Model-Based Policy Iterations for Nonlinear Systems via Controlled Hamiltonian Dynamics

Mario Sassano, *Senior Member, IEEE*, Thulasi Mylvaganam, *Senior Member, IEEE*, and Alessandro Astolfi, *Fellow, IEEE*

***Abstract*—The infinite-horizon optimal control problem for nonlinear systems is studied. In the context of model-based, iterative learning strategies we propose an alternative definition and construction of the *temporal difference error* arising in policy iteration strategies. In such architectures, the error is computed via the evolution of the Hamiltonian function (or, possibly, of its integral) along the trajectories of the closed-loop system. Herein the temporal difference error is instead obtained via two subsequent steps: first the dynamics of the underlying costate variable in the Hamiltonian system is steered by means of a (virtual) control input in such a way that the stable invariant manifold becomes externally attractive. Then, the *distance-from-invariance* of the manifold, induced by approximate solutions, yields a natural candidate measure for the *policy evaluation* step. The *policy improvement* phase is then performed by means of standard gradient descent methods that allows us to correctly update the weights of the underlying functional approximator. The above-mentioned architecture then yields an *iterative (episodic) learning* scheme based on a scalar, constant *reward* at each iteration, the value of which is insensitive to the length of the episode, as in the original spirit of reinforcement learning strategies for discrete-time systems. Finally, the theory is validated by means of a numerical simulation involving an automatic flight control problem.**

Manuscript received 7 May 2021; revised 31 December 2021; accepted 13 May 2022. Date of publication 17 August 2022; date of current version 26 April 2023. The work of Mario Sassano was supported by the University “Tor Vergata,” Rome, 2019 “Beyond the Borders” Program under Grant E89C20000600005. The work of Alessandro Astolfi was supported in part by the European Union’s Horizon 2020 Research and Innovation Programme under Grant 739551 (KIOS CoE), in part by the Italian Ministry for Research in the framework of the 2017 Program for Research Projects of National Interest (PRIN) under Grant 2017YKXYXJ, and in part by the framework of the 2020 Program for Research Projects of National Interest (PRIN) under Grant 2020RTWES4. Recommended by Senior Editor Tetsuya Iwasaki and Guest Editors George J. Pappas, Anuradha M. Annaswamy, Manfred Morari, Claire J. Tomlin, Rene Vidal, and Melanie N. Zeilinger. *(Corresponding author: Mario Sassano.)*

Mario Sassano is with the Dipartimento di Ingegneria Civile ed Ingegneria Informatica, Università di Roma Tor Vergata, 00133 Roma, Italy (e-mail: mario.sassano@uniroma2.it).

Thulasi Mylvaganam is with the Department of Aeronautics, Imperial College London, SW7 2AZ London, U.K. (e-mail: t.mylvaganam@imperial.ac.uk).

Alessandro Astolfi is with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K., and also with the Dipartimento di Ingegneria Civile ed Ingegneria Informatica, Università di Roma Tor Vergata, 00133 Roma, Italy (e-mail: a.astolfi@imperial.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2022.3199211.

Digital Object Identifier 10.1109/TAC.2022.3199211

***Index Terms*—Iterative learning methods, nonlinear systems, optimal control.**

I. Introduction

The optimal control problem consists in designing a control law that steers the state of the system from an arbitrary initial condition to a final equilibrium configuration while minimizing—along the trajectories of the resulting system—a prescribed cost functional [1], [2], [3], [4], [5]. Two alternative strategies to solve optimal control problems have been studied in the past decades: Dynamic Programming [6], [7] and Pontryagin’s Minimum Principle [8]. In the setting of infinite-horizon problems, the two approaches share in fact a common *bottleneck*, namely the explicit knowledge of the positive definite solution of the Hamilton–Jacobi–Bellman (HJB) partial differential equation (pde) (see, e.g., [1], [9]): in the former method, the solution to the pde directly yields the optimal feedback, while in the latter, it provides the correct initial condition for the costate variable of the underlying Hamiltonian dynamics.

Methods based on a direct solution (or approximation, see, for instance, [9], [10], [11], [12]) of the HJB pde have been interpreted hitherto as offline strategies, namely requiring the knowledge (or an estimate) of the solution *prior* to the evolution of the plant or the simulation model. Recent intensive research efforts have been devoted to the objective of recasting classic optimal control design techniques into online methods. In particular, iterative learning control (ILC) and reinforcement learning (RL) [13], [14] aim at *learning* the optimal control law online by conducting a sufficiently large number of experiments on the actual plant, if feasible, or by relying on a simulation model of the former. This is achieved, for instance, by borrowing ideas similar to those employed in the context of adaptive control, leading to the so-called adaptive dynamic programming [15], [16], [17]. The methods can be further categorized by distinguishing model-based techniques [18], [19] and model-free approaches [20].

Policy iteration (PI) approaches [13]—developed both in the context of model-based and model-free methods—encompass several iterative techniques that aim at providing (or approximating) the solution to the underlying HJB pde by relying on a sequence of (stabilizing) control laws designed to converge in some sense to the optimal policy. In the case of linear-quadratic (LQ) optimal control problems, for instance, the well-known

Kleinman's Lyapunov iterations [21] describe essentially a PI approach to the solution of the optimal control design. In fact, PI strategies are typically based on two main steps, *policy evaluation* and *policy improvement*, which may be intertwined in a discrete time [22] or synchronous [18] fashion. The aim of the former step consists in evaluating the *value* of the current control policy, whereas the latter stage should lead to an improvement of such a value by suitably modifying the control law for the subsequent iteration.

While the policy improvement phase is carried out by relying on fairly standard methods (such as, for instance, gradient-descent approaches), the design choice of the *measure* devoted to capturing the quality of the current control policy is crucial to envision effective approximation mechanisms. In most of the results in the literature, the above-mentioned measure is provided essentially by the Hamiltonian function evaluated along trajectories of the closed-loop system, typically referred to as *temporal difference error* in the framework of RL and related to the so-called "reinforcement" employed for learning. An intrinsically similar situation arises also in the integral reinforcement learning (IRL) framework, see, e.g., [23], where the role of the temporal difference error, namely the approximate Bellman error, is inherited by an integrated (over a moving window) version of the Hamiltonian function along trajectories of the system, with the aim of circumventing the need for the explicit knowledge of the drift vector field in the overall architecture. As a consequence, whenever the *candidate* value function is replaced by a linearly parameterized functional approximator, the computation of the temporal difference error yields a scalar (linear) equation in a certain number of unknowns. Therefore, since the latter relation does not contain *sufficient information* to characterize the correct values of the underlying coefficients, it is necessary to collect a certain number of samples of such *reinforcement* in a sequential implementation (see, e.g., [22]) or to monitor—over a receding window of prescribed length—such an error signal, which must then remain *persistently exciting*, within a synchronous learning framework (see, e.g., [18], [23]). The conflicting requirements of persistence of excitation and asymptotic stability lead to the need, on one hand, for the injection of additional (probing) control inputs to the closed-loop system. On the other hand, the sampling rate or the length of the moving window must be suitably selected to ensure that the collected data is sufficiently rich for the computation of the correct value of the approximating parameterization. Note that the above rewarding policy for the learning algorithm is a *common feature at the core of the majority of the existing methods*, regardless of their nature (i.e., model-based or model-free, synchronous or with discrete updates).

In the context of model-based, iterative learning strategies for continuous-time nonlinear systems, the main contribution of this article consists in suggesting an alternative definition and practical construction of the *temporal difference error* with respect to the one on which the majority of RL and ILC approaches are intrinsically based. The result is achieved in sequential steps that are interesting *per se*, as discussed in the following.

First, it is shown that—provided the solution to the HJB pde is known—the stable manifold of the Hamiltonian dynamics can be *externally stabilized* by suitably steering the costate dynamics via a virtual control input, while preserving the property of invariance of the latter manifold as well as the behavior of the Hamiltonian dynamics restricted therein. This preliminary result possesses interesting consequences on the implementation of open-loop optimal control laws, typically employed in practical applications. In fact, whenever only the initial condition of the plant is available, open-loop optimal control laws are computed by considering the forward propagation of the underlying Hamiltonian dynamics initialized on the stable invariant manifold. Since the latter invariant manifold is structurally *externally unstable*, the above strategy is known to be particularly *fragile*, even with respect to extremely small numerical errors in the computation of the correct initial condition, and almost impractical in applications. The implementation of such a *stabilized costate dynamics* permits the construction of numerically robust open-loop control laws, which approximate the optimal feedback with arbitrary degree of accuracy.

Then, we propose an alternative definition of the temporal difference error with respect to the one at the basis of any reinforcement or iterative learning method, based on the following, intuitive, observation. Suppose that an approximate value function is employed in the construction of the aforementioned stabilizing control law. It can be easily shown that, on one hand, the property of *asymptotic stability* of the zero equilibrium of the controlled Hamiltonian dynamics is preserved also for (sufficiently close) estimates of the solution to the HJB pde, whereas, on the other hand, the (*fragile*) property of *invariance* is not retained. Therefore, a candidate measure to capture the accuracy of the approximation of the current estimate is naturally provided by the maximal distance from the induced (approximate) manifold along the trajectories of the closed-loop (stabilized) Hamiltonian dynamics. This measure provides an alternative to more "standard" temporal difference errors. To circumvent the daunting computational task of a direct solution and in the spirit of iterative and episodic learning strategies, the above information is exploited by sequentially updating the approximate value function toward the minimization of the considered temporal difference error.

The rest of this article is organized as follows. The problem statement and a few preliminaries are discussed in Section II. The aim of Section III is to suggest and design a (virtual) control architecture for the underlying Hamiltonian dynamics, which is then employed in Section IV to propose an alternative temporal different error for IL methods. These are then specialized to the case of LQ optimal control problems, before the theory related to nonlinear systems is validated by means of a physically motivated numerical simulation involving an automatic flight control problem in Section VI. Finally, Section VII concludes this article.

Notation: $\mathbb{R}_{\geqslant 0}$ (resp. $\mathbb{R}_{>0}$) denotes the set of nonnegative (resp. positive) real numbers. Given $f : \mathbb{R}^n \to \mathbb{R}$, the mappings ∇f and $\frac{\partial f}{\partial x}$ denote the column and the row vectors, respectively, of the corresponding partial derivatives, whereas $\frac{\partial^2 f}{\partial x^2}$ defines the Hessian matrix of the second-order derivatives provided they exist. For a vector-valued function $g : \mathbb{R}^n \to \mathbb{R}^m$, $\frac{\partial g}{\partial x}$ denotes the Jacobian matrix of first-order partial derivatives. The notation

$\mathcal{C}^\kappa$, $\kappa \geqslant 0$, defines the set of functions that admit continuous derivatives up to order κ. Given $r \in \mathbb{R}_{>0}$ and $\bar{x} \in \mathbb{R}^n$, the notation $\mathbb{B}_r(\bar{x})$ defines the open set $\{x \in \mathbb{R}^n : \|x - \bar{x}\| < r\}$.

II. Preliminaries and Problem Definition

Consider continuous-time, time-invariant, nonlinear systems described by the equation

$$\dot{x} = f(x) + g(x)u\,, \quad x(0) = x_0 \tag{1}$$

with $x : \mathbb{R} \to \mathbb{R}^n$ and $u : \mathbb{R} \to \mathbb{R}^m$ denoting the state and the control input, respectively. It is assumed that the vector field $f : \mathbb{R}^n \to \mathbb{R}^n$ and the mapping $g : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ are sufficiently smooth and that, in the absence of control action, the origin of $\mathbb{R}^n$ is an equilibrium point of (1), i.e. $f(0) = 0$. The trajectories of the system are evaluated and compared by means of the infinite-horizon cost functional defined by

$$\begin{aligned} J(u) &= \frac{1}{2}\int_0^\infty (q(x(t)) + \|u(t)\|_R^2)dt \\ &=: \int_0^\infty r(x(t), u(t))dt \end{aligned} \tag{2}$$

where $q : \mathbb{R}^n \to \mathbb{R}_{\geqslant 0}$, sufficiently smooth, represents a running cost imposed on the state evolution and the second term, with $R = R^\top > 0$, represents a penalty on the control effort. Let

$$A := \frac{\partial f}{\partial x}\Big|_{x=0} \tag{3}$$

and $B := g(0)$ describe the linearization around the origin of the dynamics (1), namely $\dot{\delta x} = A\delta x + Bu$, let $Q := \frac{\partial^2 q}{\partial x^2}(x)|_{x=0}$ and consider the following, standard structural assumption.

Assumption 1: The pairs (A, B) and (A, Q) are reachable and observable, respectively. ○

Under the given conditions it is known that (locally) the *value function* V associated with the optimal control problem is at least twice differentiable and there exists a continuous function u that minimises the cost function (2) subject to the dynamic constraint (1).

Let $V : \mathbb{R}^n \to \mathbb{R}_{\geqslant 0}, V \in \mathcal{C}^\kappa, \kappa \geqslant 2$, denote a positive definite solution of the HJB pde

$$0 = \frac{1}{2}q(x) + \nabla V(x)^\top f(x) - \frac{1}{2}\nabla V(x)^\top g(x)R^{-1}g(x)^\top \nabla V(x) \tag{4}$$

with $V(0) = 0$, for all $x \in \mathbb{R}^n$ (or locally in a neighborhood of the origin). Then, V represents the value function associated with the problem and the optimal state feedback, which minimizes the cost functional (2) along the trajectories of the closed-loop system, is given by

$$u_f^\star(x) = -R^{-1}g(x)^\top \nabla V(x)\,. \tag{5}$$

As such, the control design method based on (4) and (5) constitutes an offline technique, in which the explicit knowledge of the positive definite solution of (4) is a priori instrumental for the construction of the optimal feedback (5). To circumvent the computational burden of a direct solution of the HJB pde, several alternative strategies have been explored to recast the control design into an online strategy, regardless of any assumption on the full or partial knowledge of the underlying plant. In particular, the PI approach consists of two sequential steps, i.e., *policy evaluation*—in which an estimate $\hat{V}$ of the value (i.e., the cost in a minimization problem) of the current approximate feedback is provided and quantified by a *temporal difference error* in the context of RL—and *policy improvement*—in which the estimate $\hat{V}$, and hence, the approximate feedback $\hat{u}^{k+1} = u_f^\star|_{V=\hat{V}^k}$, is updated (in continuous or discrete time) to reduce the approximation error, where k denotes the index of the iteration. In the vast majority of these methods, the *temporal difference error*, on which learning is based, is envisioned on the basis of the following consideration. Let $\mathcal{H}^p : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$ denote the *pre-minimized* Hamiltonian function, defined as $\mathcal{H}^p(x, u, \lambda) = r(x, u) + \lambda^\top(f(x) + g(x)u)$. It has been shown that the solutions of the HJB pde satisfy the condition

$$\mathcal{H}^p(x(t), u_f^\star(t), \nabla V(x(t))) = 0$$

for all $t \geqslant 0$. Therefore, if the value function V is replaced by an estimate $\hat{V}$ (e.g., described in terms of a functional approximator), a *measure* of the accuracy of such an approximation may be suggested by monitoring the *temporal difference error*

$$e(t) := \mathcal{H}^p(x(t), \hat{u}^k(t), \nabla\hat{V}^k(x(t))) \tag{6}$$

(or its integral over a moving window as in IRL [23]) along the trajectories of the closed-loop system. The objective of this manuscript is to propose an alternative definition of the temporal difference error. Toward this end, recall that the optimal feedback (5) can be equivalently designed and implemented as the output $u^\star(t) = -R^{-1}g(x(t))^\top\lambda(t)$ of the Hamiltonian dynamics

$$\begin{bmatrix} \dot{x} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} \nabla_x \mathcal{H} \\ \nabla_\lambda \mathcal{H} \end{bmatrix} := \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \nabla\mathcal{H}(x, \lambda) \tag{7}$$

with $(x(0), \lambda(0)) = (x_0, \nabla V(x_0))$, where the (minimized) Hamiltonian function $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\mathcal{H}(x, \lambda) = \frac{1}{2}q(x) + \lambda^\top f(x) - \frac{1}{2}\lambda^\top g(x)R^{-1}g(x)^\top\lambda\,. \tag{8}$$

The rationale of the proposed alternative definition of the temporal difference error is that the Hamiltonian function $\mathcal{H}$, obtained by replacing the actual value function V with an approximation thereof, provides sufficient information to completely characterize the underlying parameters of $\hat{V}$ *at a single time instant*. This is due to the dependence of the dynamics as well as of the initial conditions of the Hamiltonian dynamics on the approximating coefficients and the alternative definition is achieved, essentially, by determining the maximal error over a certain time window. Although this strategy has not been pursued hitherto, it is worth observing that a similar result may be obtained by summarizing the entire information of the classic temporal difference error in IRL with its maximal value over time. However, the ability of exploiting the above-mentioned logic may be hindered by the following consideration, which concerns the instability of a certain invariant manifold and which concludes this preliminary section. This provides in addition a motivation for the technical results in Section III.

The Hamiltonian system (7) represents the *lifted* system defined on the state/costate space. Assumption 1 implies that the Hamiltonian dynamics (7) possess a *hyperbolic* equilibrium point at $(x, \lambda) = (0, 0)$ with n-dimensional stable $\mathcal{N}_s$ and unstable $\mathcal{N}_u$ submanifolds through the origin that are invariant for system (7) (see for instance the detailed discussion in [24] and [25] in the context of disturbance attenuation problems for nonlinear systems, in which a similar pde arises). Therefore, a straightforward consequence of the above discussion is that for any approximate $\hat{V}$ such that, necessarily, $\hat{V}(x(\cdot))$ is different from the corresponding time evolution of the optimal value function $V(x(\cdot))$ one has that any temporal difference error based on the evaluation of the Hamiltonian function $\mathcal{H}(x, \lambda)$ over time would be an exponentially diverging function of time. This issue is circumvented in the following section, whereas the definition of the proposed temporal difference error is deferred to Section IV.

III. External Stabilization of the Stable Invariant Submanifold

As anticipated above, the main objective of this section is to discuss preliminary results concerning the Hamiltonian dynamics (7) such that the latter can be employed to define and construct an alternative *temporal difference error*, which quantifies the accuracy of an intermediate estimate of the optimal value function. For clarity of exposition it is convenient to explicitly describe the Hamiltonian dynamics (7), with respect to the Hamiltonian function $\mathcal{H}$ defined in (8), as

$$\begin{aligned}\dot{x} &= f(x) - g(x)R^{-1}g(x)^\top\lambda \\ &:= f_1(x, \lambda) \\ \dot{\lambda} &= -\frac{\partial f}{\partial x}(x)^\top\lambda - \frac{1}{2}\nabla_x\left(q(x) - \lambda^\top g(x)R^{-1}g(x)^\top\lambda\right) \\ &:= f_2(x, \lambda)\end{aligned} \tag{9}$$

with $(x(0), \lambda(0)) = (x_0, \nabla V(x_0))$. Define the manifold $\mathcal{M} := \{(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^n : \lambda - \nabla V(x) = 0\}$. Then, it can be shown that $\mathcal{M}$ is an *invariant*, *externally unstable*, manifold for (9), since $\mathcal{N}_s = \text{graph}(\nabla V(x)) = \mathcal{M}$ and $\mathcal{N}_s$ is tangent at the origin to the stable subspace of the linearized Hamiltonian dynamics [24], [25].

To provide a concise statement of the following result, consider the *controlled Hamiltonian dynamics*

$$\begin{aligned}\dot{x} &= f_1(x, \eta) \\ \dot{\eta} &= f_2(x, \eta) + v\end{aligned} \tag{10}$$

where $v : \mathbb{R} \to \mathbb{R}^n$ is a (virtual) control input, to be designed, in charge of arbitrarily steering the costate variable η. Note that the change in the notation between the costate of (9) (i.e., λ) and of (10) (i.e., η) underlines the different evolution of the former with respect to the latter, since they satisfy distinct dynamics, with the latter being steered by the choice of v. The main result of this section, summarized in the statement below, provides a selection of the control input v such that the trajectories of the controlled Hamiltonian dynamics (10) approximate, arbitrarily close, those of the Hamiltonian dynamics (7) [or, equivalently, (9)]. Namely, the control input is such that η is arbitrarily close to λ, while ensuring that $\mathcal{M}$ is in fact an invariant, *externally stable* manifold for (10).

Theorem 1: Consider system (10) in closed loop with

$$\begin{aligned}v^\star(x, \eta, \nabla V(x)) = &-\frac{1}{2}\nabla_x(\pi^\top g(x)R^{-1}g(x)^\top\pi)|_{\pi=\eta-\nabla V(x)} \\ &+ \left(\frac{\partial}{\partial x}f_1(x, \nabla V(x))^\top - \sigma_F I\right)(\eta - \nabla V(x)).\end{aligned} \tag{11}$$

Then, for any $t^\star \in \mathbb{R}_{>0}$, $\varepsilon \in \mathbb{R}_{>0}$ and $\mu \in \mathbb{R}_{>0}$, there exists $\sigma_F \in \mathbb{R}_{>0}$ such that[1]

(i) $\mathcal{M}$ is an invariant and externally (locally) exponentially stable manifold for (10), (11);

(ii) for all $(x(0), \eta(0))$ such that $\|\eta(0) - \nabla V(x(0))\| < \mu$ and all $t > t^\star$, the trajectories of (10), (11) *are such that*

$$\|\eta(t) - \lambda(t)\| < \varepsilon$$

where λ denotes the solution to (9) *initialized at $(x_0, \nabla V(x_0))$, namely the* optimal *costate.* ○

Proof: To begin with, since the HJB equation (4) holds for all $x \in \mathbb{R}^n$, it follows that

$$\begin{aligned}0 = &\frac{1}{2}\nabla q(x) + \frac{\partial^2 V}{\partial x^2}(x)f(x) + \frac{\partial f}{\partial x}(x)^\top\nabla V(x) \\ &- \frac{1}{2}\nabla_x(\nabla V(x)^\top g(x)R^{-1}g(x)^\top\nabla V(x))\end{aligned} \tag{12}$$

which is obtained by differentiating the right-hand side of (4) with respect to x. Consider then the change of coordinates described by $(z_1, z_2) = \Phi(x, \eta) = (x, \eta - \nabla V(x))$, which immediately yields the inverse transformation $\Phi^{-1}(z) = (z_1, z_2 + \nabla_{z_1}V(z_1))$. In the transformed coordinates, the dynamics of the first component of the state becomes

$$\begin{aligned}\dot{z}_1 &= f_1(z_1, z_2 + \nabla_{z_1}V(z_1)) \\ &= f_1(z_1, \nabla_{z_1}V(z_1)) - g(z_1)R^{-1}g(z_1)^\top z_2\end{aligned}$$

where the second equality is obtained by recalling that the function $f_1(x, \cdot)$ is affine for any $x \in \mathbb{R}^n$. Moreover, the dynamics of z_2 is derived as shown in (13) (overleaf),

$$\begin{aligned}\dot{z}_2 = &f_2(z_1, z_2 + \nabla_{z_1}V(z_1)) + v \\ &- \frac{\partial^2 V}{\partial z_1^2}(z_1)f_1(z_1, z_2 + \nabla_{z_1}V(z_1)) \\ = &v - \frac{\partial f}{\partial z_1}(z_1)^\top(z_2 + \nabla_{z_1}V(z_1)) - \tfrac{1}{2}\nabla_{z_1}q(z_1) \\ &+ \tfrac{1}{2}\nabla_{z_1}(\pi^\top g(z_1)R^{-1}g(z_1)^\top\pi)|_{\pi=z_2+\nabla_{z_1}V(z_1)} \\ &- \frac{\partial^2 V}{\partial z_1^2}(z_1)(f(z_1) - g(z_1)R^{-1}g(z_1)^\top(z_2 + \nabla_{z_1}V(z_1))) \\ = &v - \frac{\partial f}{\partial z_1}(z_1)^\top z_2 + \nabla_{z_1}(\pi^\top g(z_1)R^{-1}g(z_1)^\top)|_{\pi=\nabla_{z_1}V(z_1)}\end{aligned}$$

[1] By slightly abusing the notation the manifold $\mathcal{M}$ in item (i) must be interpreted with the costate λ replaced by the variable η as in (10).

$$\begin{aligned}
&z_2+\frac{\partial^2 V}{\partial z_1^2}(z_1)g(z_1)R^{-1}g(z_1)^\top z_2\\
&\quad+\tfrac{1}{2}\nabla_{z_1}(\pi g(z_1)R^{-1}g(z_1)^\top\pi)|_{\pi=z_2}\\
&=v-\frac{\partial}{\partial z_1}f_1(z_1,\nabla_{z_1}V(z_1))^\top z_2\\
&\quad+\tfrac{1}{2}\nabla_{z_1}(\pi g(z_1)R^{-1}g(z_1)^\top\pi)|_{\pi=z_2} \qquad (13)
\end{aligned}$$

where the third equality is obtained by recalling that (12) holds and the last equality is derived by the definition of the vector field f_1 as in (9) and noting that

$$\begin{aligned}
\frac{\partial}{\partial x}f_1(x,\nabla V(x))^\top = \frac{\partial f}{\partial x}(x)^\top + \frac{\partial^2 V}{\partial x^2}(x)g(x)R^{-1}g(x)^\top\\
-\nabla_x(\pi^\top g(x)R^{-1}g(x)^\top)|_{\pi=\nabla V(x)}\,.
\end{aligned}$$

Therefore, by replacing the control law $v^\star$, defined in (11), into the dynamics (13) one has that $\dot z_2 = -\sigma_F z_2$, hence

$$z_2(t)=\eta(t)-\nabla V(x(t)) = e^{-\sigma_F t}(\eta(0)-\nabla V(x_0))\,. \quad (14)$$

This, in turn, immediately implies item *(i)* of the statement by recalling the definition of the submanifold $\mathcal{M}$ for (10).

Let now x_λ denote the solution of (7) with initial condition $x_\lambda(0)=x_0$, namely such that $\dot x_\lambda = f(x_\lambda) - g(x_\lambda)R^{-1}g(x_\lambda)^\top\lambda$, where $\lambda(t)=\nabla V(x_\lambda(t))$ for all $t\geqslant 0$. Therefore, x_λ verifies

$$\dot x_\lambda = f_1(x_\lambda,\nabla V(x_\lambda))\,. \quad (15)$$

Moreover, since by (14) $\eta(t)=\nabla V(x(t))+e^{-\sigma_F t}\xi_0$, with $\xi_0 := \eta(0)-\nabla V(x_0)$, note that

$$\begin{aligned}
\|\eta(t)-\lambda(t)\| &= \|\nabla V(x(t))-\nabla V(x_\lambda(t))+e^{-\sigma_F t}\xi_0\|\\
&\leqslant \|\nabla V(x(t))-\nabla V(x_\lambda(t))\| + e^{-\sigma_F t}\|\xi_0\| \qquad (16)
\end{aligned}$$

where x satisfies the equation

$$\begin{aligned}
\dot x(t) &= f(x(t)) - g(x(t))g(x(t))^\top\eta(t)\\
&= f(x(t)) - g(x(t))R^{-1}g(x(t))^\top(\nabla V(x(t))+e^{-\sigma_F t}\xi_0)\\
&= f_1(x(t),\nabla V(x(t))) - g(x(t))R^{-1}g(x(t))^\top\xi_0 e^{-\sigma_F t}\,. \qquad (17)
\end{aligned}$$

Therefore, item (ii) of the claim follows by considering (15) and (17) and applying Lemma 5 in the Appendix, with $f(x) = f_1(x,\nabla V(x))$ and $s(x) = -g(x)R^{-1}g(x)^\top\xi_0$. Namely, noting that $\|x_\lambda(0)-x(0)\|=0$, by Lemma 5, for any $\varepsilon_x>0$, there exists $\sigma_F>0$ such that $\|x_\lambda(t)-x(t)\|\leqslant\varepsilon_x$ for all $t\geqslant 0$. By continuity of the mapping ∇V this in turn implies that for any $\varepsilon_V>0$ there exists $\sigma_F>0$ such that $\|\nabla V(x(t)) - \nabla V(x_\lambda(t))\|\leqslant\varepsilon_V$, for all $t\geqslant 0$. Thus, letting $\varepsilon_V=\varepsilon/2$, there exists σ_F such that the first term of the second line of (16) satisfies $\|\nabla V(x(t))-\nabla V(x_\lambda(t))\|\leqslant\varepsilon/2$ for all $t\geqslant 0$. Imposing the additional constraint that $\sigma_F > \dfrac{1}{t^\star}\ln(\varepsilon/2\|\xi_0\|)$, it follows that the second term in (16) satisfies the condition $e^{-\sigma_F t}\|\xi_0\|<\varepsilon/2$, for all $t\geqslant t^\star$, from which item (ii) follows. ■

Theorem 1 provides a *robustified* mechanism to compute the optimal costate. A further consequence of the constructions discussed in the statement—resulting essentially from the fact that the behavior on the manifold $\mathcal{M}$ is not modified by the feedback control input $v^\star$ and from Assumption 1—is summarized in the following statement.

Proposition 1: The origin $(x,\eta)=(0,0)\in\mathbb{R}^{2n}$ is a locally exponentially stable equilibrium point for system (10) in closed loop with (11). ○

Proof: In the transformed coordinates the closed-loop system is described by the equations

$$\dot z_1 = f_1(z_1,\nabla_{z_1}V(z_1)) - g(z_1)R^{-1}g(z_1)^\top z_2 \quad (18a)$$

$$\dot z_2 = -\sigma_F z_2\,. \quad (18b)$$

By the structural requirements of Assumption 1 the manifold $\mathcal{M}$ is tangent, locally around the origin of the extended state/costate space, to the stable n-dimensional invariant subspace of the linearized Hamiltonian dynamics. Therefore, it follows that the linearization of the vector field $f_1(z_1,\nabla_{z_1}V(z_1))$ in (18a) has all eigenvalues with negative real part since V solves the underlying HJB pde. The claim is then shown by noting that the dynamics in (18b) is linear, with $-\sigma_F I$ a Hurwitz matrix. Consequently, the overall linearized description of (18) is an upper triangular block matrix, which is Hurwitz. ■

Example 1. Consider the nonlinear system $\dot x = x^2+u$, with $x(t)\in\mathbb{R}$ and $u(t)\in\mathbb{R}$, and let $q(x)=x^2$ in the cost functional (2). Then, the HJB pde is solved by the positive definite function $V(x)=\frac{1}{3}(x^2+1)^{(3/2)}+\frac{1}{3}x^3-\frac{1}{3}$. Hence, $\nabla V(x) = x(\sqrt{x^2+1}+x)$. The graph of the function ∇V, $x\in\mathbb{R}$—which coincides with the stable invariant submanifold $\mathcal{N}_s$ of the underlying Hamiltonian dynamics—is depicted by the dotted gray line in Fig. 1. The dashed gray line displays the phase plot of the solution of the Hamiltonian dynamics (7) (correctly) initialized at $(x_0,\nabla V(x_0))$. It can be appreciated that the trajectory, simulated in the time interval [0,30], does not converge to the origin, even setting the relative and absolute tolerance to 10^{-12} in the MATLAB routine *ode45*. Namely, due to the manifold $\mathcal{M}$ being externally unstable, (unavoidable) numerical errors cause the trajectory to diverge. Instead, the solid black line displays the phase plot of the controlled Hamiltonian dynamics (10), with a randomly generated initial condition $\eta(0)$, in closed loop with (11). It is worth observing that, even for a random initial condition $\eta(0)$, the optimal costate is recovered arbitrarily fast via the selection of $\sigma_F=100$ and the trajectory *robustly* converges to the origin, since the invariant submanifold $\{(x,\lambda):\lambda=\nabla V(x)\}$ is rendered externally exponentially stable. △

IV. PI VIA CONTROLLED HAMILTONIAN DYNAMICS

The objective of this section consists in discussing how the results of Theorem 1, which in principle relies on the knowledge

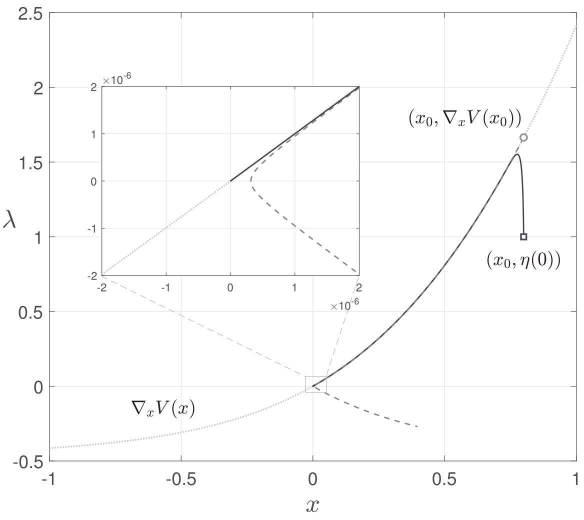

Fig. 1. Phase plot of the Hamiltonian dynamics (7) (dashed gray line) and of the controlled Hamiltonian dynamics (10), (11) (solid black line), together with the stable invariant submanifold $\mathcal{N}_s$ described by the graph of V_x (dotted gray line).

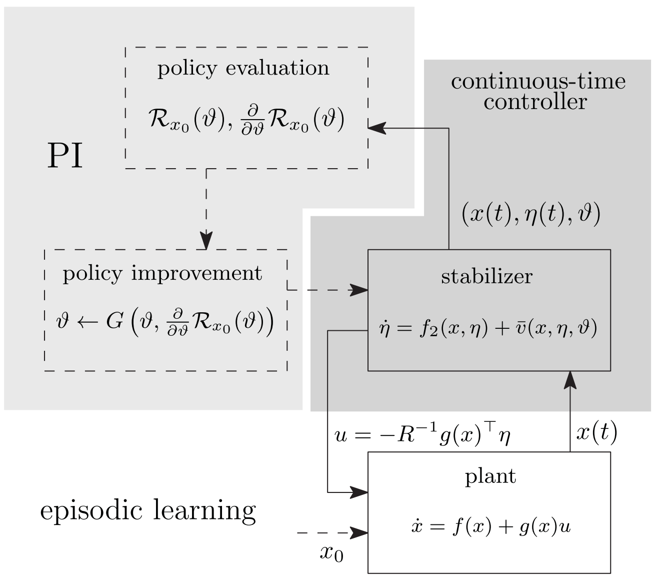
PI
policy evaluation
$\mathcal{R}_{x_0}(\vartheta), \frac{\partial}{\partial\vartheta}\mathcal{R}_{x_0}(\vartheta)$
continuous-time controller
$(x(t), \eta(t), \vartheta)$
policy improvement
$\vartheta \leftarrow G\left(\vartheta, \frac{\partial}{\partial\vartheta}\mathcal{R}_{x_0}(\vartheta)\right)$
stabilizer
$\dot{\eta} = f_2(x,\eta) + \bar{v}(x,\eta,\vartheta)$
$u = -R^{-1}g(x)^\top \eta$
$x(t)$
plant
$\dot{x} = f(x) + g(x)u$
episodic learning
x_0

Fig. 2. Schematic description of the model-based, iterative learning architecture. Solid and dashed blocks/lines represent continuous-time actions and discrete-time updates, respectively. At the end of each episode the initial condition of the plant is reset to x_0.

of V, can be employed to construct an approximate optimal control policy. Inspired by the rationale behind PI architectures, the proposed approach consists of a policy evaluation phase, followed by a policy improvement (discrete time) step. The overall iterative strategy (illustrated in Fig. 2) relies on a finite-dimensional parameterization of candidate *value functions* by means of value function approximators (VFAs), such as (single-layer) neural networks with a polynomial basis, see, e.g., [19]. Therefore, the suggested approximator may be equivalently interpreted as a sum-of-squares (SOS) estimate of the underlying value function.

To this end, let $\mathfrak{c}(x) := x^{\{m\}}$ denote a vector containing a basis for the monomials of degree less than or equal to m with respect to the variable $x \in \mathbb{R}^n$. Note that $\mathfrak{c}(x) \in \mathbb{R}^\nu$ with

$$\nu = \binom{n+m}{m} \tag{19}$$

where the notation on the right-hand side describes the binomial coefficient of $n+m$ over m. In the spirit of (model-based) online PI schemes, define the function $\bar{V} : \mathbb{R}^\mu \times \mathbb{R}^n \to \mathbb{R}$ as

$$\bar{V}(\vartheta, x) := \mathfrak{c}^\top(x)\Theta\mathfrak{c}(x) = \sum_{i=1}^{\mu} \vartheta_i \bar{\mathfrak{c}}_i(x) \tag{20}$$

with $\Theta \in \mathbb{R}^{\nu\times\nu}$ a symmetric and positive definite matrix. In (20), the parameter $\vartheta \in \mathbb{R}^\mu$, $\mu = \nu(\nu+1)/2$, denotes a vector containing the entries of the matrix Θ by columns, namely $\vartheta = \mathrm{vech}(\Theta)$, where vech denotes the *half-vectorization* operator for symmetric matrices, whereas $\bar{\mathfrak{c}}_i(x)$ describes the ith element of the vector-valued function $(\mathfrak{c}(x) \otimes_{\mathbb{S}} \mathfrak{c}(x))$, with $\otimes_{\mathbb{S}}$ denoting the symmetric Kronecker product (see [26] for more detailed discussions and insights). Note that $\bar{V}$ in (20) describes a single layer, linear neural network with polynomial activation functions $\bar{\mathfrak{c}}_i$, $i = 1, \ldots, \mu$.

Furthermore—to avoid cumbersome notation in the derivations below that involve (mixed) second-order derivatives of the function $\bar{V}$ with respect to ϑ and x—the following compact notation is introduced. Let $h(\vartheta, x) = \nabla_x \bar{V}(\vartheta, x)$ and define the approximate (virtual) control input

$$\bar{v} = v^\star|_{\nabla V(x) = h(\vartheta, x)} \tag{21}$$

namely the feedback control input (11) obtained by replacing the gradient of the optimal value function with the parameterized function h. The following sections provide the definition and some properties of a finite-dimensional cost function $\mathcal{R}_{x_0} : \mathbb{R}^\mu \to \mathbb{R}_{\geqslant 0}$ that is shown to be instrumental for the construction, or the approximation, of optimal control laws. The rationale behind the formal derivations below can be intuitively anticipated as follows: the statement of Theorem 1 entails that the implementation of $v^\star$ to the controlled Hamiltonian dynamics (10) implies that the manifold $\mathcal{M}$ is rendered externally LES, while preserving its underlying *invariance* property. The implementation of a *generic* $\bar{v}$ instead—provided the feedback $\bar{v}$ is *sufficiently close* to $v^\star$—preserves the attractivity property of the manifold $\mathcal{M}$ (due to the intrinsic robustness of Lyapunov's asymptotic stability to uncertainties) while, however, compromising the (fragile) property of invariance. Therefore, the main idea below is to employ the *distance* of the ensuing trajectories from the manifold $\mathcal{M}$ as a *reward value* at each *episode* of the learning process. In particular, given a certain initial condition $x_0 \in \mathbb{R}^n$, an *episode* is defined as a sufficiently long time interval in which the trajectories of the plant—ensuing from $x_0 \in \mathbb{R}^n$ and in closed loop with the current estimate of $v^\star$—are monitored online. Such a scalar reward is subsequently employed to measure the difference between $\bar{v}$ and $v^\star$ and, consequently, to adjust the value of the parameter[2] ϑ to minimize such a distance.

[2] It is worth observing at this stage that—while the derivations in this manuscript have been carried out by considering a SOS approximation of the

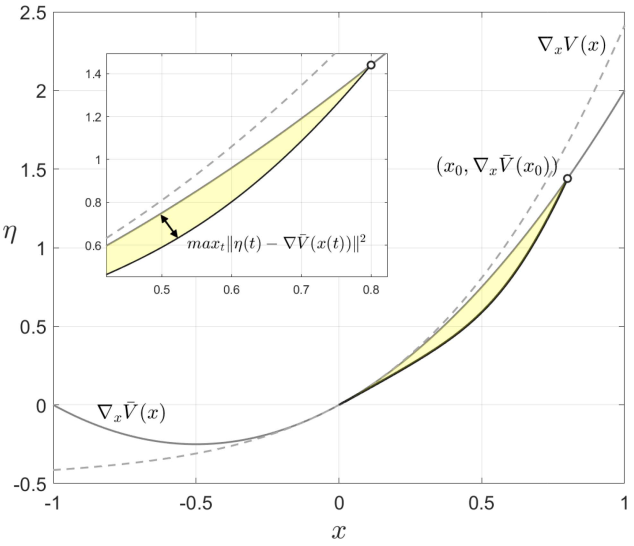

Fig. 3. Graphical illustration of the distance from invariance of the manifold $\{(x,\eta) : \eta - \nabla\bar{V}(x) = 0\}$ induced by the approximate value function $\bar{V} = (1/2)x^2 + (1/3)x^3$. The solid and dashed gray lines describes the graph of $\nabla\bar{V}$ and ∇V, respectively. The solid black line indicates the trajectory of system (10) in closed loop with $\bar{v}$.

Note that, differently from alternative PI schemes, by the nature of the episode and of the corresponding reward proposed herein, the measured signal is not required to be sufficiently rich to obtain an *informative* measure, nor is the selection of the *length of the episode* crucial for the convergence of the algorithm. The intuition above is illustrated and motivated at this preliminary stage by means of the following numerical simulation.

Example 2: Consider again the setting in Example 1 and the approximate value function provided by the third-order Taylor's expansion of V, namely $\bar{V}(x) = (1/2)x^2 + (1/3)x^3$. Fig. 3 provides a graphical illustration of the distance from invariance of the manifold $\bar{\mathcal{M}} := \{(x,\eta) : \eta - \nabla\bar{V}(x) = 0\}$ (solid gray line) induced by the approximate value function $\bar{V} = (1/2)x^2 + (1/3)x^3$. The trajectory (solid black line) of system (10) in closed loop with $\bar{v}$ ensuing from an initial condition $(x_0, \eta_0) \in \bar{\mathcal{M}}$ shows that the manifold is not invariant, although externally attractive. △

A. Maximal Distance From Invariance as Policy Evaluation

In this section, we introduce and characterize the reward signal of the model-based RL architecture that is employed to measure the distance between $\bar{v}$ and $v^\star$ along trajectories of the closed-loop state/costate dynamics. To this end, consider the finite-dimensional cost function $\mathcal{R} : \mathbb{R}^\mu \to \mathbb{R}$ defined as

$$\mathcal{R}_{x_0}(\vartheta) := \max_{t\in\mathbb{R}_{\geqslant 0}} \|\eta(t) - h(\vartheta, x(t))\|^2 \tag{22}$$

parameterized with respect to the initial condition $x_0 \in \mathbb{R}^n$. Note that the dependence of the function $\mathcal{R}_{x_0}$ on its argument ϑ is threefold, the first one being the, immediately evident, direct dependence of the function h on ϑ. Then, since $(x(t), \eta(t))$ appearing in $\mathcal{R}_{x_0}$ denote the trajectories of the controlled Hamiltonian dynamics (10) in closed loop with $\bar{v}(x, \eta, h(\vartheta, x))$ and initialized at $(x(0), \eta(0)) = (x_0, h(\vartheta, x_0))$, one has that $\mathcal{R}_{x_0}$ depends on ϑ also via the fact that the underlying (closed-loop) vector field is parameterized with respect to ϑ as well as via the specific selection of the initial conditions.

The results of this section characterize certain properties of the function $\vartheta \mapsto \mathcal{R}_{x_0}(\vartheta)$, which are derived by relying on the following standing assumption.

Assumption 2: Fix $x_0 \in \mathbb{R}^n$. Let the value of ν in (20) be given and define

$$\bar{\vartheta} := \arg\min_{\vartheta} \mathcal{R}_{x_0}(\vartheta)\,. \tag{23}$$

The origin is a LES equilibrium point of system (10) in closed loop with $\bar{v}|_{\vartheta=\bar{\vartheta}}$ in (11), (21) and $(x_0, h(\bar{\vartheta}, x_0)) \in \mathbb{R}^n \times \mathbb{R}^n$ belongs to the basin of attraction of the origin. ○

Note that Assumption 2 is verified provided μ is selected sufficiently large, since it is known that the approximation errors of $\bar{V}$ with respect to V as well as the corresponding partial derivatives are uniformly bounded in a compact set and converge uniformly to zero provided ν tends to infinity, see, e.g., [27]. Moreover, the assumption is trivially satisfied whenever the positive definite function V, solution of the HJB pde (4) is in fact a SOS with highest degree smaller than or equal to m in (20). The latter condition ensures the existence of $\vartheta^\star$ with the property that $\bar{V}(\vartheta^\star, x) = V(x)$ for all $x \in \mathbb{R}^n$ and Assumption 2 is implied by the statement of Proposition 1 in the nominal case. Under Assumption 2, the reward value (22) represents the maximal distance between the trajectory of η and the manifold $\mathcal{M}$. Intuitively, the optimal control law (or an approximation thereof) can be obtained by minimizing such a distance.

Lemma 1: Consider the function $\mathcal{R}_{x_0}$ in (22) and suppose that Assumption 2 holds. There exists a constant $r_\vartheta \in \mathbb{R}_{>0}$ and an open set $\mathcal{U} \subset \mathbb{R}^n$, containing the origin, such that $\mathcal{R}_{x_0}$ is bounded and continuous for all $\vartheta \in \mathbb{B}_{r_\vartheta}(\bar{\vartheta})$ and for all $x_0 \in \mathcal{U}$. ○

Proof: As a straightforward consequence of Assumption 2 and of the definition of the function $\mathcal{R}_{x_0}$ in (22), it follows that $\mathcal{R}_{x_0}(\bar{\vartheta})$ is finite for any x_0 in a neighborhood $\mathcal{U} \subset \mathbb{R}^n$ containing the origin. By definition, $\mathcal{R}_{x_0}(\vartheta)$ is nonnegative for all $\vartheta \in \mathbb{R}^\mu$. Let now $z_{\bar{\vartheta}} := (x, \eta)$ denote the solution of (10), (11) ensuing from $z_{\bar{\vartheta}}(0) = (x_0, h(\bar{\vartheta}, x_0))$, while z_ϑ describes, similarly, the solution of (10) in closed loop with $\bar{v}$ from $z_\vartheta(0) = (x_0, h(\vartheta, x_0))$. Then, by continuity of the solution of (10), (11) with respect to the parameter ϑ and to the initial condition (see, e.g., [28]), for any $\varepsilon > 0$ there exist a nonempty open set of initial conditions x_0 with the property that $\|z_{\bar{\vartheta}}(t) - z_\vartheta(t)\| < \varepsilon$ for all ϑ sufficiently close to $\bar{\vartheta}$. Hence, boundedness of $\mathcal{R}_{x_0}$ follows by continuity, in fact linearity, of the mapping $\vartheta \mapsto h(\vartheta, x)$ for fixed $x \in \mathbb{R}^n$. Continuity of the function $\mathcal{R}_{x_0}$, instead, follows immediately by relying on arguments identical to those employed above—showing that the *flow* (x, η) is a continuous function of the parameter ϑ appearing in the vector field and in the initial

optimal value function—the analysis is general in terms of the manipulation of the underlying parameters ϑ, hence, a similar strategy could be adapted and extended to several *functional approximators* and different activation functions.

conditions—and by recalling that the maximum of a continuous function is continuous. ■

As discussed in the proof of Lemma 1, the function $\mathcal{R}_{x_0}$ is such that $\mathcal{R}_{x_0}(\bar\vartheta)$ yields the minimal distance from invariance of the resulting attractive manifold, for any x_0, provided that Assumption 2 holds. Moreover, in the limiting case for μ that tends to infinity one has that $\lim_{\mu\to\infty}\mathcal{R}_{x_0}(\bar\vartheta) = \mathcal{R}_{x_0}(\lim_{\mu\to\infty}\bar\vartheta) = 0$. The aim of the following remark consists in showing that, in the case of LQ optimal control problems, the reward function $\mathcal{R}_{x_0}$ is equal to zero *if and only if* ϑ is equal to $\vartheta^\star = \mathrm{vech}(P^\star)$, where $P^\star$ denotes the unique positive solution of the underlying ARE, as discussed below. The above argument in turn ensures that in the LQ case the cost function $\mathcal{R}_{x_0}$ admits a unique global minimizer with respect to ϑ.

Remark 1: In the setting of LQ optimal control problems, the *controlled* Hamiltonian dynamics (10) is described by

$$\begin{aligned}\begin{bmatrix}\dot x\\ \dot\eta\end{bmatrix} &= \begin{bmatrix} A & -S\\ -Q & -A^\top\end{bmatrix}\begin{bmatrix}x\\ \eta\end{bmatrix} + \begin{bmatrix}0\\ I\end{bmatrix} v\\ &:= Hz + Gv \end{aligned}\qquad (24)$$

with $S := BR^{-1}B^\top \in \mathbb{R}^{n\times n}$, initialized at $(x(0),\eta(0)) = (x_0, Px_0)$, in closed loop with

$$\begin{aligned}\bar v &= \begin{bmatrix}-A^\top P + PSP - FP & A^\top - PS + F\end{bmatrix} z\\ &:= K(P)z\end{aligned}\qquad (25)$$

where $P = P^\top > 0$ denotes a generic positive definite matrix, which replaces the role of the matrix Θ in the parametrization (20), and $F := -\sigma_F I$. Therefore, the closed-loop system becomes

$$\begin{bmatrix}\dot x\\ \dot\eta\end{bmatrix} = \begin{bmatrix} A & -S\\ \Xi & F - PS\end{bmatrix}\begin{bmatrix}x\\ \eta\end{bmatrix}$$

with $\Xi = -Q - A^\top P - FP + PSP$. Provided that Assumption 1 holds, it can be shown that the reward function $\mathcal{R}_{x_0}$ is *positive definite* around $\vartheta^\star = \mathrm{vech}(P^\star)$, with $P^\star$ denoting the unique positive definite (maximal) solution of the algebraic Riccati equation (ARE)

$$0 = Q + A^\top P^\star + P^\star A - P^\star BR^{-1}B^\top P^\star\,. \qquad (26)$$

In fact, the function $\mathcal{R}_{x_0}$ can be zero for a certain ϑ if and only if $\vartheta = \mathrm{vech}(P)$ is such that the control law $\bar v$ in (25) is stabilizing for (24) and, simultaneously, the identity $\eta(t) = Px(t)$ holds for all $t \geqslant 0$ along the trajectory of the closed-loop system (24), (25) ensuing from the initial condition $(x(0),\eta(0)) = (x_0, Px_0)$. More precisely, $\vartheta = \mathrm{vech}(P)$ must be such that $\sigma(H + GK(P)) \subset \mathbb{C}^-$ and such that the subspace

$$\mathrm{im}\begin{bmatrix} I\\ P\end{bmatrix}$$

is invariant for $(H + GK(P))$. The latter requirement yields

$$\begin{bmatrix} A & -S\\ \Xi & F - PS\end{bmatrix}\begin{bmatrix} I\\ P\end{bmatrix} = \begin{bmatrix} I\\ P\end{bmatrix}\Lambda$$

for some $\Lambda \in \mathbb{R}^{n\times n}$. The first row-block equation leads to $\Lambda = A - SP$, while the second, recalling the definition of the matrix Ξ, becomes $-Q - A^\top P = P\Lambda = P(A - SP)$, hence recovering the ARE. Therefore, the only solution to the latter equation that additionally ensures that the matrix $H + GK(P)$ is Hurwitz is $P = P^\star$. ▲

The properties of the function $\mathcal{R}_{x_0}$ characterized in Lemma 1 ensure that the value $\bar\vartheta$ is (locally) attractive for a class of gradient descent methods (as discussed below in more detail). Therefore, the first objective of the following section is to provide the gradient of the reward function $\mathcal{R}_{x_0}$.

B. Gradient of the Reward Function

The objective of this section is to characterize and compute the gradient of the cost function $\mathcal{R}_{x_0}$ with respect to ϑ.

Remark 2: Since the definition of $\mathcal{R}_{x_0}$ involves the maximization of a function of time, it is expected that it possesses points of nondifferentiability. Moreover, the cost function (22) is in general a nonconvex function of ϑ, appearing in the initial conditions of the controlled Hamiltonian dynamics (10) and as a parameter that determines the underlying vector field. While nonsmooth, nonconvex optimization problems are known to be NP-hard, there are a range of methods available (including learning-based techniques) that render the (finite dimensional) problem of minimizing the cost (22) more readily solvable in practice (as demonstrated in Section VI in the context of optimal design for an automatic flight control system) than the original (infinite dimensional) optimal control problem. One such method, based on gradient information obtained on several points, is selected and discussed in the following section ▲

The solution of (22) is obtained via the controlled Hamiltonian dynamics[3] (10), with the crucial difference that while λ in (10) represents the *optimal* costate variable, η defines the approximate costate induced by the selection of the stabilizing control law $\bar v$ in (10), initialized at $(x(0),\eta(0)) = (x_0, h(\vartheta, x_0))$. Therefore, as mentioned in Section IV-A, the function $\mathcal{R}_{x_0}$ depends on the vector of parameters ϑ explicitly, as shown in (22), as well as via the initial condition and the dynamics itself. The computation of the gradient of $\mathcal{R}_{x_0}$ is detailed in the following formal statement. To provide a concise statement, let $\chi = (x,\eta) \in \mathbb{R}^{2n}$ and define

$$F(\vartheta,\chi) := \begin{bmatrix}\nabla_\eta \mathcal{H}(x,\eta)\\ -\nabla_x\mathcal{H}(x,\eta) + \bar v(x,\eta,h(\vartheta,x))\end{bmatrix}. \qquad (27)$$

Lemma 2: Let $(\zeta_1,\zeta_2,S_x,S_\eta) : \mathbb{R} \to \mathbb{R}^\mu \times \mathbb{R}^{2n} \times \mathbb{R}^{n\times\mu} \times \mathbb{R}^{n\times\mu}$ denote the trajectory of the dynamics

$$\dot\zeta = F(\vartheta,\zeta) \qquad (28a)$$

$$\begin{bmatrix}\dot S_x\\ \dot S_\eta\end{bmatrix} = \nabla_\zeta F(\vartheta,\zeta)\begin{bmatrix}S_x\\ S_\eta\end{bmatrix} + \begin{bmatrix}0\\ \dfrac{\partial\bar v}{\partial h}\dfrac{\partial h}{\partial\vartheta}\end{bmatrix} \qquad (28b)$$

[3] By a slight abuse of notation, the Hamiltonian function defined in (8) should be interpreted with the costate λ replaced by the variable η.

$\zeta = (\zeta_1, \zeta_2) \in \mathbb{R}^n \times \mathbb{R}^n$, initialized at

$$\zeta(0) = \begin{bmatrix} x_0 \\ h(\vartheta, x_0) \end{bmatrix} \tag{29a}$$

$$\begin{bmatrix} S_x(0) \\ S_\eta(0) \end{bmatrix} = \begin{bmatrix} 0 \\ \dfrac{\partial}{\partial\vartheta} h(\vartheta, x_0) \end{bmatrix}. \tag{29b}$$

Suppose that the set $\arg\max_{t \in \mathbb{R}_{\geqslant 0}} \|\eta(t) - h(\vartheta, x(t))\|^2 =: \mathcal{T}(\vartheta)$ is a singleton and let $\mathcal{T}(\vartheta) =: \hat{t}$. Then

$$\frac{\partial}{\partial\vartheta}\mathcal{R}_{x_0} = 2(\zeta_2(\hat{t}) - h(\vartheta, \zeta_1(\hat{t})))^\top \left(S_\eta(\hat{t}) - \frac{\partial}{\partial\vartheta} h(\vartheta, \zeta_1(\hat{t})) - \frac{\partial}{\partial x} h(\vartheta, \zeta_1(\hat{t})) S_x(\hat{t}) \right). \tag{30}$$

○

Proof: To begin with, by considering the definition of the cost function $\mathcal{R}_{x_0}$ in (22) and the discussion in the following paragraph, it immediately follows that

$$\frac{\partial}{\partial\vartheta}\mathcal{R}_{x_0} = 2(\eta(t) - h(\vartheta, x(t)))^\top \left(\frac{\partial \eta(t)}{\partial\vartheta} - \frac{\partial}{\partial\vartheta} h(\vartheta, x(t)) - \frac{\partial}{\partial x} h(\vartheta, x(t)) \frac{\partial x(t)}{\partial\vartheta} \right) \Bigg|_{t=\hat{t}}. \tag{31}$$

Therefore, only the *sensitivity* of the flow $\chi = (x, \eta)$ with respect to ϑ remains to be computed. Toward this end, by the standard fixed-point representation of the solution of an ordinary differential equation, one has that

$$\chi(t, \vartheta) = \chi(0, \vartheta) + \int_0^t F(\vartheta, \chi(\tau, \vartheta)) d\tau\,. \tag{32}$$

Thus, by considering the partial derivative with respect to ϑ

$$\frac{\partial}{\partial\vartheta}\chi(t, \vartheta) = \int_0^t \left[\frac{\partial F}{\partial\chi} \frac{\partial}{\partial\vartheta}\chi(\tau, \vartheta) + \frac{\partial F}{\partial\vartheta} \right] d\tau + \begin{bmatrix} 0 \\ \dfrac{\partial}{\partial\vartheta} h(\vartheta, x_0) \end{bmatrix} \tag{33}$$

which is obtained by recalling that $\chi(0, \vartheta) = (x_0, h(\vartheta, x_0))$. By the fundamental theorem of integral calculus, the time derivative of (33) then satisfies

$$\frac{d}{dt}\frac{\partial}{\partial\vartheta}\chi(t, \vartheta) = \frac{\partial F}{\partial\chi}\frac{\partial}{\partial\vartheta}\chi(t, \vartheta) + \frac{\partial F}{\partial\vartheta}. \tag{34}$$

Therefore, by letting $S_x(t) := \partial x(t)/\partial\vartheta$ and $S_\eta(t) := \partial\eta(t)/\partial\vartheta$, for all $t \in \mathbb{R}_{\geqslant 0}$ the gradient of the cost function $\mathcal{R}_{x_0}$ with respect to ϑ is provided by (30). In fact, the matrices $S_x : \mathbb{R} \to \mathbb{R}^{n \times \mu}$ and $S_\eta : \mathbb{R} \to \mathbb{R}^{n \times \mu}$ obtained as solutions of (28b), together with the initial conditions $S_x(0) = 0$ and $S_\eta(0) = h_\vartheta(\vartheta, x_0)$, satisfy the dynamics (34), while ζ satisfying (28a) replicates the time evolution of χ in (34). ■

Remark 3: In the statement of Lemma 2, it has been assumed, for simplicity of exposition that the set $\mathcal{T}$ is a singleton. Whenever the set $\mathcal{T}$ is multivalued for some ϑ, the gradient of the function $\mathcal{R}_{x_0}$ is in turn multivalued and it can be obtained by considering (31) evaluated at any $t \in \mathcal{T}(\vartheta)$. The constructions in the following section—yielding a gradient-descent algorithm that extends the *gradient-sampling* strategy to the setting of matrix manifolds—are applicable without changes also in the multivalued case. Furthermore, in practice, it may be possible to entirely circumvent the issue of non-differentiability by considering a modified, *truncated*, definition of the set $\mathcal{T}$ for which the gradient is computed, namely $\mathcal{T}^T(\vartheta) := \arg\max_{t \in [0, \bar{t}]} \|\eta(t) - h(\vartheta, x(t))\|^2$, for any $\bar{t} \in \mathbb{R}_{>0}$. In fact, since the measure induced by $\mathcal{R}_{x_0}$ is based on a forward invariance property, $\mathcal{R}_{x_0}(\vartheta)$ is equal to zero if and only if its restriction to any finite interval $[0, \bar{t}]$ is equal to zero. The numerical values selected for $\bar{t}$ and σ_F (which is related to the rate of external convergence) are, however, relevant in practice (see Section V-B for an illustrative example). ▲

C. Manifold Gradient Descent Algorithms for PI

The objective of this section consists in combining the intuitions and constructions of Sections III, IV-A, and IV-B to provide a hybrid dynamical system that yields an adaptive optimal control law via PI. To provide a concise statement of the main result of this section—which provides the construction of such a hybrid adaptive control law—a few preliminary definitions and tools are briefly recalled.

Definition 1: Given a vector $w = [w_1, \ldots, w_p] \in \mathbb{R}^p$, with $p = n(n+1)/2$, the *inverse half vectorization* operator, denoted $\mathrm{vech}^{-1}(w)$ maps w into the symmetric matrix $\frac{1}{2}(W + W^\top) \in \mathbb{R}^{n \times n}$ with

$$W = \frac{1}{2}\begin{bmatrix} w_1 & 2w_2 & 2w_3 & \ldots & 2w_{n-1} & 2w_n \\ 0 & w_{n+1} & 2w_{n+2} & \ldots & 2w_{2n-2} & 2w_{2n-1} \\ 0 & 0 & w_{2n} & \ldots & 2w_{3n-3} & 2w_{3n-2} \\ \vdots & & \ddots & \ddots & \vdots & \\ 0 & 0 & & & w_{p-2} & 2w_{p-1} \\ 0 & 0 & 0 & \ldots & 0 & w_p \end{bmatrix}. \tag{35}$$

○

It is worth observing that, by combining the constructions of Section IV-B with Definition 1, the matrix $\mathrm{vech}^{-1}(\zeta_1(\hat{t}))$ is a symmetric matrix that describes the gradient of the (scalar) cost function $\mathcal{R}_{x_0}$ with respect to the matrix Θ. Therefore, before presenting and discussing in detail the hybrid control architecture mentioned above, the gradient-descent algorithm is first revisited with respect to the manifold of symmetric and positive definite matrices, denoted by $\mathbb{S}^+(n) = \{X \in \mathbb{R}^{n \times n} : X = X^\top, X > 0\}$. This, in fact, is instrumental in updating the vector ϑ while ensuring that the matrix Θ does not leave $\mathbb{S}^+(n)$. The tangent space at a point $X \in \mathbb{S}^+(n)$ is defined as $T_X\mathbb{S}^+(n) = \{W \in \mathbb{R}^{n \times n} : W = W^\top\}$, namely by the space of symmetric matrices. The manifold $\mathbb{S}^+(n)$ becomes a Riemannian manifold by introducing the Riemannian metric as

$$\langle W_1, W_2 \rangle_X \triangleq \mathrm{tr}(X^{-1}W_1X^{-1}W_2) \tag{36}$$

for $W_1 \in T_X\mathbb{S}^+(n)$ and $W_2 \in T_X\mathbb{S}^+(n)$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. The geodesic curve at the point $X \in \mathbb{S}^+(n)$

in the direction $W \in T_X\mathbb{S}^+(n)$ is defined as

$$\gamma_{X,W}(\alpha) = X^{1/2}\exp(\alpha X^{-1/2} W X^{-1/2}) X^{1/2} \tag{37}$$

and it is entirely contained in $\mathbb{S}^+(n)$ for any $\alpha \in [0, 1]$. Finally, consider a few preliminary definitions, which are borrowed from [29].

Definition 2 (see [29, Def. 3.1]): Given a vector ϑ such that $\Theta = \text{vech}^{-1}(\vartheta)$ and a positive real number ϵ, $\partial_\epsilon \mathcal{R}_{x_0}(\vartheta)$ denotes the ϵ-subdifferential of $\mathcal{R}_{x_0}$. ○

While the interested reader is referred to [29], note that the above subdifferential is in fact computed by considering the convex hull of the gradients determined in (30) at several points *close* (according to the underlying Riemannian distance and the underlying exponential map) to $\Theta = \text{vech}^{-1}(\vartheta)$. By relying on [29], Thm. 3.12], it can then be concluded that a *descent direction* can be determined by selecting the element of minimal norm in $\partial_\epsilon \mathcal{Y}_{x_0}(\vartheta)$, for some $\epsilon > 0$, denoted as $\mathcal{D}(\vartheta)$. More precisely, consider $\partial_\epsilon \mathcal{R}_{x_0}(\vartheta)$ and let $w^\circ := \arg\min\{\|w\| : w \in \partial_\epsilon \mathcal{R}_{x_0}(\vartheta)\}$ define the element of minimal norm in the set. Then, $\mathcal{D}(\vartheta) := -w^\circ/\|w^\circ\|$ yields a descent direction with uniform decrease.

The above constructions essentially represent an extension of the *sampling gradient method* to the case of (matrix) manifolds. The following formal statement describes the convergence properties of the proposed episodic learning strategy, a practical implementation of which is then provided by Algorithm 1. In particular, in the following statement, the time histories induced by the episodic learning of Fig. 2 are interpreted in terms of trajectories of a hybrid system.

Theorem 2: Consider the hybrid system defined by the *flow dynamics* defined by (28) together with

$$\begin{aligned} \dot{\hat{\vartheta}} &= 0 \\ \dot{\tau} &= 1 \end{aligned} \tag{38}$$

the *jump dynamics* defined by the reset condition

$$\begin{aligned} \zeta^+ &= \begin{bmatrix} x_0 \\ h(\vartheta, x_0) \end{bmatrix} \\ \begin{bmatrix} S_x^+ \\ S_\eta^+ \end{bmatrix} &= \begin{bmatrix} 0 \\ \dfrac{\partial}{\partial \vartheta} h(\vartheta, x_0) \end{bmatrix} \end{aligned}$$

together with

$$\begin{aligned} \hat{\vartheta}^+ &= \text{vech}\left(\gamma_{\text{vech}^{-1}(\hat{\vartheta}),\text{vech}^{-1}\left(\mathcal{D}(\hat{\vartheta})\right)}(\alpha)\right) \\ \tau^+ &= 0 \end{aligned} \tag{39}$$

with $\alpha \in (0, 1)$, *flow set* $C := \{(\zeta, S_x, S_\eta, \hat{\vartheta}, \tau) \in \mathbb{R}^{2n} \times \mathbb{R}^{n\times\mu} \times \mathbb{R}^{n\times\mu} \times \mathbb{R}^{\mu} \times \mathbb{R} : \tau \leqslant \mathcal{T}(\hat{\vartheta})\}$ and *jump set* $D := \{(\zeta, S_x, S_\eta, \hat{\vartheta}, \tau) \in \mathbb{R}^{2n} \times \mathbb{R}^{n\times\mu} \times \mathbb{R}^{n\times\mu} \times \mathbb{R}^{\mu} \times \mathbb{R} : \tau \geqslant \mathcal{T}(\hat{\vartheta})\}$. Then, all trajectories of the hybrid system (28), (38), (29), (39) ensuing from the initial conditions $\hat{\vartheta}(0, 0)$ with the property that $\bar{v}|_{\hat{\vartheta}(0,0)}$ is stabilizing for (10) are such that $\hat{\vartheta}(t, k)$ converges to a stationary point of $\mathcal{R}_{x_0}(\vartheta)$. ○

Algorithm 1:

```
(0) Initialization. Fix $\tau_M \in \mathbb{R}_{>0}$, sufficiently large, and fix
    $\alpha \in (0, 1)$ and $\epsilon \in \mathbb{R}_{>0}$, sufficiently small. Fix $x_0 \in \mathbb{R}^n$
    and select $\hat{\vartheta} \in \mathbb{R}^\mu$ such that $\hat{u} = \bar{v}(x, \eta, h(\hat{\vartheta}, x))$
    asymptotically stabilizes the zero equilibrium of (10).
(1) Define a collection of points $\{\vartheta_1, \ldots, \vartheta_\kappa\}$ according
    to [29], Sec. 3.3] such that
    $\text{dist}(\text{vech}^{-1}(\hat{\vartheta}), \text{vech}^{-1}(\vartheta_i)) < \epsilon$, $i = 1, \ldots, \kappa$.
    For $i = 1$ to $\kappa$
(1.a) Integrate the closed-loop system (10), (21) in the
    interval $[0, \tau_M]$ from $(x(0), \eta(0)) = (x_0, h(\vartheta_i, x_0))$ and
    compute $\hat{t} = \arg\max_{t\in[0,\tau_M]} \|\eta(t) - h(\vartheta_i, x(t))\|^2$.
(1.b) Integrate system (28) in the interval $[0, \hat{t}]$ from the
    initial condition (29). Define the gradient
    $(\partial\mathcal{R}_{x_0}(\vartheta_i)/\partial\vartheta)$ as in (30).
    end
(2) Use the gradients $(\partial\mathcal{R}_{x_0}(\vartheta_i)/\partial\vartheta)$ to construct
    $\partial_\epsilon\mathcal{R}_{x_0}(\hat{\vartheta})$ and define $\mathcal{D}(\hat{\vartheta})$
(3) Update $\hat{\vartheta}$ according to
    $\hat{\vartheta} \leftarrow \text{vech}\left(\gamma_{\text{vech}^{-1}(\hat{\vartheta}),\text{vech}^{-1}\left(\mathcal{D}(\hat{\vartheta})\right)}(\alpha)\right)$
(4) Repeat from Step (1).
```

Proof: The claim is shown by noting that the sublevel sets of the function $\mathcal{R}_{x_0}$ are bounded and relying on the convergence properties of the algorithm discussed in [29], Th. 3.18]. ■

A systematic description of the practical implementation of the results expressed in Theorem 2 is provided in Algorithm 1, which yields a strategy based on episodic learning to construct approximate optimal control laws.

Note that Step (1) of Algorithm 1 aims at approximating the computation of the ϵ-subdifferential $\mathcal{D}(\vartheta)$, via the evaluation of the gradient at a finite collection of nearby points, in such a way that the gradient descent method can be practically implemented. Furthermore, by relying on the properties established in Section IV-B, (local) convergence of Algorithm 1 follows by somewhat standard arguments on gradient-descent strategies, see, e.g., [29]. The principles of the control design methodology are presented in the algorithm above by focusing on its essential features for clarity of exposition. However, a few, classical refinements of the implementation of the gradient descent approach may be straight-forwardly included. In addition to the use of a sampling gradient method mentioned above, one could, for instance, consider a *line search* subroutine that permits the computation of an optimized step α, in place of the constant one introduced in Algorithm 1.

V. Robust Open-Loop LQ Optimal Control

The main objective of this section consists in specializing the previous results to the case of linear systems and with quadratic cost functionals. As discussed in Remark 1, in such a setting the Hamiltonian dynamics is described by a system of linear equations, the costate of which is controlled by means of the additional, virtual control input v, as shown in (24).

A. Comparison Between Optimal and Controlled *Costate*

Before discussing the main results of this section, the classic characterization of the optimal costate in this scenario is briefly revisited. Toward this end, consider a linear, time-invariant, system described by the equation

$$\dot{x} = Ax + Bu \tag{40}$$

with $x : \mathbb{R} \to \mathbb{R}^n$ and $u : \mathbb{R} \to \mathbb{R}^m$ denoting the state and the control input, respectively, together with the quadratic cost functional

$$J(u) = \frac{1}{2}\int_0^\infty (\|x(\tau)\|_Q^2 + \|u(\tau)\|_R^2) d\tau\,. \tag{41}$$

It is well known that the optimal control policy is given by

$$u^\star(t) = -R^{-1}B^\top P^\star x(t) \tag{42}$$

where $P^\star$ denotes the symmetric, positive definite solution of the ARE (26). The optimal solution is equivalently obtained as

$$u^\star(t) = -R^{-1}B^\top \lambda^\star(t) \tag{43}$$

where $\lambda^\star$ denotes the solution of the Hamiltonian system

$$\begin{bmatrix} \dot{x} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} A & -S \\ -Q & -A^\top \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} := H \begin{bmatrix} x \\ \lambda \end{bmatrix} \tag{44}$$

initialized at $(x(0), \lambda(0)) = (x_0, P^\star x_0)$.

Lemma 3: Consider the LQ optimal control problem described by the dynamics (40) and the cost functional (41). Suppose that Assumption 1 holds. Then, the optimal costate is

$$\lambda^\star(t) = P^\star e^{(A-SP^\star)t} x_0 \tag{45}$$

with $S = BR^{-1}B^\top$. ○

Proof: Let $z = (x, \lambda)$ and define the change of coordinates $\hat{z} = T^{-1}z$ with

$$T^{-1} = \begin{bmatrix} I & 0 \\ -P^\star & I \end{bmatrix} \tag{46}$$

which depends on the solution of the ARE (26), and hence

$$T = \begin{bmatrix} I & 0 \\ P^\star & I \end{bmatrix}. \tag{47}$$

It is then straightforward to show that

$$T^{-1}HT = \begin{bmatrix} A_{\text{cl}} & -S \\ 0 & -A_{\text{cl}}^\top \end{bmatrix} \tag{48}$$

with $A_{\text{cl}} := A - SP^\star$. Therefore, in the transformed coordinates

$$\hat{z}_1(t) = e^{A_{\text{cl}}t}\left(\hat{z}_1(0) - \int_0^t e^{-A_{\text{cl}}\tau} S e^{-A_{\text{cl}}^\top \tau} d\tau\, \hat{z}_2(0)\right) \tag{49a}$$

$$:= e^{A_{\text{cl}}t}\left(\hat{z}_1(0) - M(t)\hat{z}_2(0)\right)$$

$$\hat{z}_2(t) = e^{-A_{\text{cl}}t}\hat{z}_2(0)\,. \tag{49b}$$

By noting that, since $z = T\hat{z}$ with T defined in (47), $\lambda(t) = \hat{z}_2(t) + P^\star \hat{z}_1(t)$ it follows that

$$\lambda(t) = P^\star e^{A_{\text{cl}}t}x_0 + \left(e^{-A_{\text{cl}}t} - P^\star e^{A_{\text{cl}}t}M(t)\right)(\lambda_0 - P^\star x_0)\,. \tag{50}$$

The proof is then concluded by recalling that $\lambda_0 = P^\star x_0$ and $\lambda(t)|_{\lambda_0 = P^\star x_0} = \lambda^\star(t) = P^\star e^{(A-SP^\star)t}x_0$. ■

By specializing the constructions discussed in the nonlinear setting to the case of linear systems, consider the *controlled Hamiltonian system*

$$\begin{bmatrix} \dot{x} \\ \dot{\eta} \end{bmatrix} = H \begin{bmatrix} x \\ \eta \end{bmatrix} + Gv \tag{51}$$

with G defined as in (24), where $v : \mathbb{R} \to \mathbb{R}^n$ in (11) reduces to the (linear) feedback

$$v^\star = (A^\top - P^\star S - \sigma_F I)(\eta - P^\star x) \tag{52}$$

since $f_1(x, V_x(x)) = (A - SP^\star)x$ and $g(x) = B$ do not depend on the state variable x.

Lemma 4: Consider the *controlled Hamiltonian system* (51) in closed loop with (52). Then, the resulting controlled costate is given by

$$\eta(t) = e^{-\sigma_F t}\xi + P^\star e^{A_{\text{cl}}t}(x_0 - M_s(t)\xi) \tag{53}$$

where $\xi = \eta_0 - P^\star x_0$ and $M_s(t) = \int_0^t e^{-A_{\text{cl}}\tau} S e^{-\sigma_F \tau} d\tau$. ○

Proof: The claim is shown by relying on arguments identical to those of the proof of Lemma 3 and by noting that $TG = G$ and that $v = (A_{\text{cl}}^\top - \sigma_F I)\hat{z}_2$ in the transformed coordinates. □

It is now possible to compare the optimal costate $\lambda^\star$ and that computed by considering the stabilized Hamiltonian dynamics (51). To provide a concise statement, let $c_1 \in \mathbb{R}_{>0}$ and $c_2 \in \mathbb{R}_{>0}$ be such that $\|e^{-A_{\text{cl}}t}\| < c_1 e^{c_2 t}$.

Proposition 2: Consider the Hamiltonian dynamics (44) and the closed-loop Hamiltonian dynamics (51), (52). Fix any $t^\star \in \mathbb{R}_{>0}$, $\varepsilon \in \mathbb{R}_{>0}$ and $\mu \in \mathbb{R}_{>0}$ and let

$$\sigma_F > \max\left\{\frac{2c_1\|S\|}{\varepsilon} + c_2, \frac{1}{t^\star}\log\left(\frac{\varepsilon}{2\mu}\right)\right\}.$$

Define $e(t) = \lambda^\star(t) - \eta(t)$. Then

$$\|e(t)\| \leqslant \varepsilon \tag{54}$$

for all $t > t^\star$ and for all $(x_0, \eta_0) \in \mathbb{R}^n \times \mathbb{R}^n$ such that $\|(x_0, \eta_0 - P^\star x_0)\| < \mu$. ○

Proof: By the definitions of the functions $\lambda^\star$ and η given in (45) and (53), respectively, it follows that $e(t) = M_s(t)\xi - e^{-\sigma_F t}\xi$. Consider first the norm of the matrix-valued function M_s, namely $t \mapsto \int_0^t e^{-A_{\text{cl}}\tau} S e^{-\sigma_F \tau} d\tau$. In particular

$$\|M_s(t)\| = \left\|\int_0^t e^{-A_{\text{cl}}\tau} S e^{-\sigma_F \tau} d\tau\right\| \leqslant \int_0^t \|e^{-A_{\text{cl}}\tau} S e^{-\sigma_F \tau}\| d\tau$$

$$\leqslant \|S\| \int_0^t \|e^{-A_{\text{cl}}\tau}\|\, \|e^{-\sigma_F \tau}\| d\tau$$

$$\leqslant c_1\|S\| \int_0^t e^{(c_2 - \sigma_F)\tau} d\tau = \frac{c_1\|S\|}{c_2 - \sigma_F}\left[e^{(c_2-\sigma_F)\tau}\right]_0^t$$

$$\leqslant \frac{c_1\|S\|}{\sigma_F - c_2}\,.$$

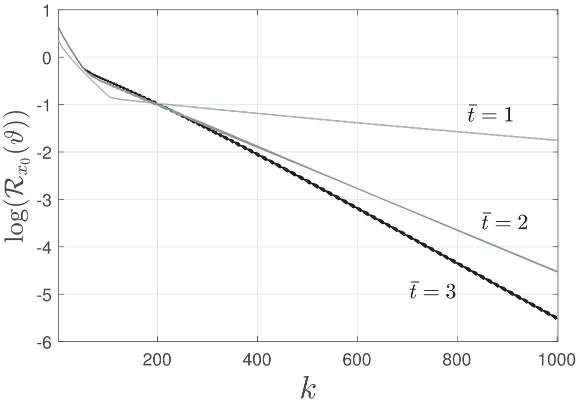

Fig. 4. Logarithm of the cost function $\mathcal{R}_{x_0}(\vartheta)$ during gradient descent iterations for different values of $\bar{t}$ in Remark 3 and with fixed step size $\alpha = 0.001$.

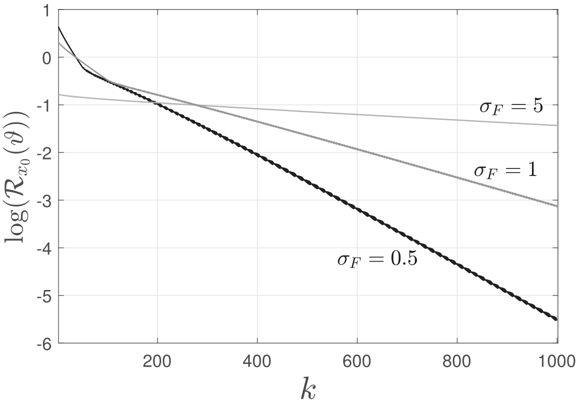

Fig. 5. Logarithm of the cost function $\mathcal{R}_{x_0}(\vartheta)$ during gradient descent iterations for different values of σ_F and with fixed step size $\alpha = 0.001$.

Therefore

$$\|e(t)\| \leqslant \|M_s(t)\|\|\xi\| + \|e^{-\sigma_F t}\|\|\xi\| \leqslant \frac{c_1\|S\|}{\sigma_F - c_2}\mu + e^{-\sigma_F t}\mu$$
$$\leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \leqslant \varepsilon$$

for all $t > t^\star$, where the third inequality follows from the selection of σ_F. ■

B. Stabilized Computation of the Optimal Costate

By building on ideas inspired by the constructions of Section IV, the main objective of this section consists in specializing the cost function (22) and the subsequent gradient descent algorithm to the case of LQ optimal control problems. Toward this end, note first that in the latter case, the candidate approximate value function $\bar{V}$ in (20) can be *a priori* effectively limited to the class of quadratic functions, namely by letting $\mathfrak{c}(x) = x \in \mathbb{R}^n$. As a consequence, the matrix $\Theta \in \mathbb{R}^{n\times n}$, with n describing the dimension of the state of system (40), represents a generic matrix $P \in \mathbb{R}^{n\times n}$ in the quadratic form

$$\bar{V}_\ell(\vartheta, x) = \frac{1}{2}x^\top P x \tag{55}$$

with $\vartheta = \mathrm{vech}(P)$, where $P = P^\top > 0$, therefore, describes a candidate solution of the ARE (26). It is worth observing that this selection is such that Assumption 2 holds with $P = P^\star$. Such a matrix P must be then determined by specializing the gradient-descent algorithm on matrix manifold discussed above to the cost function

$$\mathcal{L}_{x_0}(\vartheta) := \|\eta(\hat{t}) - Px(\hat{t})\|^2 \tag{56}$$

where (x, η) are obtained from (24), (25), and with $\hat{t} := \arg\max_{t\in\mathbb{R}_{\geqslant 0}} \|\eta(t) - Px(t)\|^2$. As commented upon in the paragraph following the definition of $\mathcal{R}_{x_0}$ in (22), the dependence of $\mathcal{L}_{x_0}$ on P is threefold: direct as immediately suggested by (56) as well as via the parameterization of the vector-field of the underlying Hamiltonian dynamics, as shown in (24) and (25), and via the initial conditions $(x(0), \eta(0)) = (x_0, Px_0)$. Interestingly, since in the LQ setting the *flow* of the Hamiltonian dynamics can be easily obtained in closed-form as a matrix exponential, such a threefold dependence can be explicitly expressed as

$$\mathcal{L}_{x_0}(\vartheta) = \left\| \begin{bmatrix} -P & I \end{bmatrix} e^{(H+GK(P))\hat{t}} \begin{bmatrix} I \\ P \end{bmatrix} x_0 \right\|^2 \tag{57}$$

with $K(P)$ defined in (25).

Remark 4: The construction in Algorithm 1, together with the discussion in Remark 1 for the LQ setting, ensures that the gradient-descent method introduced herein is such that the estimate P converges to $P^\star$. Therefore, the strategy can be interpreted as an alternative to the solution of the underlying ARE (26), which does not involve the solution of any algebraic equation, the computation of eigenvalues of any square matrix or the inversion of any non-singular matrix. ▲

Remark 5: Consider the first-order approximation of the matrix exponential function $t \mapsto e^{(H+GK(P))t}$ around the origin, evaluated at $t = \hat{t}$, namely

$$e^{(H+GK(P))\hat{t}} = I + \hat{t}(H + GK(P)) + o(\hat{t}^2)\,. \tag{58}$$

By replacing the above-mentioned approximation into the cost function (57), straightforward computations show that the cost function $\mathcal{L}_{x_0}$ can be approximated by

$$\mathcal{L}_{x_0}(\vartheta) = \hat{t}^2\|Yx_0\|^2 + o(\hat{t}^3) \tag{59}$$

where $Y := Q + PA + A^\top P - PSP$ coincides with the right-hand side of the ARE (26). ▲

To illustrate the properties mentioned in Remark 3, consider (linear) double integrator dynamics together with a cost functional as in (41) with $Q = I$ and $R = 1$. The initial condition is $x_0 = [5,\ 3]^\top$ and let Algorithm 1 be initialized with a random P and fixed step size $\alpha = 0.001$. Figs. 4 and 5 show the dependence of the convergence rate on the selection of $\bar{t}$ and σ_F, respectively: from the computational perspective, it is desirable to employ a longer time interval and a smaller value for σ_F, such that the trajectories are driven *sufficiently away* from the original

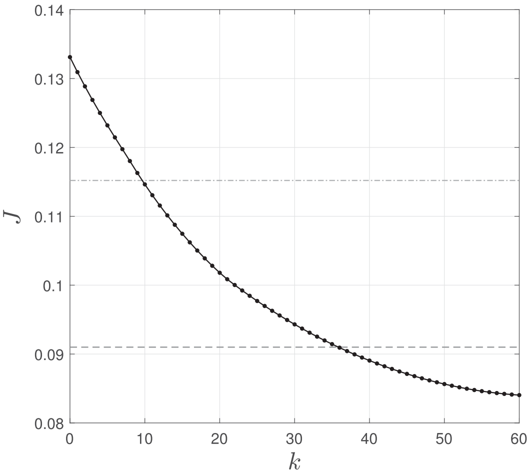

Fig. 6. Costs of the linearized control law u_ℓ (dash-dotted line) and of the nonlinear control law u_{ps} (dashed line), together with the evolution, over the iteration number k, of the cost attained by the control law $\bar{v}$ in (21) defined by considering $\bar{V}(\vartheta_k, x)$.

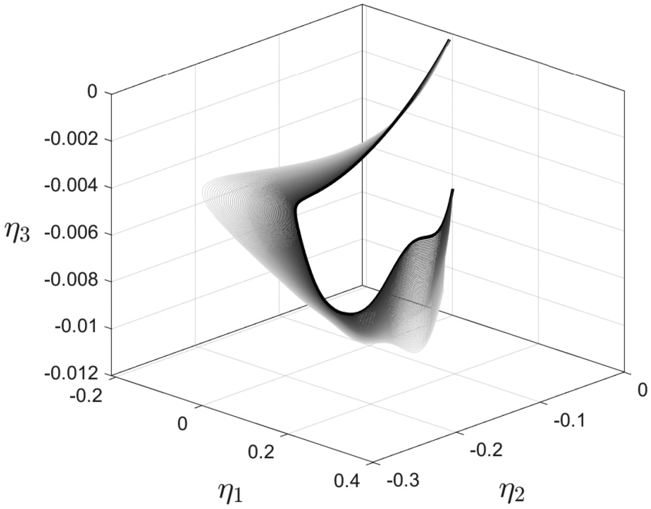

Fig. 7. Phase-plot in $\mathbb{R}^3$ of the costate variable induced by the controlled Hamiltonian dynamics (10) with $\bar{v}(x, \eta, h(\vartheta_k, x))$. The solid black line describes the phase-plot associated to the last iteration of the gradient descent algorithm.

subspace. The latter feature, in fact, yields a gradient matrix that is more reliable from the numerical point of view.

VI. Automatic Flight Control of the Longitudinal Motion of an Aircraft

The results of the previous sections are validated and corroborated by means of numerical simulations, concerning the optimal design of an automatic flight control system for the longitudinal motion of an aircraft. The model description is borrowed from [30], in which a detailed derivation of the dynamics is provided. The longitudinal motion of an aircraft is therefore captured by dynamics as in (1), described by the vector fields f and g defined in (60) (at the bottom of the page), where $x_1(t) \in \mathbb{R}$ denotes the angle of attack, in radians, while $x_2(t) \in \mathbb{R}$ and $x_3(t)$ are the angular displacement with respect to the horizon and its rate of change, respectively. The control action $u(t) \in \mathbb{R}$ describes the prescribed tail deflection angle. In [30], two control strategies are suggested: a linear static feedback—optimally designed on the basis of the *linearized* model—is subsequently compared with a nonlinear feedback obtained by a power series approximate solution of the underlying HJB pde. The latter policy is computed at the price of computationally intensive derivations that require the solutions of nonlinear algebraic equations. The effectiveness of the two strategies above is therein measured by means of a quadratic cost functional similar to (2) with $q(x) = (1/2)x^\top Qx$, where $Q = (1/2)I$ and $R = 2$. More precisely, the control law solving the linearized problem is

$$u_\ell(x) = -0.053x_1 + 0.5x_2 + 0.521x_3 \tag{61}$$

whereas the nonlinear feedback computed by approximating the HJB pde by including the cubic terms of its power series expansion is

$$\begin{aligned} u_{ps}(x) = u_\ell(x) &+ 0.04x_1^2 - 0.048x_1x_2 \\ &+ 0.374x_1^3 - 0.312x_1^2x_2 \,. \end{aligned} \tag{62}$$

Note that the coefficients of the control law (61) are related to the entries of the symmetric and positive definite matrix $P^\star$ that solves the corresponding ARE (26), i.e.,

$$P^\star = \begin{bmatrix} 0.3218 & -0.1777 & -0.0083 \\ -0.1777 & 0.7183 & 0.0495 \\ -0.0083 & 0.0495 & 0.0498 \end{bmatrix}. \tag{63}$$

The constructions discussed in the previous sections are then carried out in the setting of the automatic flight control problem described by (60) by selecting $m = 2$ in the definition of $\mathfrak{c}(x)$, yielding

$$\mathfrak{c}(x) = \left[x_1 \; x_2 \; x_3 \; x_1x_2 \; x_1x_3 \; x_2x_3 \; x_1^2 \; x_2^2 \; x_3^2 \right]^\top. \tag{64}$$

The gradient-descent method on the manifold of symmetric and positive definite matrix illustrated in Theorem 2 is then applied by initially letting the matrix $\Theta \in \mathbb{R}^{9\times 9}$ be defined as

$$\Theta_0 := \begin{bmatrix} P^\star & \mathbf{0}_{3\times 6} \\ \mathbf{0}_{6\times 3} & \varrho I_6 \end{bmatrix} \tag{65}$$

$$f(x) := \begin{bmatrix} -0.877x_1 + x_3 + 0.47x_1^2 - 0.088x_1x_3 - 0.019x_2^2 + 3.846x_1^3 - x_1^2x_3 \\ x_3 \\ -4.208x_1 - 0.396x_3 - 0.47x_1^2 - 3.564x_1^3 \end{bmatrix}, \quad g(x) := \begin{bmatrix} -0.215 \\ 0 \\ -20.967 \end{bmatrix} \tag{60}$$

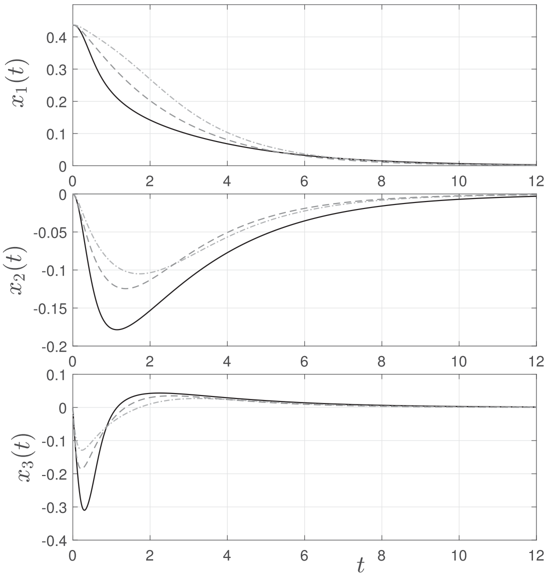

Fig. 8. Time histories of the state x induced by the value function $\bar{V}(\vartheta_{60}, x)$ (solid black lines) compared with those induced by the control law u_ℓ (dash-dotted lines) and u_{ps} (dashed line).

where ϱ is a positive parameter such that $\Theta_0 \in \mathbb{S}^+(9)$. In the following numerical simulations, the coefficient ϱ has been selected as $\varrho = 0.001$, while the parameter σ_F in the construction of the adaptive control law (21) is equal to 1. It is assumed that the initial configuration of the aircraft is such that the angle of attack corresponds to 25 degrees, hence, $x_0 = [\,(25/180)\pi\ 0\ 0]^\top$.

Fig. 6 depicts the costs yielded by the linearized solution u_ℓ in (61), i.e., $J(u_\ell) = 0.1152$, and by the nonlinear control law u_{ps} proposed in [30] and reported in (62), i.e., $J(u_{ps}) = 0.091$, dash-dotted and dashed lines, respectively, together with the evolution over the iteration number k of the cost yielded by the iterative control law (21), defined by the approximate value function $\bar{V}(\vartheta_k, x) = \mathfrak{c}(x)^\top \Theta_k \mathfrak{c}(x)$. It can be observed that even a relatively small number of iterations of the algorithm allows, essentially without solving any partial differential or even algebraic equations, to achieve a cost lower than that obtained in [30] with u_{ps}. Note in addition that the latter control law, i.e., u_{ps}, has been constructed at the price of significant computational complexity and effort in solving algebraic equations, as commented upon in [30]. The graphs in Fig. 7 instead report the evolution over the iteration number (i.e., one curve for each value of k) of the phase-plot in $\mathbb{R}^3$ of the corresponding costate variable induced by the controlled Hamiltonian dynamics (10) with $\bar{v}(x, \eta, h(\vartheta_k, x))$. The solid black line describes the phase-plot associated to the last iteration of the gradient descent algorithm.

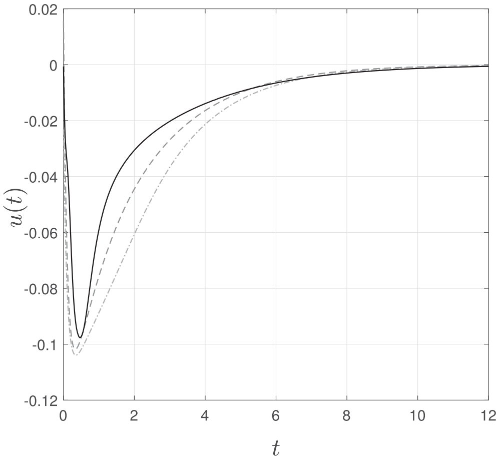

Fig. 9. Time histories of the control law u induced by the value function $\bar{V}(\vartheta_{60}, x)$ (solid black lines) compared with those induced by the control law u_ℓ (dash-dotted lines) and u_{ps} (dashed line).

Finally, the value of the matrix Θ_{60}, reported in (66) (at the bottom of the page), is then employed to compute the time histories of the state x and of the control action $\bar{v}$ induced by the value function $\bar{V}(\vartheta_{60}, x)$. These are indicated by the solid black lines in Figs. 8 and 9, respectively. The time histories are also compared with those resulting from the use of the control law u_ℓ (dash-dotted lines) and of u_{ps} (dashed line).

VII. Conclusion

An iterative learning strategy to construct optimal control laws in infinite-horizon dynamic optimization problems has been proposed and discussed. The strategy relies on a modified (controlled) version of the classic Hamiltonian dynamics arising in this context, in which the stable invariant submanifold—instrumental for the construction of the optimal policy—is rendered externally asymptotically stable. This is then exploited, in the presence of a VFA for the candidate solution, to yield an alternative temporal difference error in the episodic learning architecture that measures the distance from invariance induced by the current approximating value function. The estimate is then updated by means of a gradient descent algorithm in the manifold of symmetric positive definite matrices. The theory has

$$\Theta_{60} = \begin{bmatrix} 0.3340 & -0.1691 & -0.0032 & 0.0077 & -0.0011 & -0.0041 & -0.0068 & -0.0026 & 0.0020 \\ -0.1691 & 0.7389 & 0.0212 & -0.0041 & -0.0046 & 0.0045 & -0.0012 & -0.0019 & 0.0037 \\ -0.0032 & 0.0212 & 0.0538 & -0.0007 & -0.0014 & 0.0007 & 0.0001 & -0.0011 & 0.0013 \\ 0.0077 & -0.0041 & -0.0007 & 0.0200 & 0.0020 & -0.0145 & -0.0117 & -0.0011 & 0.0009 \\ -0.0011 & -0.0046 & -0.0014 & 0.0020 & 0.0021 & -0.0020 & -0.0004 & 0.0009 & -0.0016 \\ -0.0041 & 0.0045 & 0.0007 & -0.0145 & -0.0020 & 0.0108 & 0.0081 & 0.0007 & -0.0005 \\ -0.0068 & -0.0012 & 0.0001 & -0.0117 & -0.0004 & 0.0081 & 0.0074 & 0.0008 & -0.0008 \\ -0.0026 & -0.0019 & -0.0011 & -0.0011 & 0.0009 & 0.0007 & 0.0008 & 0.0009 & -0.0016 \\ 0.0020 & 0.0037 & 0.0013 & 0.0009 & -0.0016 & -0.0005 & -0.0008 & -0.0016 & 0.0027 \end{bmatrix}. \qquad (66)$$

been illustrated by means of a practical application involving an automatic flight control problem.

Appendix

Lemma 5: Consider the systems described by the equations

$$\dot{x} = f(x) \tag{67a}$$

$$\dot{z} = f(z) + s(z)e^{-\sigma t} \tag{67b}$$

with $x(0) = z(0) = x_0$, $s(0) = S \in \mathbb{R}^n$. Suppose that $f : \mathbb{R}^n \to \mathbb{R}^n$ and $s : \mathbb{R}^n \to \mathbb{R}^n$ are smooth mappings and that the origin is an LES equilibrium point for (67a). Then, for any $\varepsilon \in \mathbb{R}_{>0}$, there exist a neighborhood $\mathcal{U} \subset \mathbb{R}^n$ containing the origin and $\sigma \in \mathbb{R}_{>0}$ such that $\|z(t) - x(t)\| \leqslant \varepsilon$ for all $t \geqslant 0$ and all $x_0 \in \mathcal{U}$. $\circ$

Proof: To begin with, let $\chi := z - x$, hence

$$\begin{aligned}\dot{\chi} &= f(z) - f(x) + s(z)e^{-\sigma t}\\ &= f(\chi + x) - f(x) + s(\chi + x)e^{-\sigma t}\\ &:= g_1(\chi, x) + s(\chi + x)e^{-\sigma t}\\ &:= g_1(\chi, 0) + g_2(\chi, x) + s(\chi + x)e^{-\sigma t}\end{aligned} \tag{68}$$

with $g_1(\chi, x) := f(\chi + x) - f(x)$ and $g_2(\chi, x) = g_1(\chi, x) - g_1(\chi, 0)$. Therefore, it is straightforward to observe that the claim can be concluded by equivalently assessing the stability properties of the origin for the (extended) error dynamics

$$\begin{aligned}\dot{x} &= f(x)\\ \dot{\chi} &= g_1(\chi, 0) + g_2(\chi, x) + s(\chi + x)\xi\\ \dot{\xi} &= -\sigma\xi\,.\end{aligned} \tag{69}$$

By LES of the origin of (67a), there exist $r_o \in \mathbb{R}_{>0}$, $d \in \mathbb{R}_{>0}$, and $\lambda \in \mathbb{R}_{>0}$ such that $\|x(t)\| \leqslant de^{-\lambda t}\|x_0\|$ for any $t \geqslant 0$ and for all $x_0 \in D_0 := \{x \in \mathbb{R}^n : \|x\| < r_0\}$. On the other hand, the inequality $\|\xi(t)\| \leqslant e^{-\sigma t}\|\xi_0\|$ for any $t \geqslant 0$ immediately follows by the definition of the dynamics for $\xi(t) \in \mathbb{R}$. Furthermore, again by LES of the origin of (67a) and by observing that $g_1(\chi, 0) = f(\chi) - f(0) = f(\chi)$, one has that $\chi = 0$ is a locally asymptotically stable equilibrium point for the dynamics $\dot{\chi} = g_1(\chi, 0)$. Hence, by [28], Th. 4.14], there exist $r_1 \in \mathbb{R}_{>0}, c_i \in \mathbb{R}_{>0}, i = 1, \ldots, 4$, a neighborhood $D_1 := \{\chi \in \mathbb{R}^n : \|\chi\| < r_1\}$ of the origin and a function $W : D_1 \to \mathbb{R}_{>0}$ with the property that

$$\begin{gathered}c_1\|\chi\|^2 \leqslant W(\chi) \leqslant c_2\|\chi\|^2\\ \nabla W(\chi)^\top g_1(\chi, 0) \leqslant -c_3\|\chi\|^2\\ \|\nabla W(\chi)\| \leqslant c_4\|\chi\|\end{gathered} \tag{70}$$

for all $\chi \in D_1$. Moreover, by smoothness of the involved functions, there exist constants $k_i \in \mathbb{R}_{>0}$, $i = 1, 2$, such that $\|g_2(\chi, x)\| \leqslant k_1\|x\|$ and $\|s(\chi + x)\| \leqslant k_2$ for all $(x, \chi) \in D_0 \times D_1$. The time derivative of the function W along the trajectories of system (68) then yields

$$\begin{aligned}\dot{W} &= \nabla W(\chi)^\top (g_1(\chi, 0) + g_2(\chi, x) + s(\chi + x)\xi)\\ &\leqslant -c_3\|\chi\|^2 + c_4k_1\|\chi\|\|x\| + c_4k_2\|\chi\|\|\xi\|\\ &\leqslant -c_3\|\chi\|^2 + \frac{(c_4k_1)}{2\,m}\|\chi\|^2 + \frac{m}{2}\|x\|^2 + \frac{(c_4k_2)}{2\,m}\|\chi\|^2\\ &\quad + \frac{m}{2}\|\xi\|^2\end{aligned} \tag{71}$$

for any $m \in \mathbb{R}_{>0}$, where the second inequality is obtained by relying on Young's inequality. Therefore, by selecting m with the property that $c_3 - c_4(k_1 + k_2)/2\,m > 0$, it follows that

$$\begin{aligned}\dot{W}(t) &\leqslant \frac{m}{2}d^2e^{-2\lambda t}\|x_0\|^2 + \frac{m}{2}\|\xi_0\|^2e^{-2\sigma t}\\ &:= \delta_1 e^{-2\lambda t}\|x_0\|^2 + \delta_2 e^{-2\sigma t}.\end{aligned} \tag{72}$$

Since the claim is obtained provided $\|\chi(t)\| < \varepsilon$ for all $t \geqslant 0$ and by recalling that, by (70), $\|\chi\|^2 \leqslant W(\chi)/c_1$, it remains to show that $W(t) < \mu := \min\{\varepsilon^2, r_1^2\}c_1$ for all $t \geqslant 0$, where the definition of μ ensures that $\chi(t) \in D_1$ for all time. To this end, by integration and by recalling that $W(0) = W(\chi(0)) = 0$ since by assumption $x(0) = z(0)$, one has that

$$\begin{aligned}W(t) &\leqslant \delta_1\|x_0\|^2 \int_0^t e^{-2\lambda\tau}d\tau + \delta_2 \int_0^t e^{-2\sigma\tau}d\tau\\ &\leqslant \frac{\delta_1}{2\lambda}\|x_0\|^2(1 - e^{-2\lambda t}) + \frac{\delta_2}{2\sigma}(1 - e^{-2\sigma t})\\ &\leqslant \frac{\delta_1}{2\lambda}\|x_0\|^2 + \frac{\delta_2}{2\sigma}\end{aligned} \tag{73}$$

Therefore, the claim is obtained by selecting $\|x_0\|^2 < \min\{r_0^2, \mu\lambda/\delta_1\}$ and $\sigma > \delta_2/\mu$, which imply that each term of the second inequality of (73) is smaller than $\mu/2$.

References

[1] D. Liberzon, *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton, NJ, USA: Princeton Univ. Press, 2011.

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

[3] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*. New York, NY, USA: Wiley, 1972.

[4] D. A. Carlson, A. B. Haurie, and A. Leizarowitz, *Infinite Horizon Optimal Control: Deterministic and Stochastic Systems*. Berlin, Germany: Springer, 2012.

[5] R. Vinter, *Optimal Control*. Berlin, Germany: Springer, 2010.

[6] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[7] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA, USA: Athena Scientific, 2005.

[8] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*. Hoboken, NJ, USA: Wiley, 1962.

[9] D. L. Lukes, "Optimal regulation of nonlinear dynamical systems," *SIAM J. Control*, vol. 7, no. 1, pp. 75–100, 1969.

[10] A. Wernli and G. Cook, "Successive control for the nonlinear quadratic regulator problem," *Automatica*, vol. 11, pp. 75–84, 1975.

[11] M. Sassano and A. Astolfi, "Dynamic approximate solutions of the HJ inequality and of the HJB equation for input-affine nonlinear systems," *IEEE Trans. Autom. Control*, vol. 57, no. 10, pp. 2490–2503, Oct. 2012.

[12] M. Sassano, T. Mylvaganam, and A. Astolfi, "An algebraic approach to dynamic optimisation of nonlinear systems: A survey and some new results," *J. Control Decis.*, vol. 6, no. 1, pp. 1–29, 2019.

[13] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[14] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.

[15] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 32, no. 2, pp. 140–153, May 2002.

[16] F. L. Lewis and D. Vrabie, “Reinforcement learning and adaptive dynamic programming for feedback control,” *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Jul.–Sep. 2009.

[17] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, “Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof,” *IEEE Trans. Syst., Man, Cybern., Part B (Cybern.)*, vol. 38, no. 4, pp. 943–949, Aug. 2008.

[18] K. G. Vamvoudakis and F. L. Lewis, “Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem,” *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[19] M. Abu-Khalaf and F. L. Lewis, “Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach,” *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[20] K. G. Vamvoudakis, “Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach,” *Syst. Control Lett.*, vol. 100, pp. 14–20, 2017.

[21] D. Kleinman, “On an iterative technique for Riccati equation computations,” *IEEE Trans. Autom. Control*, vol. AC-13, no. 1, pp. 114–115, Feb. 1968.

[22] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, “Adaptive optimal control for continuous-time linear systems based on policy iteration,” *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[23] K. G. Vamvoudakis, D. Vrabie, and F. L. Lewis, “Online adaptive algorithm for optimal control with integral reinforcement learning,” *Int. J. Robust Nonlinear Control*, vol. 24, no. 17, pp. 2686–2710, 2014.

[24] A. J. van der Schaft, *L2-Gain and Passivity Techniques in Nonlinear Control*, 2nd ed. Berlin, Germany: Springer, 2000.

[25] A. J. Van Schaft, “L_2-gain analysis of nonlinear systems and nonlinear state-feedback H_∞ control,” *IEEE Trans. Autom. Control*, vol. 37, no. 6, pp. 770–784, Jun. 1992.

[26] C. F. Van Loan, “The ubiquitous Kronecker product,” *J. Comput. Appl. Math.*, vol. 123, no. 1/2, pp. 85–100, 2000.

[27] K. Hornik, M. Stinchcombe, and H. White, “Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks,” *Neural Netw.*, vol. 3, no. 5, pp. 551–560, 1990.

[28] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.

[29] P. Grohs and S. Hosseini, “ε-subgradient algorithms for locally Lipschitz functions on Riemannian manifolds,” *Adv. Comput. Math.*, vol. 42, no. 2, pp. 333–360, 2016.

[30] W. L. Garrard and J. M. Jordan, “Design of nonlinear automatic flight control systems,” *Automatica*, vol. 13, no. 5, pp. 497–505, 1977.

Mario Sassano (Senior Member, IEEE) was born in Rome, Italy, in 1985. He received the B.S. degree in automation systems engineering and the M.S. degree in systems and control engineering from the University of Rome “La Sapienza,” Rome, Italy, in 2006 and 2008, respectively, and the Ph.D. degree in control theory from Imperial College London, London, U.K.

He was a Research Assistant with the Department of Electrical and Electronic Engineering, Imperial College London (2009–2012). He is currently an Assistant Professor with the University of Rome “Tor Vergata,” Rome, Italy. His research interests are focused on nonlinear observer design, optimal control and differential game theory with applications to mechatronical systems and output regulation for hybrid systems.

Dr. Sassano is member of the IFAC Technical Committee on Control Design. He is currently an Associate Editor for the *European Journal of Control* and of the IEEE CSS and EUCA Conference Editorial Boards.

Thulasi Mylvaganam (Senior Member, IEEE) was born in Bergen, Norway, in 1988. She received the M.Eng. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2010 and the Ph.D. degree in control theory from the Department of Electrical and Electronic Engineering, Imperial College London, in 2014.

She was a Research Associate with Imperial College London from 2014 to 2016. In 2016–2017, she was a Research Fellow with the Department of Aeronautics, Imperial College London, where she is currently a Senior Lecturer (Associate Professor). Her research interests include nonlinear control design, optimal control, differential game theory, and distributed control.

Dr. Mylvaganam is currently an Associate Editor for the IEEE CSS Conference Editorial Board and member of the IFAC Technical Committee on Optimal Control.

Alessandro Astolfi (Fellow, IEEE) was born in Rome, Italy, in 1967. He graduated in electrical engineering from the University of Rome in 1991. In 1992 he joined ETH-Zurich where he obtained a M.Sc. in Information Theory in 1995 and the Ph.D. degree with Medal of Honor in 1995 with a thesis on discontinuous stabilization of nonholonomic systems. In 1996 he was awarded a Ph.D. from the University of Rome “La Sapienza,” for his work on nonlinear robust control.

Since 1996, he has been with the Electrical and Electronic Engineering Department, Imperial College London, London, U.K., where he is currently a Professor of Nonlinear Control Theory and the Head of the Control and Power Group. From 1998 to 2003, he was also an Associate Professor with the Department of Electronics and Information, Politecnico of Milano, Milan, Italy. Since 2005, he has also been a Professor with Dipartimento di Ingegneria Civile e Ingegneria Informatica, University of Rome Tor Vergata, Rome, Italy. He was a Visiting Lecturer in “Nonlinear Control” in several universities, including ETH-Zurich (1995–1996), Terza University of Rome (1996), Rice University, Houston (1999), Kepler University, Linz (2000), SUPELEC, Paris (2001), and Northeastern University (2013). His research interests are focused on mathematical control theory and control applications, with special emphasis for the problems of discontinuous stabilization, robust and adaptive control, observer design and model reduction. He is the author of more than 150 journal papers, 30 book chapters, and more than 240 papers in refereed conference proceedings. He is the author (with D. Karagiannis and R. Ortega) of the monograph *Nonlinear and Adaptive Control With Applications* (Springer-Verlag, 2007).

Dr. Astolfi is the recipient of the IEEE CSS A. Ruberti Young Researcher Prize (2007), the IEEE RAS Googol Best New Application Paper Award (2009), the IEEE CSS George S. Axelby Outstanding Paper Award (2012), and the Automatica Best Paper Award (2017). He was the recipient of the Medal of Honor for his first Ph.D. degree. He is currently a Distinguished Member of the IEEE CSS, IEEE Fellow, and IFAC Fellow. He was an Associate Editor for *Automatica*, *Systems and Control Letters*, IEEE TRANSACTIONS ON AUTOMATIC CONTROL, the *International Journal of Control*, the *European Journal of Control*, and the *Journal of the Franklin Institute*; an Area Editor for the *International Journal of Adaptive Control and Signal Processing*; a Senior Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL; and an Editor-in-Chief for the *European Journal of Control*. He is currently Editor-in-Chief of IEEE TRANSACTIONS ON AUTOMATIC CONTROL. He was the Chair of the IEEE CSS Conference Editorial Board (2010–2017) and in the IPC of several international conferences. He has been a Member of the IEEE Fellow Committee (2016), (2019–2022).

Open Access provided by ‘Università degli Studi di Roma “Tor Vergata”’ within the CRUI CARE Agreement