex; b) for the example considered, prior knowledge beyond convexity leads to insight on how to choose a basis for value function approximation; c) the experiments go beyond theory, using

training data collected from perturbed models in the hopes of improving robustness of the resulting control law. In the experiments considered, it is found that this approach is successful, motivating future research on robust RL.

Related Research Much of this section is taken from the first author's dissertation [11]. The control problem formulated in Section IV and the solution structure used to inform the choice of the basis is taken from the second author's dissertation [17].

Convex Q-learning is a recent technique in RL. The article [19] provided foundations for this approach for continuous time models, with infinite-horizon objective, for which a result similar to Prop. 2.1 was obtained. This prior work does not lead to a practical characterization of the HJB equation, since it requires a stationary realization of the input-state process on the two-sided time interval $\{(x_t, u_t) : -\infty < t < \infty\}$. The introduction of practical algorithms came only recently in [1], [13], [14], [1], [22] (see [20, Ch. 5] for more history, and [10] for a history of RL in continuous time). There is a weak connection with the Lyapunov function approach to control design for nonlinear control systems [4]. The recent work [16] introduces an interesting bridge between linear-programming methods and traditional approaches to Q-learning which are based on a Bellman operator.

Filtering is a common theme in the present paper and [19], and appears in the work of Frank Lewis—see for example [23], concerning reinforcement learning techniques for the linear quadratic regulator problem in continuous-time.

Typical test examples in OpenAI gym are challenging in part because the control system is based on a continuous time model with fast sampling. In [14], it is argued that the *temporal difference* sequence is dominated by a cost term since $x_{t_{k+1}} \approx x_{t_k}$, which results in numerical challenges using any RL algorithm. These findings were part of the motivation for the techniques surveyed in this paper. In particular, filtering resolves the numerical challenge described in [14].

The boundedness of the constraint region for convex Q-learning was characterized in [14] for models in discrete time, and again with infinite-horizon objective. The generalization to the continuous-time finite-horizon setting of the present paper is entirely non-trivial–see Prop. 3.1 and Prop. 3.2.

Of course, Q-learning has a much longer history. Watkins' original algorithm [28] was inspired by older temporal difference learning techniques, and versions of the temporal difference are also part of convex Q-learning architectures [24], [25], [20]. Most of these prior works involve a discrete-time setting.

The marriage of MPC and Q-learning considered in Section IV was first investigated in the dissertation [9], [8] for deterministic control systems, and contemporaneously in [29] for MDPs.

Organization Section II sets the stage, with a review of the optimality equations and how these lead to the Q-ODE. New Q-learning algorithms are introduced in Section III, along with new theory regarding exploration to ensure the boundedness of the constraint region. Application to power systems operations is surveyed in Section IV. Conclusions and directions for future research are presented in Section V.

II. HJB Representations

A starting point in the derivation of the HJB equation is Bellman's principle of optimality, which is itself described in terms of the *cost-to-go*: for each $T_0 \in [0, \mathcal{T})$, this is denoted

$$J^*(x, T_0) := \inf \Big\{ \int_{T_0}^{\mathcal{T}} c(x_t, u_t, t)\, dt + J_0(x_{\mathcal{T}}) \Big\} \tag{4}$$

where the infimum is over continuous u on $[T_0, \mathcal{T}]$, subject to dynamics (1), and with $x_{T_0} = x$.

The principle of optimality is expressed as the family of fixed point equations: for $\tau \in [0, \mathcal{T})$ and with $x_0 = x$,

$$J^*(x) = \inf_{u_0^\tau} \Big\{ \int_0^{\tau} c(x_t, u_t, t)\, dt + J^*(x_\tau, \tau) \Big\}$$

Under the assumption that the value function is continuously differentiable, we may divide each side by τ and let $\tau \downarrow 0$ to obtain the HJB equation:

$$\begin{aligned} 0 &= \min_u Q^*(x, u, t) \\ Q^*(x, u, t) &:= c(x, u, t) + J_x^*(x, t) \cdot F(x, u, t) + J_t^*(x, t)\,, \end{aligned} \tag{5}$$

with $J_x^* = \partial_x J^*$, $J_t^* = \partial_t J^*$.

The following is assumed throughout the paper.

(A0) The cost to go is continuously differentiable as a function on $[0, \mathcal{T}] \times \mathbb{R}^n$. Moreover, the minimizer in (5) defines a continuous function $\phi^*(x, t)$, and the optimal input-state pair is obtained via state feedback:

$$u_t^* = \phi^*(x_t^*, t)\,, \qquad 0 \le t \le \mathcal{T}\,.$$

Q-ODE The Q-ODE is a model-free characterization of the HJB equation (5), inspired by the sample path inequality (3).

The function Q^* that is minimized in (5) is often called the Q-function. Subject to (A0), it admits the model free representation

$$Q^*(x_t, u_t, t) = c(x_t, u_t, t) + \tfrac{d}{dt} J_x^*(x_t, t)\,, \quad t \ge 0 \tag{6}$$

To avoid differentiation of measurements we estimate instead the function

$$H^*(x, u, t) := -\sigma J^*(x, t) + Q^*(x, u, t) \tag{7}$$

in which the scalar $\sigma > 0$ is fixed. We have $\phi^*(x, t) = \arg\min_u H^*(x, u, t)$ for all x, t, and

$$\begin{aligned} H^*(x_t, u_t, t) = &-\sigma J^*(x_t, t) \\ &+ \big[c(x_t, u_t, t) + \tfrac{d}{dt} J^*(x_t, t) \big] \end{aligned} \tag{8}$$

Identities (6) and (8) are valid for *any* input-state trajectory.

The Q-ODE is obtained by eliminating J^* from (8), which requires additional notation. For any continuous function $H \colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$, denote $\underline{H}(x, t) = \min_u H(x, u, t)$.

Application of (5) gives $\underline{H}^*(x,t) = -\sigma J^*(x,t)$, which on substituting into (8) and rearranging terms implies the ODE,

$$\begin{aligned}\tfrac{d}{dt}\underline{H}^*(x_t,t) &= \sigma\underline{H}^*(x_t,t)\\ &\quad + \sigma\big[c(x_t,u_t,t) - H^*(x_t,u_t,t)\big]\\ \underline{H}^*(x_\tau,\mathcal{T}) &= -\sigma J^*(x_\tau,\mathcal{T}) = -\sigma J_0(x_\tau)\end{aligned} \tag{9}$$

in which the second equation is treated as a boundary condition for the first. This motivates a time-reversal: For any function $H\colon \mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}\to\mathbb{R}$, its time-reversal along an input-state trajectory is denoted $\overleftarrow{H}_r := H\big(x_{\mathcal{T}-r}, u_{\mathcal{T}-r}, \mathcal{T}-r\big)$. When applied to $\underline{H}^*$, this becomes $\overleftarrow{\underline{H}}{}^*_r = \underline{H}^*(x_{\mathcal{T}-r},\mathcal{T}-r)$.

Equation (9) then justifies

Q-ODE: With boundary condition $\overleftarrow{\underline{H}}{}^*_0 = -\sigma J_0(x_\tau)$,

$$\tfrac{d}{dr}\overleftarrow{\underline{H}}{}^*_r = -\sigma\overleftarrow{\underline{H}}{}^*_r - \sigma[\overleftarrow{c}_r - \overleftarrow{H}{}^*_r]\,,\quad 0\le r\le\mathcal{T}\,. \tag{10}$$

This admits the *algebraic representation*

$$\overleftarrow{\underline{H}}{}^*_r = -\xi_r J_0(x_\tau) + \overleftarrow{\mathcal{H}}{}^*_r - \overleftarrow{\mathcal{C}}_r\,,\quad 0\le r\le\mathcal{T}\,. \tag{11}$$

in which filtering of observations is explicit,

$$\overleftarrow{\mathcal{H}}{}^*_r := \int_0^r \xi_{r-s}\overleftarrow{H}{}^*_s\,ds\,,\qquad \overleftarrow{\mathcal{C}}_r := \int_0^r \xi_{r-s}\overleftarrow{c}_s\,ds \tag{12}$$

with impulse response $\xi_t = \sigma e^{-\sigma t}$.

The proposition that follows inspires the MPC-Q algorithms introduced in the next section.

Proposition 2.1: Suppose that (A0) holds. Consider any continuous function $H\colon \mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}\to\mathbb{R}$ that satisfies the following bound for each input-state trajectory:

$$\begin{aligned}&\overleftarrow{\underline{H}}_r \ge -\xi_r J_0(x_\tau) + \overleftarrow{\mathcal{H}}_r - \overleftarrow{\mathcal{C}}_r\,,\\ &\textit{with}\ \ \overleftarrow{\mathcal{H}}_r = \int_0^r \xi_{r-s}\overleftarrow{H}_s\,ds\,,\qquad 0\le r\le\mathcal{T}\,.\end{aligned} \tag{13}$$

Then $H(x,u,r)\ge H^*(x,u,r)$ for all x,u,r.

The proof of Prop. 2.1 and all other technical results are contained in the Appendix.

III. Q-learning Algorithms

The algorithms introduced here are based on a family of approximations $\{H^\theta : \theta\in\mathbb{R}^d\}$. For each θ, the H^θ-*greedy policy* is defined by

$$\phi^\theta(x,t) = \arg\min_u H^\theta(x,u,t) \tag{14}$$

The ultimate goal of Q-learning is to find the parameter θ^* that leads to the best performance among these policies. An indirect approach is usually applied, such as the *projected Bellman equation* favored in much of academic research. If we are so fortunate that H^{θ^*} approximately solves (11), then inverse dynamic programming arguments yield bounds on the performance of the H^{θ^*}-greedy policy [2], [20].

Prop. 2.1 motivates the definition of the *Bellman error*,

$$\mathcal{B}^\theta_r := -\overleftarrow{\underline{H}}{}^\theta_r - \xi_r J_0(x_\tau) + \overleftarrow{\mathcal{H}}{}^\theta_r - \overleftarrow{\mathcal{C}}_r\,. \tag{15}$$

in which the filtered signal $\{\overleftarrow{\mathcal{H}}{}^\theta_r : 0\le r\le\mathcal{T}\}$ is defined as in (12). The inequality (13) using $H = H^\theta$ is equivalently expressed $\mathcal{B}^\theta_r\le 0$ for each $r\in[0,\mathcal{T}]$.

The following assumptions are imposed in all of the technical results that follow:

Assumption A1: The function class is linear,

$$H^\theta(x,u,r) = \theta^\intercal\psi(x,u,r), \tag{16}$$

The basis $\psi\colon \mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}_+\to\mathbb{R}^d$ and the cost function $c\colon \mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}_+\to\mathbb{R}_+$ are continuously differentiable (C^1).

Moreover, for each $\theta\in\mathbb{R}^d$, the minimum in (14) defines a continuous feedback law $\phi^\theta\colon \mathbb{R}^n\times\mathbb{R}_+\to\mathbb{R}^m$. And, with $u_t = \phi^\theta(x_t,t)$ for $0\le t\le\mathcal{T}$ there is a solution to the state equation (1) on $[0,\mathcal{T}]$ for each initial condition.

Choice of meta-parameters. The d-dimensional basis ψ might be chosen based on known structure of the control problem—see Section IV.

The choice of $\sigma>0$ will depend on signal to noise ratio in any online application. A large value of σ may be justified when data is collected using a simulator.

Another design choice is the input used for training. In the following, it is assumed that the input is a (possibly randomized) stationary policy. See Section IV for an example.

A. Algorithms.

Q-learning algorithms are typically designed to ensure that a *projected Bellman error* is zero under θ^* [24], [20]. We describe here natural analogs for the continuous time model, based on the Q-ODE. The first is a batch algorithm:

For each n, given the current estimate θ_n, the parameter update is obtained as the solution to the nonlinear program,

$$\theta_{n+1} = \arg\min_\theta\big\{\|\mathcal{B}^{\theta|\theta_n}\|^2_{L_2} + \tfrac{1}{\alpha_{n+1}}\|\theta-\theta_n\|^2\big\} \tag{17a}$$

$$\mathcal{B}^{\theta|\theta_n}_r := -\overleftarrow{\underline{H}}{}^{\theta_n}_r - \xi_r J_0(x_\tau) + \overleftarrow{\mathcal{H}}{}^\theta_r - \overleftarrow{\mathcal{C}}_r \tag{17b}$$

in which the non-negative sequence $\{\alpha_n : n\ge 1\}$ is analogous to the usual step-size sequence in RL. The term (17b) is defined as in (15), with the first appearance of θ frozen. The L_2 norm in (17a) is the standard, $\|\mathcal{B}^{\theta|\theta_n}\|^2_{L_2} := \int_0^{\mathcal{T}}[\mathcal{B}^{\theta|\theta_n}_r]^2\,dr$.

For the linear parameterization (16) we write

$$\overleftarrow{\Psi}_r := \int_0^r \xi_{r-s}\overleftarrow{\psi}_s\,ds\,,\quad 0\le r\le\mathcal{T}\,, \tag{18}$$

with $\overleftarrow{\psi}_s := \psi(x_{\mathcal{T}-s}, u_{\mathcal{T}-s}, \mathcal{T}-s)$. This gives $\overleftarrow{\mathcal{H}}{}^\theta_r = \theta^\intercal\overleftarrow{\Psi}_r$, and (17b) becomes

$$\mathcal{B}^{\theta|\theta_n}_r = -\overleftarrow{\underline{H}}{}^{\theta_n}_r - \xi_r J_0(x_\tau) + \theta^\intercal\overleftarrow{\Psi}_r - \overleftarrow{\mathcal{C}}_r \tag{19}$$

Substituting (19) in (17a) and taking the gradient with respect to θ leads to the fixed point equation,

$$0 = \langle\mathcal{B}^{\theta_{n+1}|\theta_n}, \overleftarrow{\Psi}\rangle + \tfrac{1}{\alpha_{n+1}}[\theta_{n+1}-\theta_n]$$

in which the first term depends linearly on θ_{n+1}:

$$\langle\mathcal{B}^{\theta_{n+1}|\theta_n}, \overleftarrow{\Psi}\rangle := \int_0^{\mathcal{T}} \overleftarrow{\Psi}_r\mathcal{B}^{\theta_{n+1}|\theta_n}_r\,dr\,.$$

If the resulting sequence of estimates $\{\theta_n\}$ is bounded, it follows that $\|\theta_{n+1}-\theta_n\| = O(\alpha_{n+1})$, which justifies the following *approximation*:

$$\theta_{n+1} = \theta_n - \alpha_{n+1}\langle\mathcal{B}^{\theta_n|\theta_n}, \overleftarrow{\Psi}\rangle \tag{20}$$

This recursion is similar to Watkins' algorithm.

There is little theory to predict the success of (17) or the recursion (20). In particular, outside of very special cases, the stability of Q-learning has been open topic of research for more than three decades. For history and recent criteria for stability based on sufficient exploration see [21].

Convex Q-Learning Prop. 2.1 is motivation for the following "ideal" algorithm: Choose a probability measure μ on $\mathbb{R}^n \times \mathbb{R}^m \times [0, \mathcal{T}]$, and solve the nonlinear program,

$$\theta^* = \arg\min_{\theta} \ \langle \mu, H^\theta \rangle \qquad \text{s.t.} \quad \mathcal{B}_r^\theta \le 0, \quad r \in [0, \mathcal{T}] \tag{21}$$

In this paper, the infinite number of constraints (21) are relaxed by the single constraint,

$$\frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \big[\mathcal{B}_r^\theta\big]_+ \, dr \le \text{Tol} \tag{22}$$

where Tol > 0 is a small constant, and $[s]_+ = \max(0, s)$. This is a convex constraint, subject to (16).

B. Exploration and Constraint Geometry

The constraint set associated with (22) is denoted

$$\Theta = \Big\{ \theta \in \mathbb{R}^d : \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \big[\mathcal{B}_r^\theta\big]_+ \, dr \le \text{Tol} \Big\} \tag{23}$$

Necessary and sufficient conditions for boundedness will be obtained based on algebraic conditions on the basis along input-output sample paths obtained for training. To ease analysis and save space, we adopt the notation,

$$\psi_t := \psi(x_t, u_t, t)\,, \qquad \overleftarrow{\psi}_r := \psi_{\mathcal{T}-r}\,, \quad 0 \le t, r \le \mathcal{T}\,.$$

The covariance matrix is denoted

$$\Sigma := \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \widetilde{\psi}_s \widetilde{\psi}_s^{\intercal} \, ds\,, \quad \textit{with} \ \ \widetilde{\psi}_s := \psi_s - \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \psi_t \, dt \tag{24}$$

The conditions that follow are the focus of analysis in the remainder of this section. The third is a standard assumption intended to capture "sufficient exploration" in temporal difference learning [27], [20]. In the context of this paper, it is Condition E1 that is most valuable: Prop. 3.2 tells us that Θ is bounded under this condition, and hence what should be considered "good exploration".

Condition E1: The set $\{\psi_t : 0 \le t \le \mathcal{T}\}$ is not restricted to any half space in $\mathbb{R}^d$.

Condition E2: The only vector $v \in \mathbb{R}^d$ satisfying $\underline{\overleftarrow{H}}_r^v \ge \overleftarrow{\mathcal{H}}_r^v$ for all $0 \le r \le \mathcal{T}$ is $v = 0$.

Condition E3: $\Sigma > 0$, with Σ defined in (24).

Proposition 3.1: If Condition E1 holds then Conditions E2 and E3 follow.

The relationship between E1 and E3 is straightforward, since the latter is equivalent to the statement that $\{\psi_t : 0 \le t \le \mathcal{T}\}$ is not restricted to any *hyperplane* in $\mathbb{R}^d$.

Prop. 3.1 combined with the following establishes that boundedness of Θ is equivalent to Condition E2.

Proposition 3.2: If Condition E1 holds then Θ is bounded. Conversely, if Θ is bounded then Condition E2 holds.

IV. Optimal Dispatch of Energy Resources

We survey here results from the application of convex Q-learning to the optimal allocation of distributed energy resources. The goal is to schedule generation and other "balancing assets" to meet supply-demand constraints while minimizing cost. It is assumed that the balancing assets are derived from flexible loads (such as water heaters or water pumping) alongside batteries. We use the term *virtual energy storage* (VES) for both real and virtual batteries.

A. Dispatch model

There are $M \ge 2$ classes of VES along with generation $\{g_t\}$ (the aggregation of all generators in the balancing area). The goal is to optimally allocate these resources to balance the net load $\{\ell_t\}$ over the time horizon $[0, \mathcal{T}]$.

Following [6], [3], [18], the *state of charge* (SoC) for the ith VES class evolves according to the linear dynamics

$$\tfrac{d}{dt} x_t^i = -\alpha_i x_t^i - z_t^i \qquad 1 \le i \le M\,, \tag{25}$$

in which $-z_t^i$ is power deviation at time t, and α_i is a non-negative leakage parameter. For a a refrigerator or water heater, the SoC x_t^i is an affine function of internal temperature, and α_i corresponds to the thermal time constant.

Optimal control formulation. The cost function is designed based on three goals: maintain the SoC within bounds, and penalize peaks and ramps in generation. To model cost on ramping, we introduce

$$u_t^i := \tfrac{d}{dt} z_t^i \tag{26}$$

The generation variable g can be eliminated in the optimization problem by imposing the supply-demand constraint, $g_t + z_t^\sigma = \ell_t$ with $z_t^\sigma = \sum z_t^i$.

The terminal cost J_0 in (2) is chosen to be quadratic, of the form $J_0(x, z) = x^{\intercal} D x + k_\ell (z^\sigma - \ell_\tau)^2$ with $k_\ell > 0$ and $D > 0$ diagonal ($M \times M$). The cost function c is the sum of three components, reflecting the three goals:

$$c(x_t, z_t, u_t, t) = c_{\mathsf{x}}(x_t) + \kappa \big[u_t^\sigma - \tfrac{d}{dt}\ell_t\big]^2 + \kappa_\ell [z_t^\sigma - \ell_t]^2$$

with $u_t^\sigma = \sum u_t^i$, and κ, κ_ℓ positive constants. A soft constraint on capacity is imposed via

$$c_{\mathsf{x}}(x) = \sum_{i=1}^M c^i(x^i)\,, \quad x \in \mathbb{R}^M\,, \tag{27}$$

where each $c^i \colon \mathbb{R} \to \mathbb{R}_+$ is smooth and strongly convex.

The goal is to solve the optimal control problem,

$$\min_u \quad \int_0^{\mathcal{T}} c(x_t, z_t, u_t, t)\, dt + J_0(x_\tau^a) \tag{28a}$$

$$\text{s.t.} \quad \tfrac{d}{dt} x_t^i = -\alpha_i x_t^i - z_t^i \tag{28b}$$

$$\tfrac{d}{dt} z_t^i = u_t^i, \quad 1 \le i \le M\,,\ 0 \le t \le \mathcal{T} \tag{28c}$$

This is of the form (2) with augmented state $x^a := (x, z)$, and M-dimensional input u. It falls in the category of singular optimal control because the cost is not coercive in u (there is a cost on the sum u_t^σ, and not on the individual terms u_t^i) [5], [7].

Function approximation architecture A form of *state space collapse* is established in [18]: under mild conditions on J_0, the cost to go for any time T_0 can be expressed as a convex function of $x^{\sigma,a} := (x^\sigma, z^\sigma)$ with $x^\sigma = \sum x^i$. State space collapse provides motivation for the choice of function class in Q-learning.

The construction of function class is model based, in which we begin with an affine function class for the value function:

$$J^\theta(x^a, t) = J_0(x^a) + \theta^\intercal \psi(x^{\sigma,a}, t)\,, \quad \theta \in \mathbb{R}^d\,, \tag{29}$$

with $\psi\colon \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}^d$. The representation (7) then motivates the linear function class,

$$\begin{aligned} H^\theta(x^a, u, t) &:= -\sigma J^\theta(x^a, t) + Q^\theta(x^a, u, t) \\ Q^\theta(x^a, u, t) &:= c(x^a, u, t) \\ &\quad + J^\theta_x(x^a, t)\cdot F(x^a, u, t) + J^\theta_t(x^a, t) \end{aligned} \tag{30}$$

To match the ideal $J^\theta(x^a, \mathcal{T}) = J_0(x^a)$, the basis was designed to ensure $\psi(x^{\sigma,a}, \mathcal{T}) = 0$ for each $x^{\sigma,a} \in \mathbb{R}^2$. It is convenient to take a typical basis function of the form

$$\psi_{i,j}(x^{\sigma,a}, t) = q_i(x^{\sigma,a}) p_j(t) \tag{31}$$

in which $q_i \in \{(x^\sigma)^2, x^\sigma, (z^\sigma)^2, z^\sigma, 2x^\sigma z^\sigma, 1\}$ for $1 \le i \le 6$. The functions $\{p_j\}$ were taken to be a mixture of Fourier basis elements and polynomials. Through trial and error we arrived at three possibilities: we took $p_1(t) = t^2$, and for $j \ge 2$ the function p_j was an element of the set

$$\{1 - \cos(\omega_i t) : 1 \le i \le n_f\}$$

with $n_f = 30$ in all experiments. Thus, $d = 5 \times 31 = 155$.

The basis was chosen so that the functions of time are non-negative. Writing $\theta \in \mathbb{R}^d$ in compatible form so that $\theta^\intercal \psi = \sum_{i,j} \theta_{i,j}\psi_{i,j}$, the constraint $\theta_{i,j} \ge 0$ was imposed in implementations of convex Q-learning for any i, j for which $\psi_{i,j}(x^{\sigma,a}, t) = (x^\sigma)^2 p_j(t)$ or $(z^\sigma)^2 p_j(t)$. It was found that this helped to ensure that the solution θ^* would result in a cost to go approximation $J^{\theta^*}(x^a, t)$ that is convex in its first variable for each t.

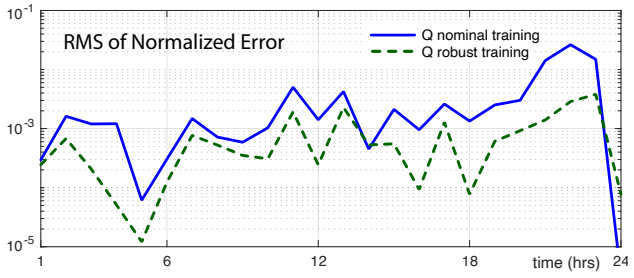

Fig. 1. Root mean square error for the normalized error for the cost to go, evaluated along an optimal trajectory, with value function obtained using convex Q-learning.

B. Three policies

Three policies were compared, all based on MPC. The policies are described in continuous time only to avoid changes in notation. In practice and in our simulations we employed sampling and used an Euler approximation for integrals.

Each policy is model based, and requires a "look-ahead" time horizon τ and "τ-terminal cost" $c^\bullet$. For any time $t_0 \ge 0$, the input u_{t_0} is obtained through the following steps. First, the optimization problem is solved:

$$\min\Big\{\int_{t_0}^{t_0+\tau} c(x_{t_0+t}, u_{t_0+t}, t_0 + t)\, dt + c^\bullet(x_{t_0+\tau})\Big\} \tag{32}$$

with x_{t_0} given. The optimizer is a function of time $\{u^0_t : t_0 \le t \le t_0 + \tau\}$. The MPC input at time t_0 is $u_{t_0} = u^0_{t_0}$. This may be expressed as time varying state-feedback, though this is not how MPC is implemented in practice.

The three policies are distinguished by the choice of $c^\bullet$.

1. MPC-0. This uses $c^\bullet := 0$, and the resulting policy is denoted $u_t = \phi^\circ(x^a_t, t)$ for $0 \le t \le \mathcal{T}$.

Typically τ will be much smaller than $\mathcal{T}$, in which case MPC-0 is unlikely to be close to optimal.

There is of course an optimal choice for $c_\bullet$, provided we allow dependency on time:

$$c^\bullet(x, t_0{+}\tau) = J^*(x, \min(\mathcal{T}, t_0{+}\tau)), \quad \textit{for each } x,\ \tau \textit{ and } t_0.$$

The principle of optimality tells us that the MPC solution is then optimal, $u_t = \phi^*(x^a_t, t)$ for $0 \le t \le \mathcal{T}$.

This motivates the terminal cost $c^\bullet(x, t_0{+}\tau) = J^{\theta^*}(x, t_0{+}\tau)$, with J^θ defined in (29). The next two policies make use of this time-dependent τ-terminal cost in MPC, with θ^* obtained using the convex program (21). The policies are differentiated by the data used for training.

The following steps are common to each of the two approaches: data is collected from N_r independent runs; with $N_r = 44$ used in the simulation experiments. The entries of the initial condition $\{x^a_i : 1 \le i \le 2M\}$ were sampled uniformly and independently on the interval $[-5, 5]$.

The input $u_{t_n} = \phi^\circ(x_{t_n}, t_n) + W_{t_n}$ for adopted for training, with $\{W_{t_n}\}$ i.i.d., sampled uniformly from $[-1, 1]^M$.

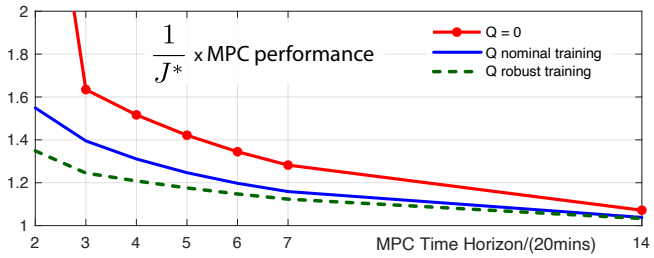

Fig. 2. Performance of MPC-Q on the nominal model: with $Q = 0$ the performance is poor unless the time horizon is long.

2. MPC-Q with nominal training For each of N_r runs, the solution to (25) is obtained to generate $\{x^a_t : 0 \le t \le \mathcal{T}\}$. This data is used in the convex program (21) to obtain

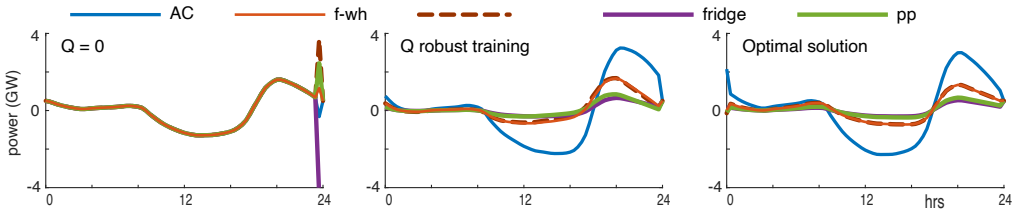

Fig. 3. Trajectories of power deviation from each load class: comparison of three policies.

θ^*; we denote $J^{\text{Nom}} = J^{\theta^*}$. The resulting policy is denoted $u_t = \phi^{\text{Nom}}(x_t^a, t)$ for $0 \le t \le \tau$.

3. MPC-Q with robust training In addition to sampling initial conditions x^a, for each of N_r runs, the model parameter is perturbed via

$$\tilde{\alpha}^{n,i} = \alpha^i \times V^{n,i}\,, \qquad 1 \le i \le M\,, 1 \le n \le N_r \tag{33}$$

where $\{V^{n,i}\}$ i.i.d. and sampled independently of x^a from $[1-\varepsilon, 1+\varepsilon]$, with ε ranging from 0 to 1. The solution to (25) with the parameter $\tilde{\alpha}^n$ is used to generate the state trajectory in the nth run. This data is used in the convex program (21) to obtain θ^*, and we denote $J^{\text{Rob}} = J^{\theta^*}$.

The resulting policy is denoted $u_t = \phi^{\text{Rob}}(x_t^a, t)$ for $0 \le t \le \tau$. We do not yet have theory to support the use of ϕ^{Rob}, but we believe this type of training is valuable. Lacking theory, we have resorted to an empirical study.

C. Simulations

The system parameters for VES, cost function c_{x} appearing in (27), and net load ℓ were taken from [18]. The optimal dispatch problem (28a) was considered with $M = 5$ VES classes: ACs, residential WHs (fwh), commercial WHs (swh), refrigerators (rfg), and pool pumps (pp). The time horizon τ was set to 24 hours, and $\sigma = 5 \times 10^{-4}$ in the convex Q learning algorithm and the definition of $\mathcal{H}^\theta$.

Investigation of policy performance requires additional experiments.

For any feedback policy $u_t = \phi(x_t^a, t)$, the associated total cost from initial condition $(x_0, z_0) = x^a$ is denoted

$$J^\Phi(x^a) = \int_0^\tau c(x_t^a, u_t, t)\, dt \tag{34}$$

In the numerical results summarized below we compared this with the optimal $J^*(x^a)$ from specific initial conditions, as well as the cost to go.

Experimental results 1: testing on nominal model. The first experiment was designed to investigate the loss in performance introduced from perturbations of the model during training.

The normalized error between the approximation J^{θ^*} and the optimal cost to go J^* was obtained for the two training approaches with J^{θ^*} indicating either J^{Nom} (nominal training) or J^{Rob} (robust training).

In these experiments the initial condition was fixed at a typical value, and the true optimal solution $\{x_t^*, z_t^*, u_t^* : 0 \le t \le \tau\}$ was obtained numerically. For each t, the cost-to-go $J^*(x_t^*, z_t^*, t)$ was compared with $J^{\text{Nom}}(x_t^*, z_t^*, t)$ and $J^{\text{Rob}}(x_t^*, z_t^*, t)$, and in each case the normalized errors were obtained:

$$\mathcal{E}_t^{\text{Nom}} = \tfrac{1}{J^*(x_t^*, z_t^*, t)}[J^{\text{Nom}}(x_t^*, z_t^*, t) - J^*(x_t^*, z_t^*, t)]$$

with $\mathcal{E}_t^{\text{Rob}}$ defined analogously. On hundred independent runs were obtained, and the RMSE was obtained in each case. Fig. 1 shows that the errors are very small in each case.

The next experiments compare policy performance using MPC-Q. The outcomes are surprising.

Fig. 2 shows data from one typical experiment, performed on the nominal model. It is surprising to see that the policy ϕ^{Rob} gave the smallest error (compared to optimal) for each look-ahead horizon considered (as small as 40 minutes). Performance for ϕ^{MPC} (with $c^\bullet \equiv 0$) was far worse.

Fig. 3 shows the power trajectories obtained using MPC-Q with look-ahead horizon $\tau = 40$mins, with Q robust training, mirrors the optimal solution. We omit plots for MPC-Q with Q nominal training since the performance is similar. The performance of MPC fails dramatically with zero terminal cost.

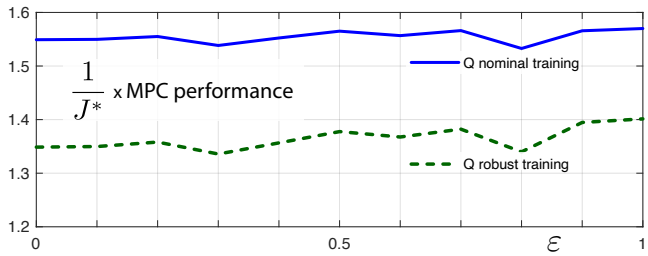

Fig. 4. Robustness of MPC-Q: normalized averaged total cost in MPC-Q as a function of ε with look-ahead horizon 40mins.

Experimental results 2: testing on perturbed models The impact of model uncertainty is investigated next.

To test a given policy ϕ we conducted N_p independent trials for a range of $\varepsilon \ge 0$, and averaged the resulting total cost obtained in each trial to obtain

$$\widehat{J}_\varepsilon^\Phi = \frac{1}{N_p}\sum_{k=1}^{N_p} J^\Phi(x_k^a) \tag{35}$$

For each k the initial condition was chosen randomly, as well as the perturbation of the model defined by $\tilde{\alpha}_k$ via (33) for $1 \le k \le N_p$, with $N_p = 50$.

Fig. 4 shows that $\widehat{J}_\varepsilon^\Phi$ is nearly independent of ε for either policy $\phi = \phi^{\text{Nom}}$ or $\phi = \phi^{\text{Rob}}$, with the latter giving better performance for each value of ε tested. The results for MPC-0 are not shown since the ratio was always greater than 3.

V. Conclusions

The Q-ODE for continuous-time finite-horizon optimal control is a new model-free characterization of the HJB equation that lends itself to the formulation of reinforcement learning algorithms.

Theory concerning the impact of disturbances and measurement noise is an important area for future research. New theory for average cost convex Q-learning in a stochastic environment is contained in [12]. We believe the value of filtering in convex Q-learning will be apparent when we include measurement noise in simulation experiments, but currently have no guidelines to choose σ.

The use of state space collapse to design a function approximation architecture was successful in the example considered. This will likely prove valuable in other applications. Such extensions may require techniques to characterize or approximate the manifold on which an optimal solution evolves, or perhaps we can create algorithms that will "learn" this structure.

References

[1] 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*, 130:3610–3618, 13–15 Apr 2021.

[2] D. P. Bertsekas. Dynamic programming and suboptimal control: A survey from ADP to MPC. *European Journal of Control*, 11(4-5):310–334, 2005.

[3] N. Cammardella, J. Mathias, M. Kiener, A. Bušić, and S. Meyn. Balancing California's grid without batteries. In *Proc. of the IEEE Conf. on Dec. and Control*, pages 7314–7321, Dec 2018.

[4] W. E. Dixon, A. Behal, D. M. Dawson, and S. P. Nagarkatti. *Nonlinear control of engineering systems: a Lyapunov-based approach*. Springer Science & Business Media, 2003.

[5] B. Francis. The optimal linear-quadratic time-invariant regulator with cheap control. *IEEE Trans. Automat. Control*, 24(4):616–621, 1979.

[6] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent. Aggregate flexibility of thermostatically controlled loads. *IEEE Trans. on Power Systems*, 30(1):189–198, Jan 2015.

[7] M. L. Hautus and L. M. Silverman. System structure and singular control. *Linear algebra and its applications*, 50:369–402, 1983.

[8] A. Kowli, E. Mayhorn, K. Kalsi, and S. P. Meyn. Coordinating dispatch of distributed energy resources with model predictive control and Q-learning. Technical report, Coordinated Science Laboratory technical report UILU-ENG-12-2204, May 2012.

[9] A. S. Kowli. *Reinforcement Learning Techniques for Controlling Resources in Power Networks*. PhD thesis, University of Illinois at Urbana Champaign, University of Illinois, Urbana, IL, USA, May 2013.

[10] F. L. Lewis and D. Liu. *Reinforcement learning and approximate dynamic programming for feedback control*, volume 17. John Wiley & Sons, 2013.

[11] F. Lu. *Convex Q-learning: theory and applications*. PhD thesis, University of Florida, 2023.

[12] F. Lu and S. Meyn. Convex Q-learning in a stochastic environment. In *Proc. of the IEEE Conf. on Dec. and Control*, 2023 (extended version, arXiv:2309.05105).

[13] F. Lu, P. G. Mehta, S. P. Meyn, and G. Neu. Convex Q-learning. In Proc. of the *American Control Conf.*, pages 4749–4756, 2021.

[14] F. Lu, P. G. Mehta, S. P. Meyn, and G. Neu. Convex analytic theory for convex Q-learning. In *Proc. of the IEEE Conf. on Dec. and Control*, pages 4065–4071, Dec 2022.

[15] A. Martinelli, M. Gargiani, M. Draskovic, and J. Lygeros. Data-driven optimal control of affine systems: A linear programming perspective. *IEEE Control Systems Letters*, 6:3092–3097, 2022.

[16] A. Martinelli, M. Gargiani, and J. Lygeros. Data-driven optimal control with a relaxed linear program. *Automatica*, 136:110052, 2022.

[17] J. Mathias. *Balancing the power grid with distributed control of flexible loads*. PhD thesis, University of Florida, Gainesville, FL, USA, 2022.

[18] J. Mathias, R. Moye, S. Meyn, and J. Warrington. State space collapse in resource allocation for demand dispatch. In *Proc. of the IEEE Conf. on Dec. and Control*, pages 6181–6188 (and arXiv:1909.06869), Dec 2019.

[19] P. G. Mehta and S. P. Meyn. Q-learning and Pontryagin's minimum principle. In *Proc. of the IEEE Conf. on Dec. and Control*, pages 3598–3605, Dec. 2009.

[20] S. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, Cambridge, 2022.

[21] S. Meyn. Stability of Q-learning through design and optimism. *arXiv 2307.02632*, 2023.

[22] G. Neu and N. Okolo. Efficient global planning in large MDPs via stochastic primal-dual optimization. In *International Conference on Algorithmic Learning Theory*, pages 1101–1123, 2023.

[23] M. Palanisamy, H. Modares, F. L. Lewis, and M. Aurangzeb. Continuous-time Q-learning for infinite-horizon discounted cost linear quadratic regulator problems. *IEEE Trans. on Cybernetics*, 45(2):165–176, 2014.

[24] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

[25] C. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

[26] A. Tanzanakis and J. Lygeros. Data-driven control of unknown systems: A linear programming approach. *ArXiv*, abs/2003.00779, 2020.

[27] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control*, 42(5):674–690, 1997.

[28] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

[29] M. Zhong, M. Johnson, Y. Tassa, T. Erez, and E. Todorov. Value function approximation and model predictive control. In *IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*, pages 100–107, 2013.

Appendix

The proof of Prop. 2.1 requires Grönwall's inequality in this simplified form:

Lemma 1.1 (Bellman-Grönwall): Let w be a continuous real-valued function on the interval $[0, \mathcal{T}]$. Suppose that the following integral bound holds with the constants $\alpha, \beta \geq 0$:

$$w_r \leq \alpha + \beta \int_0^r w_s \, ds \,, \qquad 0 \leq r \leq T$$

Then, $w_r \leq \alpha e^{\beta r}$ for $0 \leq r \leq T$.

Proof of Prop. 2.1: Since $H \geq \underline{H}$, it follows from (13) that for any trajectory,

$$\overleftarrow{H}_r \geq -\xi_r J_0(x_\tau) + \overleftarrow{\mathcal{H}}_r - \overleftarrow{\mathcal{C}}_r. \tag{36}$$

For any optimal trajectory $\{x_r^*, u_r^*\}$ we have $\overleftarrow{H}_r^* = \underline{\overleftarrow{H}}_r^*$, so from (11),

$$\overleftarrow{H}^* = -\xi_r J_0(x_\tau) + \overleftarrow{\mathcal{H}}_r^* - \overleftarrow{\mathcal{C}}_r \tag{37}$$

Denote $\Delta_r := \overleftarrow{H}_r^* - \overleftarrow{H}_r$. Subtracting (36) from (37) then yields along the optimal input-state trajectory,

$$\Delta_r \leq \overleftarrow{\mathcal{H}}_r^* - \overleftarrow{\mathcal{H}}_r = \int_0^r \xi_{r-s} \Delta_s \, ds,$$

where the equality on the right follows from the definitions of $\mathcal{H}$ and $\mathcal{H}^*$. Setting $w_r = e^{\sigma r} \Delta_r$ and applying Lemma 1.1,

$$w_r = e^{\sigma r} \Delta_r \leq 0, \qquad 0 \leq r \leq \mathcal{T},$$

which in turn implies $\Delta_r \le 0$, thereby yielding $H \ge H^*$ along this optimal trajectory.

It follows that $H(x,u,r) \ge H^*(x,u,r)$ for any (x,u,r), since there is an optimizing trajectory that passes through any such triple. ∎

Proof of Prop. 3.1: To establish that Condition E1 implies E2, we establish the contrapositive: if there is a non-zero vector v satisfying $\underline{\overleftarrow{H}}{}^v_r \ge \overleftarrow{\mathcal{H}}{}^v_r$ for each r, then the set $\{\overleftarrow{\psi}_r : 0 \le r \le \mathcal{T}\}$ is restricted to a half space in $\mathbb{R}^d$.

If such v exists, then by definition of $\underline{H}$,

$$H^v(x_{\mathcal{T}-r},u,r) \ge \underline{\overleftarrow{H}}{}^v_r \ge \overleftarrow{\mathcal{H}}{}^v_r\,, \qquad u \in \mathbb{R}^m.$$

Letting $p_r = v^{\intercal}\overleftarrow{\psi}_r$, and $y_r = v^{\intercal}\overleftarrow{\Psi}_r$, this inequality implies that $p_r \ge y_r$ and by definition,

$$\begin{aligned} y_r &= \int_0^r \xi_{t-r} p_\tau d\tau \\ \frac{d}{dr} y_r &= -\sigma(y_r - p_r), \end{aligned} \tag{38}$$

On applying the boundary condition $y_0 = 0$,

$$y_r = -\sigma \int_0^r (y_\tau - p_\tau) d\tau \ge 0.$$

Letting $\delta_r = p_r - y_r$, which is non-negative, gives $p_r = y_r + \delta_r$, and for each $0 \le r \le \mathcal{T}$,

$$v^{\intercal}\overleftarrow{\psi}_r = -\sigma \int_0^r (y_\tau - p_\tau) d\tau + \delta_r \ge 0\,,.$$

Hence Condition E1 fails when E2 fails, as claimed.

To show that Condition E1 implies E3, we again establish the contrapositive: if $\det(\Sigma) = 0$, then the set $\{\psi_t : 0 \le t \le \mathcal{T}\}$ is restricted to a half space in $\mathbb{R}^d$.

If $v \in \text{Null}(\Sigma)$ with $v \ne 0$, then

$$0 = v^{\intercal}\Sigma v = \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} (v^{\intercal}\widetilde{\psi}_t)^2 \, dt$$

Since $\{\widetilde{\psi}_t\}$ is continuous in t, it follows that

$$v^{\intercal}\widetilde{\psi}_t = 0\,, \qquad \text{for } 0 \le t \le \mathcal{T}.$$

This implies that $\{\psi_t\}$ is restricted to a half space, so that Condition E1 fails. ∎

Proof of Prop. 3.2: **Step 1: E1 implies boundedness of Θ.** Prop. 3.1 tells us that E2 follows from E1, so it suffices to establish boundedness of Θ subject to E2. We establish its contrapositive: if Θ is unbounded, then there is a non-zero vector v satisfying $\underline{\overleftarrow{H}}{}^v_r \ge \overleftarrow{\mathcal{H}}{}^v_r$ for $0 \le r \le \mathcal{T}$.

Unboundedness of Θ means that for each $m \ge 0$, there exists θ^m such that $\|\theta^m\| \ge m$ and

$$\frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \max\left\{0, \mathcal{J}_r - \underline{\overleftarrow{H}}{}^{\theta^m}_r + \overleftarrow{\mathcal{H}}{}^{\theta^m}_r\right\} dr \le \text{Tol} \tag{39}$$

with $\mathcal{J}_r := -\xi_r J_0(x_\tau) - \overleftarrow{\mathcal{C}}_r$.

Dividing (39) by $\|\theta^m\|$ gives:

$$\frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \max\left\{0, \frac{\mathcal{J}_r}{\|\theta^m\|} - \frac{\underline{\overleftarrow{H}}{}^{\theta^m}_r}{\|\theta^m\|} + \frac{\overleftarrow{\mathcal{H}}{}^{\theta^m}_r}{\|\theta^m\|}\right\} dr \le \frac{\text{Tol}}{\|\theta^m\|} \tag{40}$$

Denote $\check{\theta}^m = \theta^m / \|\theta^m\|$. By the definition of $\underline{\overleftarrow{H}}{}^{\theta^m}$,

$$\frac{1}{\|\theta^m\|} \underline{\overleftarrow{H}}{}^{\theta^m}_r = \min_u \Big\{ \frac{1}{\|\theta^m\|} H^{\theta^m}(\overleftarrow{x}_r, u, r) \Big\} = \underline{\overleftarrow{H}}{}^{\check{\theta}^m}_r$$

Thus, we can write (40) as

$$\frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \max\left\{0, \frac{\mathcal{J}_r}{\|\theta^m\|} - \underline{\overleftarrow{H}}{}^{\check{\theta}^m}_r + \overleftarrow{\mathcal{H}}{}^{\check{\theta}^m}_r\right\} dr \le \frac{\text{Tol}}{\|\theta^m\|} \tag{41}$$

Since $\|\check{\theta}^m\| = 1$ for each m, there exists a convergent subsequence $\{\theta^{m_i}\}$ with limit satisfying $\|\check{\theta}\| = 1$:

$$\lim_{i\to\infty} \frac{\theta^{m_i}}{\|\theta^{m_i}\|} = \lim_{i\to\infty} \check{\theta}^{m_i} = \check{\theta}$$

The inequality (41) then gives

$$\begin{aligned} &\frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \max\left\{0, -\underline{\overleftarrow{H}}{}^{\check{\theta}}_r + \overleftarrow{\mathcal{H}}{}^{\check{\theta}}_r\right\} dr \\ &\quad = \lim_{i\to\infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \max\left\{0, \tfrac{1}{\|\theta^{m_i}\|}\mathcal{J}_r - \underline{\overleftarrow{H}}{}^{\check{\theta}^{m_i}}_r + \overleftarrow{\mathcal{H}}{}^{\check{\theta}^{m_i}}_r\right\} dr \\ &\quad \le 0 \end{aligned}$$

Continuity of $\{\underline{\overleftarrow{H}}{}^{\check{\theta}}_r, \overleftarrow{\mathcal{H}}{}^{\check{\theta}}_r : 0 \le r \le \mathcal{T}\}$ implies the desired conclusion: E2 fails, with $v = \check{\theta}$,

$$\underline{\overleftarrow{H}}{}^{\check{\theta}}_r \ge \overleftarrow{\mathcal{H}}{}^{\check{\theta}}_r\,, \qquad 0 \le r \le \mathcal{T}\,.$$

Step 2: Boundedness of Θ implies E2. We once again establish the contrapositive: if Condition E2 fails, we show that Θ is unbounded.

Failure of E2 implies that there is $v \ne 0$ satisfying $\underline{\overleftarrow{H}}{}^v_r \ge \overleftarrow{\mathcal{H}}{}^v_r$ for $0 \le r \le \mathcal{T}$. To show that Θ is unbounded we fix $\theta^0 \in \Theta$, and show that $\theta^\omega := \theta^0 + \omega v \in \Theta$ for each $\omega \ge 0$. Because the function class is linear, we have

$$\begin{aligned} \underline{\overleftarrow{H}}{}^{\theta^\omega}_r &:= \min_u \{H^{\theta^\omega}(x_{\mathcal{T}-r}, u, \mathcal{T}-r)\} \\ &= \min_u \{H^{\theta^0}(x_{\mathcal{T}-r}, u, \mathcal{T}-r) + \omega H^v(x_{\mathcal{T}-r}, u, \mathcal{T}-r)\} \end{aligned}$$

This and sub-linearity of the minimum gives for each r,

$$\underline{\overleftarrow{H}}{}^{\theta^\omega}_r \ge \underline{\overleftarrow{H}}{}^{\theta^0}_r + \omega \underline{\overleftarrow{H}}{}^v_r.$$

It follows that the Bellman error for θ^ω admits the bound,

$$\begin{aligned} \mathcal{B}^{\theta^\omega}_r &:= \mathcal{J}_r - \underline{\overleftarrow{H}}{}^{\theta^\omega}_r + \overleftarrow{\mathcal{H}}{}^{\theta^\omega}_r \\ &\le \mathcal{J}_r - [\underline{\overleftarrow{H}}{}^{\theta^0}_r + \omega \underline{\overleftarrow{H}}{}^v_r] + [\overleftarrow{\mathcal{H}}{}^{\theta^0}_r + \omega \overleftarrow{\mathcal{H}}{}^v_r] \end{aligned}$$

and on rearranging terms, $\mathcal{B}^{\theta^\omega}_r \le \mathcal{B}^{\theta^0}_r + \omega[-\underline{\overleftarrow{H}}{}^v_r + \overleftarrow{\mathcal{H}}{}^v_r]$.

By assumption, we have $-\underline{\overleftarrow{H}}{}^v_r + \overleftarrow{\mathcal{H}}{}^v_r \le 0$ and thus $\mathcal{B}^{\theta^\omega}_r \le \mathcal{B}^{\theta^0}_r$. Consequently, $\theta^\omega \in \Theta$ for every ω, as claimed:

$$\frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \max\{0, \mathcal{B}^{\theta^\omega}_r\} \, dr \le \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \max\{0, \mathcal{B}^{\theta^0}_r\} \, dr \le \text{Tol}$$

∎