Linear-like Policy Iteration based Optimal Control for Continuous-time Nonlinear Systems

Adnan Tahirovic[1] and Alessandro Astolfi[2]

***Abstract*— We propose a novel strategy to construct optimal controllers for continuous-time nonlinear systems by means of linear-like techniques, provided that the optimal value function is differentiable and *quadratic-like*. This assumption covers a wide range of cases and holds locally around an equilibrium under mild assumptions. The proposed strategy does not require solving the Hamilton-Jacobi-Bellman equation, that is a nonlinear partial differential equation, which is known to be hard or impossible to solve. Instead, the Hamilton-Jacobi-Bellman equation is replaced with an easy-solvable state-dependent Lyapunov matrix equation. We exploit a linear-like factorization of the underlying nonlinear system and a policy-iteration algorithm to yield a linear-like policy-iteration for nonlinear systems. The proposed control strategy solves optimal nonlinear control problems in an asymptotically exact, yet still linear-like manner. We prove optimality of the resulting solution and illustrate the results via four examples.**

I. INTRODUCTION

The solution of optimal control problems for nonlinear systems hinges upon the solution of the Hamilton-Jacobi-Bellman (HJB) partial differential equations (PDE), which can be extremely difficult or impossible to solve. Many approximation methods for solving the HJB PDE have been developed, under a variety of assumptions, at the cost of some optimality loss [1]. An alternative way for solving optimal control problems for nonlinear systems is based on Pontryagin maximum principle, which provides necessary conditions of optimality. Direct discretization is another approach for solving optimal control problems; it is often used for problems over finite horizon and to handle constraints. The resulting problem can be efficiently solved due to the existence of fast and reliable nonlinear programming solvers, which make this the most widely used and popular approach. However, the HJB equation gives both necessary and sufficient conditions for an optimal feedback control solution and provides the optimal value function over the entire state space [1]. This makes a solution based on the HJB approach unique. For this reason, in this paper we study unconstrained optimal control problems and their solutions via the HJB equation.

A first class of techniques used to solve HJB equations is based on the theory of viscosity solutions [2]. This solution is proved to be the value function of the underlying optimal control problem. It is required to be continuous, but not necessarily differentiable, as it is assumed for classical solutions. To obtain an approximate viscosity solution, finite-difference and finite-element methods have been used: both require a discretization of the state space, hence the computational cost increases exponentially with the dimension of the state space.

A second class of techniques is based on the principle of model-based reinforcement learning with policy-iteration (PI) algorithm, which reduces a nonlinear HJB PDE to a linear PDE [3], [4]. This is used to find the cost associated to an admissible control. The PI algorithm also provides an incremental improvement of the control policy and ensures convergence to the optimal control. In many cases, solving a linear PDE is still not easy. In [5], Galerkin approximations have been used to approximately solve optimal control problems by combining this approximation with the PI algorithm. Some other approaches developed to approximate the solution of the HJB PDE, up to a desired degree of accuracy, have been presented in [12]–[14].

A third class of techniques is based on results obtained for linear systems and for a cost in quadratic form. For such systems the HJB PDE reduces to an algebraic Riccati equation (ARE), which is easy to solve. The methods based on Jacobian linearization of the nonlinear system, feedback linearization [6], [15], dynamic extensions [16], and state-dependent Riccati equations (SDRE) [17]–[21], provide techniques to approximate the optimal control by avoiding solving nonlinear PDEs. The linearization-based approach is feasible only in the vicinity of an equilibrium, while feedback-linearization may cancel "useful" nonlinearities and may not provide a near-to-optimal control law. The dynamic extension-based approach relies on a modified cost to avoid solving the HJB PDE, providing thus a suboptimal control law. It is worth noting that the dynamic extension-based control is capable to extract an upper bound of the modified cost to provide a measure of the sub-optimality level of the solution. The SDRE-based control approach relies upon a *linear-like factorization of the nonlinear system*. Its main disadvantage is the lack of stability guarantee.

This paper provides a thorough theoretical extension of our previous work [22]. We propose a control strategy for input-affine continuous-time nonlinear systems which is based on the PI paradigm combined with the linear-like factorization used in the SDRE approach. We use the PI algorithm to ensure convergence of the policy to the optimal control. Unlike

*This work has been partially supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 739551 (KIOS CoE); by the Italian Ministry for Research in the framework of the 2017 and of the 2020 Program for Research Projects of National Interest (PRIN), Grant No. 2017YKXYXJ and 2020RTWES4; and by the EPSRC grant "Model Reduction from Data", Grant No. EP/W005557.

[1] Adnan Tahirovic is with Faculty of Electrical Engineering, University of Sarajevo, atahirovic@etf.unsa.ba

[2] Alessandro Astolfi is with Imperial College London, U.K., and also with Dipartimento di Informatica, Sistemi e Produzione, Universita di Roma, Italy. a.astolfi@imperial.ac.uk

other PI approaches, we use a linear-like factorization of the nonlinear system to avoid solving any PDE, thus replacing the PDE with a state-dependent Lyapunov matrix equation (SDLE). In this way the proposed control strategy solves the optimal nonlinear control problem in an asymptotically exact, but still linear-like, manner, provided the optimal cost has a quadratic-like form. If this is not the case, the obtained results suggest that the proposed approach has a potential to find an optimal solution in the vicinity of an equilibrium.

The paper is organized as follows. In Section II we define the problem and recall a general form of the PI algorithm. In Section III we recall the SDRE approach with its associated factorization technique and redefine the optimal control problem. In Section IV, we define the linear-like PI which computes the optimal control with a modified cost. Section V introduces the modified linear-like PI to solve the considered nonlinear optimal control problem. Section VI provides an illustration of the results via four examples, while Section VII concludes the paper.

II. Control based on policy iteration for continuous-time systems

A. Problem description

Consider a class of continuous-time nonlinear systems described by an equation of the form

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

with state $x(t) \in \mathbb{R}^n$, input $u(t) \in \mathbb{R}^m$ and f and g Lipschitz continuous on a compact set $\tilde{\Omega} \subset \mathbb{R}^n$ that contains the origin. Suppose in addition that the system (1) has an equilibrium at the origin for $u = 0$, that is $f(0) = 0$. Finally, assume that the system is controllable in $\tilde{\Omega}$, that is, it is possible to find an input signal u which steers the state of the system to the origin from any initial condition x_0 in $\tilde{\Omega}$ in some time $\bar{t} \geq 0$.

Consider now the cost function

$$V(x_0, u) = \int_0^{\infty} (l(x) + \|u\|_R^2) dt, \tag{2}$$

where the state penalty function l is a positive function on $\tilde{\Omega}$, such that $l(0) = 0$. Assume that the system (1) with output $y = l(x)$ is zero-state observable, and $R \in \mathbb{R}^{mxm}$ is a symmetric positive definite matrix. Typically, $l(x)$ is quadratic, that is $l(x) = x^T Q x$, where $Q = Q^T$ is a positive semidefinite matrix.

A feedback control $u = u(x)$ is called an *admissible control*, $u \in \mathscr{A}(\Omega)$, with respect to l on Ω, if u is continuous on Ω, $u(0) = 0$, the zero equilibrium of the closed-loop system is locally asymptotically stable with basin of attraction $\Omega \subseteq \tilde{\Omega}$, and the cost (2) is finite for all $x_0 \in \Omega$. The minimal value of the cost function V, obtained for an admissible control $u^* = u^*(x)$ (the optimal control), is denoted as the optimal cost $V^*(x)$, $\forall x \in \Omega$. This optimal cost V^*, called the value function, is the solution of the HJB equation

$$\frac{\partial V^*(x)}{\partial x} f(x) - \frac{1}{4} \frac{\partial V^*(x)}{\partial x} g(x) R^{-1} g(x)^T \frac{\partial V^*(x)}{\partial x}^T + l(x) = 0, \tag{3}$$

provided it is differentiable. Equation (3) is in general hard to solve even in those cases in which a (unique) solution is known to exist. The requirement to solve a PDE makes the optimal control problem virtually impossible to solve in closed-form. If a solution exists, the optimal control is

$$u^* = u^*(x) = -\frac{1}{2} R^{-1} g^T(x) \frac{\partial V^*(x)}{\partial x}^T. \tag{4}$$

B. Policy iteration for nonlinear systems

To compute the value of the cost for a given initial condition x_0 and an admissible control $\hat{u}$, one has to solve (1) with $u = \hat{u}$, which is not always possible, and compute the integral (2) along the corresponding solution. Another way to deal with this problem is to differentiate (2) along the trajectories of the system yielding the linear PDE

$$\frac{\partial \hat{V}(x)}{\partial x} (f(x) + g(x)\hat{u}(x)) + l(x) + \|\hat{u}\|_R^2 = 0, \tag{5}$$

which represents an incremental expression of the cost of the admissible control $\hat{u}$, and it does not depend on the trajectories of the system (1). If the optimal control (4) is used, *i.e.* $\hat{u} = u^*$, then (5) transforms into the nonlinear PDE (3), the solution of which directly provides the optimal cost V^* and the optimal control law u^*. For more detail, see, e.g. [4] and [5].

The optimal PI for continuous-time nonlinear systems has been proposed in [4]. The main idea of this iterative algorithm is to choose an arbitrarily initial admissible control $\hat{u} = \hat{u}(x) \in \mathscr{A}(\Omega)$ and solve the linear PDE (5) for $\hat{V}$, which should be easier than solving the nonlinear PDE (3). In order to improve the performance of the arbitrarily selected control $\hat{u}$, one then defines the policy-update

$$\hat{u}^*(x) = \arg\min_u \frac{\partial \hat{V}(x)}{\partial x} (f(x) + g(x)\hat{u}(x)) + l(x) + \|\hat{u}\|_R^2 = -\frac{1}{2} R^{-1} g^T(x) \frac{\partial \hat{V}(x)}{\partial x}^T, \forall x \in \Omega. \tag{6}$$

Having a new and improved control $\hat{u}^*$ (see e.g. [4] and [5]), one can again solve (5) to obtain the value function $\hat{V}$. By iteratively updating the value function and the control law iterating (5) and (6), the optimal PI algorithm ensures, in principle, the desired convergence, i.e. $\lim_{k\to\infty} \hat{V}_k(x) = V^*(x)$ and $\lim_{k\to\infty} \hat{u}_k(x) = u^*(x)$, $\forall x \in \Omega$, where k is the index of the iteration.

Choosing an arbitrarily initial admissible control in an analytical form as a first step of the policy iteration algorithm can be difficult for some nonlinear systems. However, different techniques for constructing such a control law for classes of nonlinear systems can be found, e.g., in [6]–[11].

Although equation (5) should be easier to solve for $\hat{V}$ than solving (1) and (2), it is still a PDE. For this reason different approaches to approximately deal with equation (5) have been proposed, see, e.g. [4], [5]. The goal of this paper is to show how PI can be exploited to find the optimal control solution without the need to solve any PDE on the basis of a simple linear-like procedure.

C. Policy iteration for linear systems

In this section we consider linear systems, that is the system (1) with $f(x) = Ax$, with $A \in \mathbb{R}^{nxn}$, $g(x) = B$, with $B \in \mathbb{R}^{nxm}$ and a quadratic cost, that is $l(x) = x^T Qx$, with $Q = Q^T \geq 0$, in (2). Assume that the pair (A,B) is stabilizable and the pair $(Q^{1/2},A)$ is detectable.

Assuming that the optimal value function is of the form

$$V^*(x) = x^T P^* x, \tag{7}$$

where $P^* = P^{*T}$ is a positive definite matrix, the HJB equation (3) becomes the ARE

$$A^T P^* + P^* A - P^* BR^{-1} B^T P^* + Q = 0, \tag{8}$$

which is easily solvable and has a unique positive definite solution P^*. The optimal control action can then be computed from (4) yielding

$$u^*(x) = -R^{-1} B^T P^* x = \Pi^* x, \tag{9}$$

where Π^* is the optimal control policy.

Although the solution to the optimal control problem for continuous-time linear systems can be given in the closed-form (9), we recall the optimal PI algorithm to understand how to construct the optimal control in an iterative manner.

In the simplified version of the optimal PI algorithm for linear systems the cost-update equation (5) becomes the Lyapunov Matrix Equation (LME)

$$(A + B\hat{\Pi})^T \hat{P} + \hat{P}(A + B\hat{\Pi}) + Q + \hat{\Pi}^T R\hat{\Pi} = 0, \tag{10}$$

which can be easily solved for a positive definite matrix $\hat{P}$, provided an admissible control $\hat{u} = \hat{\Pi}x$ is given. Additionally, the policy-update equation (6) for linear systems becomes

$$\hat{u}^* = \hat{\Pi}^* x = -R^{-1} B^T \hat{P} x, \quad \hat{\Pi}^* = -R^{-1} B^T \hat{P}. \tag{11}$$

The proof of convergence of the PI for the linear case is provided in [23], where it has also been shown that the PI is actually Kleiman-Newton's method, which ensures convergence to the solution of the ARE whenever the initial control is admissible.

III. Pointwise factorization of the optimal control problem

Under mild regularity assumptions the nonlinear system (1) can be rewritten in the form

$$\dot{x} = A(x)x + g(x)u, \tag{12}$$

where $A : \mathbb{R}^n \to \mathbb{R}^{nxn}$ is a smooth matrix valued function. The main idea behind the factorizations of the function f as $f(x) = A(x)x$ is to represent the nonlinear system (1) as a pointwise linear system by assuming that A and g are constant matrices for each state x along the trajectories of the system, see e.g. [19].

In the spirit of the above factorization, similarly to the linear case, we assume a pointwise quadratic form for the optimal value function, namely

$$V^*(x) = x^T P^*(x) x, \tag{13}$$

where $P^*(x) = [P^*(x)]^T$ for all $x \in \Omega$ is a state-dependent matrix valued function and it is positive definite for all $x \in \Omega$.

For clarity, we first define the solution to the SDRE [19], which represents the factorized version of the ARE (8).

Definition 1 [SDRE] *A positive definite matrix $\bar{P}$ is the pointwise solution to the SDRE for the state x if*

$$A(x)^T \bar{P} + \bar{P}A(x) - \bar{P}g(x)R^{-1} g(x)^T \bar{P} + Q = 0. \tag{14}$$

As in the case of the ARE, the SDRE is easily solvable for each fixed $x \in \Omega$. By mimicking the linear-like procedure presented in II-C, the control action can be computed in the pointwise form

$$u^*(x) = -R^{-1} g^T(x) \bar{P}(x) x = \bar{\Pi}(x) x. \tag{15}$$

Equations (14) and (15) form the SDRE-based control method: (14) is solved for each x along the trajectories of the system and the control law is computed as in (15).

Note that the SDRE-based control does not provide the optimal solution to the optimal control problem for the nonlinear system, since (14) has not been derived from the HJB equation (3). Another issue pertains to the matrix $\bar{P}$, for which we do not have a closed form solution, that is $\bar{P} = \bar{P}(x)$, but only the pointwise value for each state x along the trajectories of the system. This prevents $V(x) = x^T \bar{P}(x) x$ from being a Lyapunov function candidate, since its time derivative along the trajectories of the system, namely

$$\dot{V}^*(x) = \dot{x}^T \bar{P}(x) x + x^T \bar{P}(x) \dot{x} + x^T \dot{\bar{P}}(x) x, \tag{16}$$

has the additional term $\dot{\bar{P}}(x)$, which is impossible to obtain analytically and to be used for further analysis. To address this issue consider the following statement.

Lemma 1 [Direct optimal control] *Assume that the optimal value function for the optimal control problem for the nonlinear system (12) is given in the quadratic-like form (13), where $P^*(x) = [P^*(x)]^T$ is a positive definite matrix for all $x \in \Omega$. Then $P^*(x)$ is the solution of the HJB equation*

$$\begin{aligned} &x^T \{A(x)^T P^* + P^* A(x) - P^* g(x) R^{-1} g(x)^T P^* + Q\} x \\ &+ u_{corr}^T R u_{corr} + x^T \dot{P}^* x = 0, \end{aligned} \tag{17}$$

while the optimal control is given by $u^* = \bar{u} + u_{corr}$, where

$$\bar{u} = -R^{-1} g^T(x) P^*(x) x = \bar{\Pi} x, \tag{18}$$

$$u_{corr} = -\frac{1}{2} R^{-1} [\sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j g^T(x) \frac{\partial p_{i,j}}{\partial x}], \tag{19}$$

and $p_{i,j}$ indicates the $(i,j)^{th}$ element of the matrix $P^*(x)$.

Proof: Starting from (5), one obtains (dropping arguments)

$$\begin{aligned} &(Ax + gu^*)^T P^* x + x^T P^* (A + gu^*) + x^T Qx \\ &+ u^{*T} R u^* + x^T \dot{P}^* x = 0, \end{aligned} \tag{20}$$

where the optimal control u^* is obtained as the control that minimizes the left-hand-side of (20), giving the two components (18) and (19). Note that the last term of (20) is the time derivative along the trajectories of the system once u^* is used, that is $x^T \dot{P}^* x = x^T \dot{P}^*|_{A(x)+g(x)u^*} x$, which gives the second term (19) of the minimizing control u^*.

If we replace u^* in the left-hand-side of (20) with $\bar{u} + u_{corr}$, we obtain

$$\begin{aligned}&(Ax+g\bar{u})^T P^* x + x^T P^*(A+g\bar{u}) + (\bar{u}+u_{corr})^T R(\bar{u}+u_{corr})\\&+x^T Qx + (gu_{corr})^T P^* + P^* g u_{corr} + x^T \dot{P}^* x = 0,\end{aligned} \tag{21}$$

which gives

$$\begin{aligned}&x^T\left[(A+g\bar{\Pi})^T P^* + P^*(A+g\bar{\Pi}) + Q + \bar{\Pi}^T R\bar{\Pi}\right] x\\&+u_{corr}^T R u_{corr} + x^T \dot{P}^* x = 0.\end{aligned} \tag{22}$$

By replacing the control policy $\bar{\Pi}$ in accordance with (18), one gets (17) which completes the proof. ■

Although Lemma 1 provides the exact solution to the optimal control problem, the HJB equation (17), which is itself a PDE, is as hard to solve for P^* as the initial HJB equation (3). However, equation (17) allows for a separation of the optimal control problem into two simpler problems, one aimed at finding the solution $\bar{u}$, which is the counterpart of (14), and the second one aimed at finding a correction term from the last two terms in (17), which are discussed in Sections IV and V, respectively.

IV. An approximate control based on linear-like policy iteration

A. The State-dependent Lyapunov Equation - SDLE

The main idea behind the linear-like PI is to use the PI algorithm for nonlinear systems by avoiding using PDEs, *i.e.* by using only Lyapunov matrix equations as in the linear case discussed in Section II-C. To do so, we conduct the PI by omitting the last two terms in (17) to obtain a Lyapunov equation instead of the PDE at the cost of optimality loss. For clarity, we define the State-dependent Lyapunov Equation (SDLE) which is used as the approximate cost-update equation in the PI algorithm.

Definition 2 [Approximate cost-update] *Consider the admissible control $\hat{u} = \hat{\Pi}(x)x \in \mathscr{A}(\Omega)$. A differentiable function $\hat{V} = x^T \hat{P}(x) x : \Omega \to \mathbb{R}$ ($\hat{V}(0) = 0$), where $\hat{P}(x)$ is a positive definite matrix for all $x \in \Omega$, is the approximate cost function of $\hat{u}$ if $\hat{P}(x)$ satisfies the SDLE*

$$(A(x)+g(x)\hat{\Pi})^T \hat{P} + \hat{P}(A(x)+g(x)\hat{\Pi}) + Q + \hat{\Pi}^T R\hat{\Pi} = 0. \tag{23}$$

We call (23) the approximate cost-update equation for the nonlinear system and write $\hat{P}(x) = CU_{SDLE}(\hat{\Pi}(x))$, where the index SDLE indicates that one has to solve the state-dependent Lyapunov matrix equation (23) to obtain $\hat{P}(x)$.

Note first that this equation is easy solvable as in the linear case (10). Moreover, unlike the idea behind the SDRE (14), where P is computed pointwise for each single x along the trajectories of the system, the SDLE provides an analytical form of $\hat{P}$. Having $\hat{P}$ in closed form, it is then possible to compute the time derivative $\dot{\hat{P}}$ along the trajectories of the system, thus circumventing one of the main limitations of the SDRE-based approach.

Note also that the SDLE can be derived from (5) as in (20)-(22), by letting $\hat{u} = \bar{u} + u_{corr}$, where the terms equal to the last two terms in (17) are omitted for simplicity. This would mean that the SDLE can be considered as the cost-update equation when taking $u_{corr}(x) = 0$, for all x, and by omitting the time derivative $\dot{P}$. For this reason we call (23) *the approximate cost-update equation*, and we write $\hat{P}(x) = CU_{SDLE}(\hat{\Pi})$.

B. The control based on the SDLE

Along with *Definition* 2, we introduce a new definition and two results to define the control based on the SDLE.

Definition 3 [Approximate policy-update] *Consider the differentiable function $\hat{V}(x) = x^T \hat{P}(x) x : \Omega \to \mathbb{R}$ ($\hat{V}(0) = 0$), in which for each x, $\hat{P}(x)$ is a positive definite matrix obtained from (23). The control $\hat{u}^*$ is said to update the control $\hat{u}$ (or the policy $\hat{\Pi}^*$ updates the policy $\hat{\Pi}$) in accordance with the approximate policy-update equations for nonlinear systems*

$$\hat{u}^* = -R^{-1} g(x)^T \hat{P}(x) x, \quad \hat{\Pi}^* = -R^{-1} g(x)^T \hat{P}(x), \tag{24}$$

and we write $\hat{\Pi}^* = PU_{SDLE}(\hat{P}(x))$.

Note that (24) includes only the first term (18) of the optimal control given by (18)-(19). For this reason, we also call $\hat{\Pi}^* = PU_{SDLE}(\hat{P}(x))$ *the approximate policy-update equation*.

Lemma 2 [Stabilizability of the approximate policy-update] *Consider an admissible control $\hat{u}_k(x) = \hat{\Pi}_k x \in \mathscr{A}(\Omega)$ and the positive definite solution $\hat{P}_k$ obtained from (23) in accordance to Def. 2. Then the updated control $\hat{u}_{k+1} = \hat{\Pi}_{k+1}(x) x = -R^{-1} g(x)^T \hat{P}_k(x) x$ is admissible on Ω as well.*

Proof: To prove the statement we need to consider two different state space regions, $\mathscr{R}_1 \in \mathbb{R}^n$ and $\mathscr{R}_2 \in \mathbb{R}^n$, in which $x^T \dot{\hat{P}}_k x \geq 0$ and $x^T \dot{\hat{P}}_k x < 0$, respectively, and the time derivative $\dot{\hat{P}}$ is obtained along the trajectories of the system (1) in closed-loop with $\hat{u}_{k+1}$.

Let V_k be a candidate control Lyapunov function which is positive-definite, that is

$$V_k(x) = \begin{cases} \hat{V}_k(x) = x^T \hat{P}_k x & \text{if } x \in \mathscr{R}_2,\ x^T \dot{\hat{P}}_k x < 0, \\ \\ \hat{V}_k^m(x) = \hat{V}_k + \int_0^p x^T \dot{\hat{P}}_k x dt & \text{if } x \in \mathscr{R}_1,\ x^T \dot{\hat{P}}_k x \geq 0, \end{cases} \tag{25}$$

where p is an arbitrary positive constant. This function is continuous by definition and differentiable for all $x \in \Omega$, including the states x along the switching hypersurface $x^T \dot{\hat{P}}_k x = 0$. Namely, $\lim \dot{\hat{V}}_k^m = \lim \dot{\hat{V}}_k$ also holds in the limiting case when $x^T \dot{\hat{P}}_k x \to 0_{\pm}$.

In the region $\mathscr{R}_1$, the time derivative of V_k along the trajectories of the system (1) becomes

$$\dot{V}_k = \dot{\hat{V}}_k^m = x^T [(A + g\hat{\Pi}_{k+1})^T \hat{P}_k + \hat{P}_k (A + g\hat{\Pi}_{k+1})] x, \tag{26}$$

which can be rewritten in the form

$$\begin{aligned}\dot{\hat{V}}_k^m &= x^T[(A+g\hat{\Pi}_k)^T\hat{P}_k+\hat{P}_k(A+g\hat{\Pi}_k)]x\\ &\quad +x^T[(g\hat{\Pi}_{k+1})^T\hat{P}_k+\hat{P}_k g\hat{\Pi}_{k+1}-(g\hat{\Pi}_k)^T\hat{P}_k-\hat{P}_k g\hat{\Pi}_k]x.\end{aligned}$$

Since the pair $(\hat{\Pi}_k,\hat{P}_k)$ satisfies (23) and $g^T\hat{P}_k=-R\hat{\Pi}_{k+1}$, we have

$$\begin{aligned}\dot{\hat{V}}_k^m &= -x^T[Q+\hat{\Pi}_k^T R\hat{\Pi}_k]x-\\ &\quad x^T[\hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1}+\hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1}-\hat{\Pi}_k^T R\hat{\Pi}_{k+1}-\hat{\Pi}_{k+1}^T R\hat{\Pi}_k]x,\end{aligned}$$

hence

$$\dot{\hat{V}}_k^m=-x^T[Q+\hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1}+(\hat{\Pi}_{k+1}-\hat{\Pi}_k)^T R(\hat{\Pi}_{k+1}-\hat{\Pi}_k)]x. \tag{27}$$

As a result, $\dot{\hat{V}}_k^m$ is negative-definite in the region $\mathscr{R}_1$.

In the region $\mathscr{R}_2$, the time derivative of V_k along the trajectories of the system (1) is

$$\dot{V}_k=\dot{\hat{V}}_k=x^T[(A+g\hat{\Pi}_{k+1})^T\hat{P}_k+\hat{P}_k(A+g\hat{\Pi}_{k+1})]x+x^T\dot{\hat{P}}_k x,$$

which can be written in the form

$$\begin{aligned}\dot{\hat{V}}_k &= -x^T[Q+\hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1}+(\hat{\Pi}_{k+1}-\hat{\Pi}_k)^T R(\hat{\Pi}_{k+1}-\hat{\Pi}_k)]x\\ &\quad +x^T\dot{\hat{P}}_k x.\end{aligned} \tag{28}$$

This means that $\dot{\hat{V}}_k$ is negative-definite in $\mathscr{R}_1$.

This proves that V_k is a Lyapunov function, the origin is asymptotically stable and the control law $\hat{u}_k$ is admissible.■

Lemma 3 [Cost of the approximate policy-update] *Consider an admissible control $\hat{u}_k=\hat{\Pi}_k x\in\mathscr{A}(\Omega)$ and its corresponding positive definite solution $\hat{P}_k$ obtained from (23) in accordance to Def. 2. Then the cost of $\hat{u}_k$ is $V_k^m=\hat{V}_k+\int_0^\infty x^T\dot{\hat{P}}_k x dt$, where $\hat{V}_k=x^T\hat{P}_k x$.*

Proof: By assumption, we use (23) to construct $\hat{P}_k$ as

$$(A(x)+g(x)\hat{\Pi}_k)^T\hat{P}_k+\hat{P}_k(A(x)+g(x)\hat{\Pi}_k)+Q+\hat{\Pi}_k^T R\hat{\Pi}_k=0, \tag{29}$$

which can be modified into the form

$$\frac{\partial\hat{V}_k(x)}{\partial x}(f(x)+g(x)\hat{u}_k(x))-x^T\dot{\hat{P}}_k x+x^T Qx+\hat{u}(x)_k^T R\hat{u}_k(x)=0. \tag{30}$$

For all $x\in\mathscr{R}_2$, it is seen from (30) that $\hat{V}_k$ can be considered the cost for $\hat{u}_k$ which is related to the modified positive semidefinite state-cost matrix $Q^m=Q-\dot{\hat{P}}_k\geq 0$, that is

$$\hat{V}_k=\int_0^\infty(x^T(Q-\dot{\hat{P}}_k)x+\hat{u}_k^T R\hat{u}_k)dt. \tag{31}$$

It follows from (31) that

$$\hat{V}_k+\int_0^\infty x^T\dot{\hat{P}}_k x dt=\int_0^\infty(x^T Qx+\hat{u}_k^T R\hat{u}_k)dt=V_k^m\geq 0, \tag{32}$$

which is the claim of Lemma 3 for all $x\in\mathscr{R}_2$. For all $x\in\mathscr{R}_1$, it is not possible to conduct the same analysis as for $x\in\mathscr{R}_2$, since Q^m might be negative definite for some $x\in\mathscr{R}_1$.

However, for all $x\in\mathscr{R}_1$, $x^T\dot{\hat{P}}_k x$ and $\frac{\partial\hat{V}_k}{\partial x}(f+g\hat{u}_k)$ in (30) can form a new term, $\frac{\partial V_k^m}{\partial x}(f+g\hat{u}_k)$, where $V_k^m=\hat{V}_k+\int_0^\infty x^T\dot{\hat{P}}_k x$. This modification gives (30) in the form of (5), that is

$$\frac{\partial V_k^m(x)}{\partial x}(f(x)+g(x)\hat{u}_k(x))+x^T Qx+\hat{u}(x)_k^T R\hat{u}_k(x)=0, \tag{33}$$

which is the incremental expression of the cost of the admissible control. In addition, since V_k^m is positive definite in $\mathscr{R}_1$, it represents the cost of $\hat{u}_k$, which completes the proof of Lemma 3. ■

It should be noted that it is possible to conduct the same analysis for all $x\in\mathscr{R}_2$ as for $x\in\mathscr{R}_1$ since, due to (32), it is now possible to assume that V_k^m is a positive definite solution of (33) for every $x\in\mathscr{R}_2$, which is required to show, that it is the cost of the control $\hat{u}_k$.

C. The approximate linear-like PI based on the SDLE

The approximate linear-like PI based on the SDLE iteratively uses the approximate cost-update (Def. 2) and the approximate policy-update (Def. 3) in order to construct the final form of the approximate control. The following result states that such a procedure is convergent.

Theorem 1 [Convergence of the approximate linear-like PI] *Consider the $(k-1)^{th}$ iteration of the approximate linear-like PI based on the SDLE procedure, that is the pair $(\hat{P}_{k-1},\ \hat{\Pi}_k)$. Assume the control is obtained in accordance to Def. 3, that is $\hat{u}_k=\hat{\Pi}_k x=PU_{SDLE}(\hat{P}_{k-1})$, while $\hat{P}_{k-1}$ is the positive definite solution of (23) in accordance to Def. (2), that is $\hat{P}_{k-1}=CU_{SDLE}(\hat{\Pi}_{k-1})$, and the initial control $\hat{u}_1$ is admissible. If the pair $(\hat{P}_k,\ \hat{\Pi}_{k+1})$ is constructed at the k^{th} iteration step, then the approximate linear-like PI based on the SDLE procedure converges.*

Proof: In accordance to the approximate cost-update (23), $\hat{P}_k$ is the unique and positive definite solution of

$$x^T[(A+g\hat{\Pi}_k)^T\hat{P}_k+\hat{P}_k(A+g\hat{\Pi}_k)+Q+\hat{\Pi}_k^T R\hat{\Pi}_k]x=0. \tag{34}$$

If $\hat{P}_k(\hat{u}_k)$ related to the control $\hat{u}_k$ is considered as an update of $\hat{P}_{k-1}(\hat{u}_{k-1})$, then it can be replaced in (34) by

$$\hat{P}_k(\hat{u}_k)=\hat{P}_{k-1}(\hat{u}_{k-1})+\Delta\hat{P}_k(\hat{u}_k), \tag{35}$$

where $\Delta\hat{P}_k(\hat{u}_k)$ is by definition the variation of the matrix $\hat{P}_{k-1}(\hat{u}_{k-1})$ once the new control $\hat{u}_k$ is used. This further means that the form of the variation of the cost (Lemma 3) for the control $\hat{u}_k$ is given as

$$\Delta V_k^m(\hat{u}_k)=x^T\Delta\hat{P}_k(\hat{u}_k)x+\int_0^\infty x^T\Delta\dot{\hat{P}}_k(\hat{u}_k)x dt, \tag{36}$$

where

$$V_k^m(\hat{u}_k)=V_{k-1}^m(\hat{u}_{k-1})+\Delta V_k^m(\hat{u}_k). \tag{37}$$

Combining (34) and (35) leads to

$$\begin{aligned}&x^T[(A+g\hat{\Pi}_k)^T\hat{P}_{k-1}+\hat{P}_{k-1}(A+g\hat{\Pi}_k)+Q+\hat{\Pi}_k^T R\hat{\Pi}_k]x+\\ &x^T[(A+g\hat{\Pi}_k)^T\Delta\hat{P}_k+\Delta\hat{P}_k(A+g\hat{\Pi}_k)]x=0.\end{aligned} \tag{38}$$

In case of a linear system, the first term would represent the Hamiltonian function. However, for a nonlinear system we call this term a linear-like Hamiltonian function for the pair $(\hat{P}_{k-1},\ \hat{\Pi}_k)$, that is $\mathscr{H}_{k-1}$. The second term in (38) represents the time derivative of the cost variation (36) when the control $\hat{u}_k$ is used, that is

$$\mathscr{H}_{k-1} + \frac{d}{dt}\Delta V_k^m(\hat{u}_k) = 0. \tag{39}$$

Starting again from (34), we write

$$\begin{aligned} &x^T[(A+g\hat{\Pi}_{k+1})^T\hat{P}_k + \hat{P}_k(A+g\hat{\Pi}_{k+1}) + Q + \hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1}]x \\ &+ x^T[(g\hat{\Pi}_k - g\hat{\Pi}_{k+1})^T\hat{P}_k + \hat{P}_k(g\hat{\Pi}_k - g\hat{\Pi}_{k+1}) \\ &+ \hat{\Pi}_k^T R\hat{\Pi}_k - \hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1}]x = 0. \end{aligned}$$

Using now $g^T\hat{P}_k = -R\hat{\Pi}_{k+1}$, which follows from the approximate policy-update $\hat{\Pi}_{k+1} = -Rg^T\hat{P}_k$, we obtain

$$\begin{aligned} &\mathscr{H}_k + x^T[-\hat{\Pi}_k^T R\hat{\Pi}_{k+1} + \hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1} - \hat{\Pi}_{k+1}^T R\hat{\Pi}_k \\ &+ \hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1} + \hat{\Pi}_k^T R\hat{\Pi}_k - \hat{\Pi}_{k+1}^T R\hat{\Pi}_{k+1}]x = 0, \end{aligned}$$

that is

$$\mathscr{H}_k = -(\hat{\Pi}_{k+1} - \hat{\Pi}_k)^T R(\hat{\Pi}_{k+1} - \hat{\Pi}_k) \le 0, \tag{40}$$

which by back-tracking becomes

$$\mathscr{H}_{k-1} = -(\hat{\Pi}_k - \hat{\Pi}_{k-1})^T R(\hat{\Pi}_k - \hat{\Pi}_{k-1}) \le 0. \tag{41}$$

Combining now (41) with (39), yields

$$\frac{d}{dt}\Delta V_k^m(\hat{u}_k) \ge 0, \tag{42}$$

which can be integrated over $[0,\infty)$ to get

$$\lim_{t\to\infty}\Delta V_k^m(\hat{u}_k) - \Delta V_k^m(\hat{u}_k) \ge 0. \tag{43}$$

Due to Lemma 2 the controls $\hat{u}_k$ and $\hat{u}_{k-1}$ are admissible, hence the state of the system is zero as $t \to \infty$ for both controls. This means that there is no variation in the cost at the origin between any two admissible controls, that is $\lim_{t\to\infty}\Delta V_k^m(u_k) = 0$. From (43) and (37), it now holds that

$$V_k^m \le V_{k-1}^m. \tag{44}$$

As a result, the sequence $\{V_k^m\}_{k=1}^n$ is monotonically decreasing, while being bounded from below by zero due to Lemma 3, that is $V_k^m \ge 0$, hence it is convergent when $n \to \infty$.■

We call the solution based on this approach the PI-SDLE control. One of the main advantages of the proposed PI-SDLE control is that the linear-like PI can also be computed pointwise using (23), instead of finding a closed form solution. In such a case, one needs to conduct the whole PI algorithm for every single x along the trajectories of the system. Such a procedure is similar to the pointwise computation of the ARE solution when the SDRE-based control is used. Unlike the SDRE-based control, the PI-SDLE based control is proven to be stabilizing in Ω provided the initial control is admissible.

V. Optimal control based on linear-like policy iteration

We now show how to use the linear-like PI proposed in *Theorem* 1 to obtain the optimal solution of optimal control problems for continuous-time nonlinear systems.

Algorithm 1 PLAIN-PI($A(x)$, $g(x)$, Q, R, $\hat{u}_0^1 = \hat{\Pi}_0^1(x)x \in \mathscr{A}(\Omega)$)

```
for $k \leftarrow 0$ to $N^1$ do                                   ▷ loop with $N^1$ steps
    $\hat{P}_k^1 \leftarrow$ CuSdle($A(x)$, $g(x)$, $Q$, $R$, $\hat{\Pi}_k^1$)      ▷ eq. (45)
    $\hat{\Pi}_{k+1}^1 \leftarrow -R^{-1}g^T(x)\hat{P}_k^1(x)x$                      ▷ eq. (46)
end for
return $\bar{u}^1 \leftarrow \hat{\Pi}_{k+1}^1 x$; $\bar{P}^1 \leftarrow \hat{P}_k^1$
```

Definition 4 [The plain-PI] *Consider the linear-like PI*

$$(A(x)+g(x)\hat{\Pi}_k^1)^T\hat{P}_k^1 + \hat{P}_k^1(A(x)+g(x)\hat{\Pi}_k^1) + \hat{\Pi}_k^{1^T}R\hat{\Pi}_k^1 + Q = 0, \tag{45}$$

$$\hat{u}_{k+1}^{*1} = -R^{-1}g^T(x)\hat{P}_k^1(x)x. \tag{46}$$

We call (45)-(46) the linear-like plain-PI, the pair $(\bar{u}^1(x), \bar{P}^1(x))$ its limit solution, and $\mathscr{P}_1$ and $\mathscr{P}_2$ the regions in which $x^T\bar{P}^1(x)x \ge 0$ and $x^T\bar{P}^1(x)x < 0$, respectively.

The structure for the plain-PI is given in Algorithm 1.

Definition 5 [The $\mathscr{P}_1$-PI] *Consider for all $x \in \mathscr{P}_1$ the linear-like PI*

$$\begin{aligned} &(A(x)+g(x)\hat{\Pi}_{k,i}^2)^T\hat{P}_{k,i}^2 + \hat{P}_{k,i}^2(A(x)+g(x)\hat{\Pi}_{k,i}^2) + \hat{\Pi}_{k,i}^{2^T}R\hat{\Pi}_{k,i}^2 \\ &+ Q + \dot{\bar{P}}_{i-1}^2|_{A(x)x+g(x)\bar{u}_{i-1}^2} = 0, \end{aligned} \tag{47}$$

$$\hat{u}_{k+1,i}^{*2} = -R^{-1}g^T(x)\hat{P}_{k,i}^2(x)x, \tag{48}$$

where the index i indicates the outer iteration (Lines 1-11 in Algorithm 2) and one completed linear-like PI (47)-(48) over the index k (Lines 5-8 in Algorithm 2), with k indicating the inner iteration and one PI step for a fixed index i. $\dot{\bar{P}}_{i-1}^2|_{A(x)x+g(x)\bar{u}_{i-1}^2}$ is the time derivative of $\bar{P}_{i-1}^2$ along the trajectories of the system when $\bar{u}_{i-1}^2 = \bar{\Pi}_{i-1}^2 x$. Both $\dot{\bar{P}}_{i-1}^2$ and $\bar{u}_{i-1}^2$ are obtained from the $(i-1)^{th}$ PI (47)-(48) as the respective solutions, meaning that $\dot{\bar{P}}_{i-1}^2|_{A(x)x+g(x)\bar{u}_{i-1}^2}$ is a fixed matrix function during the i^{th} PI (47)-(48). The initial admissible control $\bar{u}_0^2$ and the matrix $\dot{\bar{P}}_0^2$ required for the first PI (47)-(48) ($i = 1$, $k = 0$), are taken from the solutions of the PI (45)-(46), as $\bar{u}_0^2 = \bar{u}^1$ and $\dot{\bar{P}}_0^2 = \dot{\bar{P}}^1$. We call (47)-(48) the linear-like $\mathscr{P}_1$-PI and the pair $(\bar{u}^2(x), \bar{P}^2(x))$ its limit solution.

The control construction based on the linear-like $\mathscr{P}_1$-PI from Def. 5 can be interpreted for each fixed value i as a plain-PI from Def. 4 with a modified state-cost, $Q_i^m = Q + \dot{\bar{P}}_{i-1}^2|_{A(x)x+g(x)\bar{u}_{i-1}^2}$, see (47) and Lines 4 and 6 in Algorithm 2. Whenever Q_i^m is a positive definite matrix, it is then possible to find a unique and positive definite solution $\hat{P}_{k,i}^2$ from (47) at each iteration step over the index k. Lemma 4 shows that this is the case for all $x \in \mathscr{P}_1$ and every k and j.

This means that the linear-like $\mathscr{P}_1$-PI aims to find the solution through the state-cost modification. The underlying rationale is to see whether convergence can be obtained by replacing $\dot{\bar{P}}_i^2|_{A(x)x+g(x)\bar{u}_i^2}$ in (17) with $\dot{\bar{P}}_{i-1}^2|_{A(x)x+g(x)\bar{u}_{i-1}^2}$ from the preceding $(i-1)^{th}$ PI iteration. In the latter case, we still

Algorithm 2 $\mathscr{P}_1$-PI($A(x)$, $g(x)$, Q, R, $\bar{u}_0^2 = \bar{u}^1$, $\bar{P}_0^2 = \bar{P}^1$)

```
 1: for $i \leftarrow 1$ to $M^2$ do                          ▷ outer loop with $M^2$ steps
 2:     $\dot{\bar{P}}_{i-1}^2|_{A(x)x+g(x)\bar{u}_{i-1}^2} \leftarrow$ TimeDerivative($\bar{P}_{i-1}^2$, $\bar{u}_{i-1}^2$)
 3:     $\hat{\Pi}_{0,i}^2 \leftarrow \bar{\Pi}_{i-1}^2$                      ▷ policy term from $\bar{u}_{i-1}^2$
 4:     $Q_i^m \leftarrow Q + \dot{\bar{P}}_{i-1}^2$                  ▷ cost state modification
 5:     for $k \leftarrow 0$ to $N^2$ do        ▷ inner loop with $N^2 + 1$ steps
 6:         $\hat{P}_{k,i}^2 \leftarrow$ CuSdle($A(x)$, $g(x)$, $Q_i^m$, $R$, $\hat{\Pi}_{k,i}^2$)
                                                              ▷ eq. (47)
 7:         $\hat{\Pi}_{k+1,i}^2 \leftarrow -R^{-1} g^T(x) \hat{P}_{k,i}^2(x)$          ▷ eq. (48)
 8:     end for
 9:     $\bar{P}_i^2 \leftarrow \hat{P}_{k,i}^2$
10:     $\bar{u}_i^2 \leftarrow \hat{\Pi}_{k+1,i}^2 x$
11: end for
12: return $\bar{u}^2 \leftarrow \bar{u}_i^2$; $\bar{P}^2 \leftarrow \bar{P}_i^2$
```

Algorithm 3 $\mathscr{P}_2$-PI($A(x)$, $g(x)$, Q, R, $\bar{u}_0^3 = \bar{u}^1$, $\bar{P}_0^3 = \bar{P}^1$)

```
 1: for $j \leftarrow 1$ to $M^3$ do                          ▷ outer loop with $M^3$ steps
 2:     $\dot{\bar{P}}_{j-1}^3|_{A(x)x+g(x)\bar{u}_{j-1}^3} \leftarrow$ TimeDerivative($\bar{P}_{j-1}^3$, $\bar{u}_{j-1}^3$)
 3:     $\hat{\Pi}_{0,j}^3 \leftarrow \bar{\Pi}_{j-1}^3$                      ▷ Policy term from $\bar{u}_{j-1}^3$
 4:     $u_{j,corr} \leftarrow$ QuadraticEq($R$, $\dot{\bar{P}}_{j-1}^3$)               ▷ eq. (51)
 5:     for $k \leftarrow 0$ to $N^3$ do        ▷ inner loop with $N^3 + 1$ steps
 6:         $\hat{P}_{k,j}^3 \leftarrow$ CuSdle($A(x)$, $g(x)$, $Q$, $R$, $\hat{\Pi}_{k,j}^3$)
                                                              ▷ eq. (49)
 7:         $\hat{u}_{k+1,j}^{*3} \leftarrow -R^{-1} g^T(x) \hat{P}_{k,j}^3(x) x + u_{j,corr}$
                                                              ▷ eq. (50)
 8:     end for
 9:     $\bar{P}_j^3 \leftarrow \hat{P}_{k,j}^3$
10:     $\bar{u}_j^3 \leftarrow \hat{u}_{k+1,j}^{*3}$
11: end for
12: return $\bar{u}^3 \leftarrow \bar{u}_j^3$; $\bar{P}^3 \leftarrow \bar{P}_j^3$
```

solve the SDLE instead of the HJB equation.

Once convergence is achieved over k (Lines 5-8 in Algorithm 2), that is the new pair $(\bar{u}_i^2, \bar{P}_i^2)$ is obtained (Lines 9 and 10 in Algorithm 2), we repeat the procedure until convergence of the outer PI over the index i is achieved as well (Lines 1-11 in Algorithm 2). The proof of convergence is given in Lemma 6.

Definition 6 [The $\mathscr{P}_2$-PI] *Consider, for all $x \in \mathscr{P}_2$, the linear-like PI*

$$(A(x) + g(x)\hat{\Pi}_{k,j}^3)^T \hat{P}_{k,j}^3 + \hat{P}_{k,j}^3 (A(x) + g(x)\hat{\Pi}_{k,j}^3) + \hat{\Pi}_{k,j}^{3^T} R \hat{\Pi}_{k,j}^3 + Q = 0 \tag{49}$$

$$\hat{u}_{k+1,j}^{*3} = \hat{\Pi}_{k+1,j}^3 x = -R^{-1} g^T(x) \hat{P}_{k,j}^3(x) x + u_{j,corr}\,, \tag{50}$$

where $u_{j,corr}$ is the correction component obtained as solution to the quadratic matrix equation

$$u_{j,corr}^T R u_{j,corr} + x^T \dot{\bar{P}}_{j-1}^3|_{A(x)x+g(x)(\bar{u}_{j-1}^3 + u_{j,corr})} x = 0. \tag{51}$$

The index j indicates the outer iteration (Lines 1-11 in Algorithm 3) and one completed linear-like PI (49)-(50) over the index k (Lines 5-8 in Algorithm 3), with k indicating the inner iteration and one PI step for a fixed index j. $\dot{\bar{P}}_{j-1}^3|_{A(x)x+g(x)(\bar{u}_{j-1}^3+u_{j,corr})}$ is the time derivative of $\bar{P}_{j-1}^3$ along the trajectories of the system when the control $\bar{u}_{j-1}^3 + u_{j,corr}$ is used. Both $\dot{\bar{P}}_{j-1}^3$ and $\bar{u}_{j-1}^3$ are obtained from the $(j-1)^{th}$ PI (49)-(50) as the respective solutions, hence $u_{j,corr}$ is consequently as well. This means that $\dot{\bar{P}}_{j-1}^3|_{A(x)x+g(x)(u_{j-1}^3+u_{j,corr})}$ is a fixed matrix function during the j^{th} PI (49)-(50). The initial admissible control $\bar{u}_0^3$ and the matrix $\dot{\bar{P}}_0^3$ required for the first PI (49) and (50) ($j = 1$, $k = 0$) are taken from the solutions of the PI (45)-(46), as $\bar{u}_0^3 = \bar{u}^1$ and $\dot{\bar{P}}_0^3 = \dot{\bar{P}}^1$. We call (49)-(50) the linear-like $\mathscr{P}_2$-PI and the pair $(\bar{u}^3(x), \bar{P}^3(x))$ its limit solution.

The control construction based on the linear-like $\mathscr{P}_2$-PI from Def. 6 can be interpreted for each fixed value j as a plain-PI from Def. 4 with a modified control, see (50) and (51), as well as Lines 4 and 7 in Algorithm 3. We show in Lemma 5 that it is possible to find a real-valued solution u_{corr} from (51) for all $x \in \mathscr{P}_2$ and every j.

The algorithm for the $\mathscr{P}_2$-PI (49) and (50) is as follows. For a fixed $\dot{\bar{P}}_{j-1}^3$ and u_{j-1}^3 (Lines 2 and 3 in Algorithm 3), we construct the correction term $u_{j,corr}$ in accordance to (51) (Line 4 in Algorithm 3). Then, (49) is solved (Line 6 in Algorithm 3) to construct the modified control $\hat{u}_{k+1,j}^{*3}$ (50) for every k of the inner iteration (Line 7 in Algorithm 3). The underlying rationale is to see whether convergence can be obtained by modification of the control with respect to the plain-PI. The HJB equation (17) indicates that u_{corr} can be used to cancel out the last term of the left-hand side of (17) whenever it is possible to find a real-valued solution u_{corr} from $u_{corr}^T R u_{corr} + x^T \dot{P}^* x = 0$. In order to simplify the problem, we use (51) based on the pair $(\bar{u}_{j-1}^3, \bar{P}_{i-1}^3)$ obtained from the preceding $(i-1)^{th}$ PI iteration. Since the next k steps of the inner iterations (Lines 5-8) requires the policy $\hat{\Pi}_{k+1,j}^3$ of the modified control $\hat{u}_{k+1,j}^{*3}$ obtained in Line (7), one can easily derive it from (50) and (51).

Once convergence is achieved over k (Lines 5-8 in Algorithm 3), that is the new pair $(\bar{u}_j^3, \bar{P}_j^3)$ is obtained (Lines 9 and 10 in Algorithm 3), we repeat the procedure until convergence of the outer PI over the index j (Lines 1-11 in Algorithm 3) is achieved as well. The proof of convergence is given in Lemma 7.

Lemma 4 [Domain of the $\mathscr{P}_1$-PI] *Consider the $\mathscr{P}_1$- PI from Def. 5. Then,*

$$x^T \dot{\bar{P}}^1 x \geq 0 \Rightarrow x^T \dot{\bar{P}}_{i-1}^2|_{A(x)x+g(x)\bar{u}_{i-1}^2} x \geq 0, \forall i. \tag{52}$$

Proof: In accordance with Def. 5, the initial admissible control $\bar{u}_0^2$ and the matrix $\dot{\bar{P}}_0^2$, required by the $\mathscr{P}_1$-PI (47) and (48), are the solutions of the plain-PI from Def. 4. The limit form of (45) and (46) can be written as

$$(A + g\bar{\Pi}^1)^T \bar{P}^1 + \bar{P}^1 (A + g\bar{\Pi}^1) + \bar{\Pi}^{1^T} R \bar{\Pi}^1 + Q = 0, \tag{53}$$

from which we select $\bar{u}_0^2 = \bar{u}^1$ and $\dot{\bar{P}}_0^2 = \dot{\bar{P}}^1$ to form the $\mathscr{P}_1$-PI for $i=1$. By comparing (53) with the form of the incremental expression of the cost (5) for $\bar{u}^1=\bar{\Pi}^1 x$, that is

$$x^T((A+g\bar{\Pi}^1)^T\bar{P}^1+\bar{P}^1(A+g\bar{\Pi}^1)+\bar{\Pi}^{1^T}R\bar{\Pi}^1+Q+\dot{\bar{P}}^1)x=0, \tag{54}$$

one can see that the optimal cost and the optimal control are already achieved for all x along the hyperplane $x^T\dot{\bar{P}}^1x=0$.

For the i^{th}-outer iteration of the $\mathscr{P}_1$- PI (47) and (48), the inner iteration over the index k can also be considered as a plain-PI, so the convergence can be achieved provided the initial control is admissible. Starting from the first $\mathscr{P}_1$-PI ($i=1$) and the first inner iteration ($k=1$), we take $\bar{\Pi}^1$ for the initial control policy, that is $\hat{\Pi}^2_{k=1,i=1}=\bar{\Pi}^1$, which is already optimal in the set $x^T\dot{\bar{P}}^1x=x^T\dot{\bar{P}}^2_0x=0$. Furthermore, once we complete the iteration over the index k for $i=1$, the optimality in this set is preserved due to the convergence of the plain-PI.

More generally, once convergence of the i^{th} iteration is achieved over the index k, the form (55) follows from (47) and (48) and

$$\begin{aligned}&(A+g\bar{\Pi}^2_i)^T\bar{P}^2_i+\bar{P}^2_i(A+g\bar{\Pi}^2_i)+\bar{\Pi}^{2^T}_iR\bar{\Pi}^2_i\\&+Q+\dot{\bar{P}}^2_{i-1}|_{Ax+g\bar{u}^2_{i-1}}=0.\end{aligned} \tag{55}$$

It is therefore possible to conclude from (55) that the optimal cost and the optimal control are already achieved for all x in the set $x^T\dot{\bar{P}}^2_{i-1}x=0$. On the other hand, any optimal pair $(\bar{\Pi}^2_i,\bar{P}^2_i)$ related to the original state-cost x^TQx must satisfy

$$(A+g\bar{\Pi}^2_i)^T\bar{P}^2_i+\bar{P}^2_i(A+g\bar{\Pi}^2_i)+\bar{\Pi}^{2^T}_iR\bar{\Pi}^2_i+Q+\dot{\bar{P}}^2_i=0, \tag{56}$$

which holds only for x along $x^T\dot{\bar{P}}^2_ix=0$. Assuming now that the convergence is not yet achieved along the index i, this further means that the optimality is not achieved for all other x which are outside the set $x^T\dot{\bar{P}}^2_ix=0$. However, since the optimality is already achieved for $x^T\dot{\bar{P}}^2_{i-1}x=0$, we conclude that the sets $x^T\dot{\bar{P}}^2_{i-1}x=0$ and $x^T\dot{\bar{P}}^2_ix=0$ represent the same state-space set. This means that the initial hypersurface $x^T\dot{\bar{P}}^1x=0$, along which the optimal control and the optimal cost are achieved by the plain-PI, is preserved for all i of the $\mathscr{P}_1$-PI. This discussion proves that $x^T\dot{\bar{P}}^1x=0\Rightarrow x^T\dot{\bar{P}}^2_ix=0,\forall i$.

Additionally, one can interpret from (53) and (54) that, in order to achieve the optimal cost and the optimal control by a form of linear-like PI, one needs to properly modify the state-cost x^TQx. It can be seen that, for all x for which $x^T\dot{\bar{P}}^1(x)x>0$, the modified state-cost has to be larger than x^TQx. Assume now that $\exists x: x^T\dot{\bar{P}}^2_ix<0$ for $x^T\dot{\bar{P}}^2_{i-1}x>0$. This would mean that for such x the modified state-cost has to be smaller than x^TQx, which is in contradiction with the indication of the preceding iterations. This conclusion holds for the initial iteration as well, where $x^T\dot{\bar{P}}^2_0x\geq 0$, that is $x^T\dot{\bar{P}}^1x\geq 0$, which completes the proof.■

Lemma 5 [Domain of the $\mathscr{P}_2$-PI] *Consider the $\mathscr{P}_2$- PI from Def. 6. Then,*

$$x^T\dot{\bar{P}}^1(x)x<0\Rightarrow\dot{\bar{P}}^3_{j-1}|_{A(x)x+g(x)(\bar{u}^3_{j-1}+u_{j,corr})}<0,\forall j. \tag{57}$$

Proof: Starting from (50), we can write the preceding control of the inner iteration in the form

$$\hat{u}^{*3}_{k,j}=\hat{\Pi}^3_{k,j}x=-R^{-1}g^T\hat{P}^3_{k-1,j}x+u_{j,corr}=\hat{\Pi}^{3'}_{k,j}x+u_{j,corr}. \tag{58}$$

If we now plug the term $\hat{\Pi}^{3'}_{k,j}x+u_{j,corr}$ into (49) in a similar manner as in (20)-(22), we obtain

$$\begin{aligned}&x^T((A+g\hat{\Pi}^{3'}_{k,j})^T\hat{P}^3_{k,j}+\hat{P}^3_{k,j}(A+g\hat{\Pi}^{3'}_{k,j})+\hat{\Pi}^{3'^T}_{k,j}R\hat{\Pi}^{3'}_{k,j}+Q)x\\&+u^T_{j,corr}Ru_{j,corr}=0.\end{aligned} \tag{59}$$

Using (51), we can replace the last term to obtain

$$\begin{aligned}&x^T((A+g\hat{\Pi}^{3'}_{k,j})^T\hat{P}^3_{k,j}+\hat{P}^3_{k,j}(A+g\hat{\Pi}^{3'}_{k,j})+\hat{\Pi}^{3'^T}_{k,j}R\hat{\Pi}^{3'}_{k,j}+Q)x\\&-x^T\dot{\bar{P}}^3_{j-1}|_{A(x)x+g(x)(\bar{u}^3_{j-1}+u_{j,corr})}x=0.\end{aligned} \tag{60}$$

This means that (49)-(51) can be replaced with (60) together with a non-modified control law

$$\hat{u}^{*3}_{k+1,j}=-R^{-1}g^T(x)\hat{P}^3_{k,j}(x)x=\hat{\Pi}^{3'}_{k+1,j}x, \tag{61}$$

and (51) in order to construct the linear-like $\mathscr{P}_2$-PI. However, it is worth noting that the last term of (60) exists only in case $u_{j,corr}\neq 0$ due to (51). Observing now (60) and (61), the proof of Lemma 5 can be completed in a similar manner as in Lemma 4.■

In the following, we show the convergence of both the $\mathscr{P}_1$ and $\mathscr{P}_2$ policy-iterations and introduce the proposed optimal control.

Lemma 6 [Convergence of the $\mathscr{P}_1$-PI] *The linear-like $\mathscr{P}_1$-PI is convergent for all $x\in\mathscr{P}_1$.*

Proof: The i^{th}-outer iteration of the $\mathscr{P}_1$-PI can be considered a plain-PI with a modified state-cost matrix $\bar{Q}=Q+\dot{\bar{P}}^2_{i-1}$. This means that it is convergent for every inner iteration over the index k provided the initial control is admissible (Theorem 1). Since the initial control policy $\hat{\Pi}^2_{k=1,i}$ for the i^{th}-outer iteration is taken as the convergent solution of the preceding $(i-1)^{th}$-outer iteration, that is $\hat{\Pi}^2_{k=1,i}$=$\bar{\Pi}^2_{i-1}$, we conclude the convergence of each inner-iteration cycle over the index k. Moreover, by recalling Lemma 3 and taking into account that the $\mathscr{P}_1$-PI uses a modified state-cost $\bar{Q}$, the real cost $V^m_{k,i}(\bar{Q})$ of the control $\bar{u}^2_i$ converges to

$$V^{m,2}_i(\bar{Q})=x^T\bar{P}^2_ix+\int_0^\infty x^T\dot{\bar{P}}^2_ixdt \tag{62}$$

when $k\to\infty$ during the i^{th} -outer iteration. This leads to the real cost of the control $\bar{u}^2_i$ which is related to the original state-cost Q, that is

$$V^{m,2}_i(Q)=x^T\bar{P}^2_ix+\int_0^\infty x^T\dot{\bar{P}}^2_ixdt-\int_0^\infty x^T\dot{\bar{P}}^2_{i-1}xdt. \tag{63}$$

To complete the proof, we show that the cost $V_i^{m,2}(Q)$ does at least not increase during the transition from the i^{th} to the $(i+1)^{th}$-outer iteration. Once convergence of the i^{th}-outer iteration is achieved over the index k, (47) becomes (55), where the first expression for $k=1$ of the $(i+1)^{th}$-outer iteration is given as

$$\begin{aligned}&(A+g\bar{\Pi}_i^2)^T\hat{P}^2_{k=1,i+1}+\hat{P}^2_{k=1,i+1}(A+g\bar{\Pi}_i^2)+\bar{\Pi}_i^{2^T}R\bar{\Pi}_i^2\\&+Q+\dot{\bar{P}}_i^2|_{Ax+g\bar{u}_i^2}=0,\end{aligned} \tag{64}$$

The difference between (64) and (55) cab be written in the form

$$\begin{aligned}&(A+g\bar{\Pi}_i^2)^T\Delta P^2_{i+1}+\Delta P^2_{i+1}(A+g\bar{\Pi}_i^2)\\&+\dot{\bar{P}}_i^2|_{Ax+g\bar{u}_i^2}-\dot{\bar{P}}_{i-1}^2|_{Ax+g\bar{u}_{i-1}^2}=0,\end{aligned} \tag{65}$$

where $\Delta P^2_{i+1}=\hat{P}^2_{k=1,i+1}-\bar{P}_i^2$. Observing now the real cost of the control $\bar{u}_i^2$ for $k=1$ during the i^{th}-outer iteration, which is

$$V^{m,2}_{k=1,i+1}(Q)=x^T\hat{P}^2_{k=1,i+1}x+\int_0^\infty x^T\dot{\hat{P}}^2_{k=1,i+1}xdt-\int_0^\infty x^T\dot{\bar{P}}_i^2xdt, \tag{66}$$

it can be seen that (65) can be written in compact form as

$$\frac{d}{dt}\Delta V^{m,2}_{i+1}(Q)=0, \tag{67}$$

where the first time derivative is taken along the trajectories of the system when the control policy $\bar{\Pi}_i^2$ is used, and $\Delta V^{m,2}_{i+1}(Q)=V^{m,2}_{k=1,i+1}(Q)-V_i^{m,2}(Q)$. Similar to the discussion provided in Theorem 1, here one obtains $V^{m,2}_{k=1,i+1}(Q)=V_i^{m,2}(Q)$.■

Once the overall convergence is achieved, where $\lim_{i\to\infty}\Delta V^{m,2}_{i+1}(Q)=0$ and $\lim_{i\to\infty}\Delta P^2_{i+1}=0$, the modified cost (63) and the control (48) become

$$\bar{V}^2=\lim_{i\to\infty}V_i^{m,2}(Q)=x^T\bar{P}^2x, \tag{68}$$

$$\bar{u}^2=-R^{-1}g^T(x)\bar{P}^2(x)x. \tag{69}$$

Lemma 7 [Convergence of the $\mathscr{P}_2$-PI] *The linear-like $\mathscr{P}_2$-PI is convergent for all $x\in\mathscr{P}_2$.*

Proof: In Lemma 5, one can see that (49)-(51), which form the proposed linear-like $\mathscr{P}_2$, can be replaced with (60), (61) and (51). Starting from the latter, one can complete the proof of Lemma 7 in a similar manner as for the proof of Lemma 6.■

It is worth noting that here, at the end of the $j^{th}-outer$ iteration, the cost of the $\bar{u}_j^3$ which is related to the original state-cost matrix Q converges to

$$\begin{aligned}V_j^{m,3}(Q)=&x^T\bar{P}_j^3x+\int_0^\infty x^T\dot{\bar{P}}_j^3xdt|_{A(x)x+g(x)\bar{u}_j^3}\\&+\int_0^\infty x^T\dot{\bar{P}}^3_{j-1}x|_{A(x)x+g(x)(\bar{u}^3_{j-1}+u_{j,corr})}dt.\end{aligned} \tag{70}$$

Once the overall convergence is achieved, the correction term $u_{j,corr}$ vanishes, that is $\lim_{j\to\infty}u_{j,corr}=0$. If we now recall (51), one can write

$$\begin{aligned}&\lim_{j\to\infty}u_{j,corr}=0\Rightarrow\lim_{j\to\infty}\int_0^\infty x^T\dot{\bar{P}}^3_{j-1}x|_{A(x)x+g(x)(\bar{u}^3_{j-1}+u_{j,corr})}=\\&\lim_{j\to\infty}\int_0^\infty x^T\dot{\bar{P}}_j^3x|_{A(x)x+g(x)\bar{u}_j^3}=\lim_{j\to\infty}\int_0^\infty x^T\dot{\bar{P}}^3x|_{A(x)x+g(x)\bar{u}^3}=0.\end{aligned} \tag{71}$$

Accordingly, the last two terms from (70) vanish as well, so the modified cost (70) and the control (50) become

$$\bar{V}^3=\lim_{j\to\infty}V_j^{m,3}(Q)=x^T\bar{P}^3x, \tag{72}$$

$$\bar{u}^3=-R^{-1}g^T(x)\bar{P}^3(x)x. \tag{73}$$

Theorem 2 [Optimal control] *Assume that the optimal value function is in the form (13). Then the optimal control $u^*(x)\in\mathscr{A}(\Omega)$ is given as*

$$u^*(x)=\begin{cases}\bar{u}^2 & \mathit{if}\ x\in\mathscr{P}_1\\ \bar{u}^3 & \mathit{if}\ x\in\mathscr{P}_2,\end{cases} \tag{74}$$

where $\bar{u}^2$ and $\bar{u}^3$ are the final solutions of the linear-like $\mathscr{P}_1$ and $\mathscr{P}_2$ policy iterations, respectively.

Proof: In accordance with Lemma 6, (63) represents the cost of the control $\bar{u}_i^2$ for $x\in\mathscr{P}_1$, which means that the pair $(\bar{u}_i^2,V_i^{m,2}(Q))$ satisfies the incremental expression of the cost (5) which can be written as (55). On the other hand, it can be seen that the control $\bar{u}_i^2$ minimizes (55), which means that it is optimal with respect to the cost $V_i^{m,2}(Q)$. Since in the limiting case when $i\to\infty$, (55) converges to

$$\begin{aligned}&(A+g\bar{\Pi}^2)^T\bar{P}^2+\bar{P}^2(A+g\bar{\Pi}^2)+\bar{\Pi}^{2^T}R\bar{\Pi}^2\\&+Q+\dot{\bar{P}}^2|_{Ax+g\bar{u}^2}=0,\end{aligned} \tag{75}$$

which is the original incremental expression of the cost of the control $\bar{u}^2$ while the cost is given in the quadratic form (68), this leads to the conclusion that the limit pair $(\bar{u}^2,\bar{V}^2)$ is optimal for $x\in\mathscr{P}_1$.

In accordance with Lemma 7, (70) represents the cost of the control $\bar{u}_j^3$ for $x\in\mathscr{P}_2$, which means that the pair$(\bar{u}_j^3,V_j^{m,3}(Q))$ satisfies the incremental expression of the cost (5) which can be written as

$$\begin{aligned}&(A+g\bar{\Pi}_j^3)^T\bar{P}_j^3+\bar{P}_j^3(A+g\bar{\Pi}_j^3)+\bar{\Pi}_j^{3^T}R\bar{\Pi}_j^3\\&+Q-\dot{\bar{P}}^3_{j-1}|_{A(x)x+g(x)(\bar{u}^3_{j-1}+u_{j,corr})}=0.\end{aligned} \tag{76}$$

In the limiting case when $j\to\infty$, (76) converges to

$$\begin{aligned}&(A+g\bar{\Pi}^3)^T\bar{P}^3+\bar{P}^3(A+g\bar{\Pi}^3)+\bar{\Pi}^{3^T}R\bar{\Pi}^3\\&+Q=0,\end{aligned} \tag{77}$$

and the last two terms of (70) vanish as well. It can be seen that the control $\bar{u}^3$ minimizes (77), while (77) can be rewritten in the form of the original incremental expression of the cost of the control $\bar{u}^3$ with the quadratic function (72), that is

$$\begin{aligned}&(A+g\bar{\Pi}^3)^T\bar{P}^3+\bar{P}^3(A+g\bar{\Pi}^3)+\bar{\Pi}^{3^T}R\bar{\Pi}^3\\&+Q+x^T\dot{\bar{P}}^3|_{A(x)x+g(x)\bar{u}^3}x=0,\end{aligned} \tag{78}$$

since $x^T \dot{\bar{P}}^3|_{A(x)x+g(x)\bar{u}^3} x = 0$ holds in the limiting case. This leads to the conclusion that the limit pair $(\bar{u}^3, \bar{V}^3)$ is optimal for $x \in \mathscr{P}_2$ as well.■

VI. Illustrative examples

We provide simulation results by considering four nonlinear systems. For the first and the second system the optimal control and the optimal value function are known, so it is possible to assess the proposed approach against the optimal solution. However, the nonlinear system used in the second example is obtained through nonlinear transformation of a linear system providing a testing case with a non-constant matrix $Q = Q(x)$. In the third example we cover a nonlinear system for which the optimal control and the optimal value function are also known, however, the value function is not quadratic-like. In the fourth example we compare our approach against the control based on the Galerkin approximation (GAC) by considering a nonlinear system with an unknown optimal control policy and an unknown optimal value function. These last two examples illustrate an additional capability of the proposed approach to even solve such nonlinear control problems.

In all examples, after the plain-PI (45)-(46) is used, we complete only one outer iteration of the $\mathscr{P}_1$-PI (47)-(48) and the $\mathscr{P}_2$-PI (49)-(50), that is $i = 1$ and $j = 1$, and then three inner iterations of the $\mathscr{P}_1$-PI and the $\mathscr{P}_2$-PI. The number of the inner iterations used for the plain-PI will be indicated in each example.

A. Optimal control of the Van Der Pol oscillator

Consider the Van Der Pol oscillator

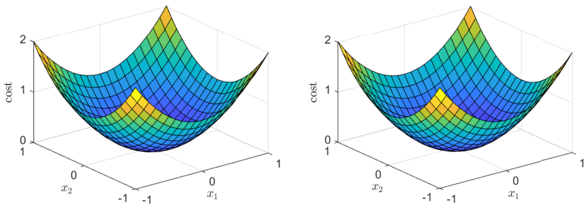
$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -x_1 - \mu(1 - x_1^2)x_2 + x_1 u, \tag{79}$$

with $\mu = 0.5$, the state cost $l(x) = x_2^2$ and assume $R = 1$. The optimal control is $u^* = -x_1 x_2$, with optimal value function $V^* = x_2^1 + x_2^2$. The system can be easily factorized with

$$A(x) = \begin{bmatrix} 0 & 1 \\ -1 & -\frac{1}{2}(1 - x_1^2) \end{bmatrix}, \quad B(x) = \begin{bmatrix} 0 \\ x_1 \end{bmatrix}. \tag{80}$$

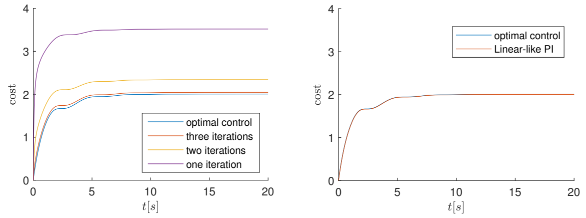
The initial admissible control for the linear-like PI can be selected to be the one that cancels out the nonlinearities and stabilizes the system, that is $u = -\frac{1}{2}x_1 x_2$ (e.g., $\Pi = [0 \;\; -\frac{1}{2}x_1]$). In such a case, the initial control used for the PIs is similar to the optimal control, so the convergence is expected to be achieved quickly. Fig. 1 shows the optimal value function (left) and the value function obtained by the proposed approach in three inner iterations of the plain-PI and three inner iterations of each $\mathscr{P}_1$-PI and the $\mathscr{P}_2$-PI.

However, in order to illustrate how convergence improves based on three iterations of the plain-PI without using the $\mathscr{P}_1$-PI and the $\mathscr{P}_2$-PI, we start from a different initial control $u = -\frac{1}{2}x_1 x_2 - 0.1 x_1^3 x_2$ (e.g., $\Pi = [-0.5x_2 \;\; -0.1x_1^3]$) which also ensures the asymptotic stability of the feedback system (79). Fig. 2 illustrates how the associated cumulative cost (left) converges depending on the number of iterations used in the plain-PI (45)-(46). In some examples, the plain-PI can achieve an optimal solution without modifications based on the $\mathscr{P}_1$-PI and the $\mathscr{P}_2$-PI. In any case, the proposed approach based on the plain-PI with three iterations of the $\mathscr{P}_1$-PI and the $\mathscr{P}_2$-PI achieves the optimal cost value (right).



(a) (b)

Fig. 1: The Van Der Pol oscillator: The optimal value function (left) and the value function obtained by the proposed approach using only three inner iterations (right) of the plain-PI.

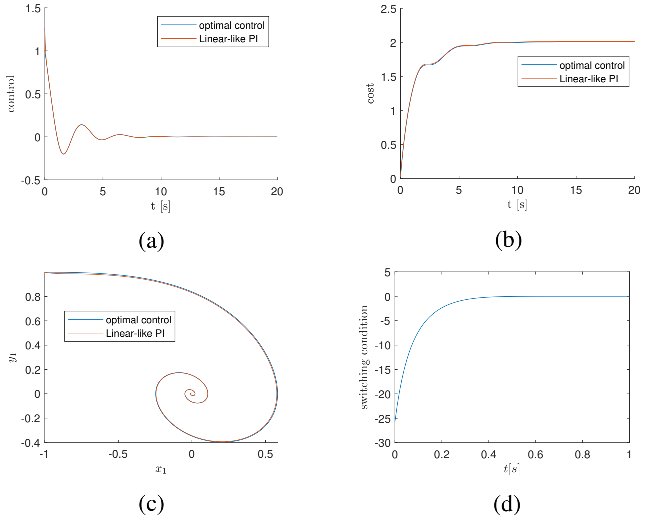


(a) (b)

Fig. 2: Cumulative costs obtained for the initial condition $x = [-1; 1]$ using only the plain PI with a different number of inner iterations (left); and the cumulative cost of the proposed approach based on three iterations of the $\mathscr{P}_1$-PI and the $\mathscr{P}_2$-PI (right).

Fig. 3 provides a comparison between the proposed approach and the optimal control for the system for $x_0 = [-1; 1]$ and three inner iterations of the plain-PI with three inner iterations of the $\mathscr{P}_1$-PI and the $\mathscr{P}_2$-PI. From the control signals, cumulative costs and phase portrait, we conclude optimality of the proposed solution. In Fig. 3 one can also see the switching function $x^T \dot{\bar{P}}^1 x$ along the trajectories of the system, which is used in (74). This function indicates the time intervals in which the two different forms of the optimal control (74) have been used. Another interesting observation is that this function becomes zero before the state reaches the origin. This phenomenon has not been investigated in this work, and it can be a promising direction for further understanding of the proposed framework. This means that the system trajectory has approached the hyper-surface $x^T \dot{\bar{P}}^1 x = 0$ (in this example, a curve) and then has moved along this surface towards the equilibrium. Somewhat surprisingly, once the system state is on this hyper-surface, along which the PIs (45)-(46), (47)-(48) and (49)-(50) are equivalent, one only needs the linear-like PI (45)-(46) to obtain the remaining part of the optimal control. Fig. (4) show different shapes of the switching function obtained for different initial conditions.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2022.3226671



(a) (b)

(c) (d)

Fig. 3: Comparison between the optimal and the proposed controls in terms of control signals (a), cumulative costs (b), and phase portrait (c) obtained along the trajectory from the initial condition $x = [-1;1]$ and three iterations of the plain-PI ($k = 3$). Subfigure (d) shows the values of the boundary function used in (74), that is $x^T\dot{\bar{P}}^1 x$.

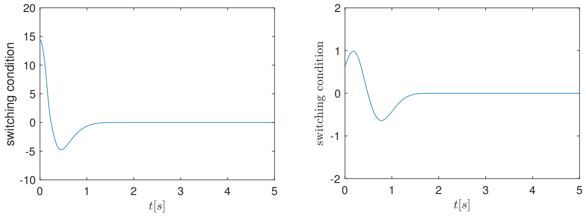


Fig. 4: Switching function for the initial condition $x = [-1;-1]$ (left) and $x = [0.8;0.8]$ (right).

B. A nonlinear system with a non quadratic cost

Consider the nonlinear system

$$\dot{x} = \begin{bmatrix} x_2 - 3x_2^2 - 2x_1x_2 + x_2^3 \\ -2x_1 - 3x_2 + x_2^2 \end{bmatrix} + \begin{bmatrix} x_2 \\ 1 \end{bmatrix} u, \tag{81}$$

with the state cost $l(x) = (x_1 - \frac{1}{2}x_2)^2 + x_2^4$ and assume $R = 1$. This system can be obtain by transforming the linear system

$$\dot{\bar{x}}_1 = \bar{x}_2, \quad \dot{\bar{x}}_2 = -2\bar{x}_1 - 3\bar{x}_2 + u, \tag{82}$$

with $\bar{Q}$ being the identity matrix and $\bar{R} = 1$, using the nonlinear transformation $x_2 = \bar{x}_2$ and $x_1 = \bar{x}_1 + \frac{1}{2}\bar{x}_2^2$. The optimal control of the system (81) can be computed through the same nonlinear transformation starting from the optimal control of the linear system (82).

The nonlinear system (81) can be easily factorized with

$$A(x) = \begin{bmatrix} -2x_2 & 1 - 3x_2 + x_2^2 \\ -2 & -3 + x_2 \end{bmatrix}, \quad B(x) = \begin{bmatrix} x_2 \\ 1 \end{bmatrix}, \tag{83}$$

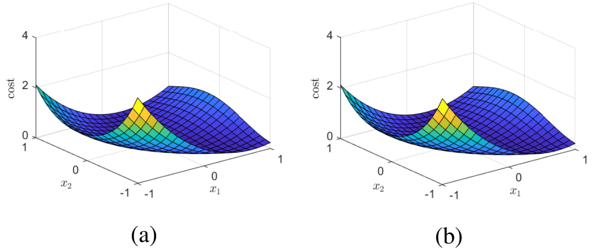
where the state cost can be rewritten in the state-dependent quadratic form $l(x) = x^T Q(x) x$ with

$$Q(x) = \begin{bmatrix} 1 & -\frac{1}{2}x_2 \\ -\frac{1}{2}x_2 & \frac{1}{4} + x_2^4 \end{bmatrix}. \tag{84}$$

In order to illustrate the proposed PI algorithm, the initial admissible control is obtained through the same nonlinear transformation starting from the control of the linear system (82) which allocates the system poles to $(-3, j0)$ and $(-1, j0)$, that is

$$u = -x_1 - x_2 + \frac{1}{2}x_2^2, \;\; \Pi = [-1 \;\; -1 + \frac{1}{2}x_2]. \tag{85}$$

Fig. 5 shows the optimal value function (left) and the value function obtained by the proposed approach (right) using only two inner iterations of the plain-PI indicating the convergence is locally achieved.



(a) (b)

Fig. 5: The optimal value function (left) and the value function obtained by the linear-like PI approach using only two inner iterations of the plain-PI (right) for the nonlinear system with a state-dependent cost.

C. A nonlinear system without a quadratic-like form of the optimal value function

Example 1: Consider the nonlinear system

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = -x_1 + x_2 \sinh(x_1^2 + x_2^2) + u, \tag{86}$$

with the state cost $l(x) = x_2^2$ and assume $R = 1$ [24]. The optimal control of this system

$$u^* = -x_2 e^{x_1^2 + x_2^2} \tag{87}$$

has an associated value function which is not quadratic-like, that is

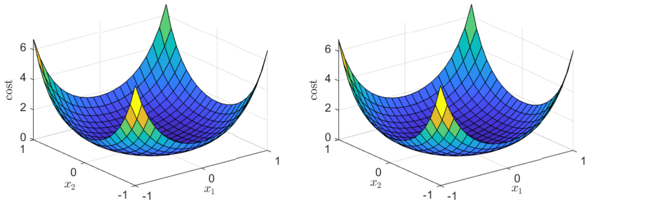
$$V^* = e^{x_1^2 + x_2^2} - 1. \tag{88}$$

The nonlinear system (86) can be easily factorized with

$$A(x) = \begin{bmatrix} 0 & 1 \\ -1 & \sinh(x_1^2 + x_2^2) \end{bmatrix}, \quad B(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{89}$$

for which the initial admissible control for the linear-like PI is selected to be the one that cancels out the nonlinearities and stabilizes the system, that is $u = -x_2 - x_2 \sinh(x_1^2 + x_2^2)$ (e.g., $\Pi = [0 \;\; -1 - \sinh(x_1^2 + x_2^2)]$).

Fig. 6 shows the optimal value function (left) and the value function obtained by the proposed approach (right) using only two inner iterations of the plain-PI indicating the convergence is locally achieved. The obtained results



(a) (b)

Fig. 6: The optimal value function (left) and the value function obtained by the proposed approach using only two inner iterations of the plain PI (right) for the system without quadratic-like form of optimal value function.

also suggests that the proposed PI algorithm has a potential to construct optimal control even in case an optimal value function is not a quadratic-like.

D. A nonlinear system with unknown optimal control

Consider the nonlinear system

$$\dot{x} = \begin{bmatrix} -x_1^3 - x_2 \\ x_1 + x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \tag{90}$$

with the state cost $l(x) = x_1^2 + x_2^2$ and assume $R = 1$. The system can be easily factorized with

$$A(x) = \begin{bmatrix} -x_1^2 & -1 \\ 1 & 1 \end{bmatrix}, \quad B(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{91}$$

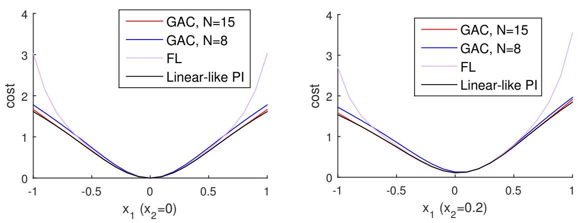
The initial admissible control for the linear-like PI is selected to be the control based on feedback linearization (FL) which is obtained in the form [5]

$$u(x) = 3x_1^5 + 3x_1^2 x_2 - x_2 + 0.4142x_1 - 1.3522(x_1^3 + x_2), \tag{92}$$

for which one possible control policy is

$$\Pi = [3x_1^4 + 0.4142 - 1.3522x_1^2 + 3x_1x_2 \quad -1 - 1.3522]. \tag{93}$$

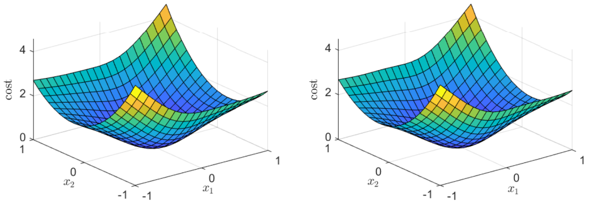
The solution based on Galerkin approximation (GAC) has been obtained for different orders of the approximation and those can be found in [5]. In this example, we use two such controls obtained for $N = \{8, 15\}$.



(a) (b)

Fig. 7: The cost values for different initial conditions for x_1, where $x_2 = 0$ (a) and $x_2 = 0.2$ (b).

We illustrate the comparison of the GAC, FL and the proposed approach in terms of their associated costs as in [5]. Fig. 7 shows the costs that have been obtained for different initial conditions in x_1, while x_2 is constant, that is $x_2 = 0$ (a) and $x_2 = 0.2$ (b). One can observe that the proposed linear-like policy-iteration generates the minimal cost, although the GAC with $N = 15$ is similar. However, we stress that the GAC requires a number of preconditions for a valid implementation [5].



(a) (b)

Fig. 8: The optimal value function (left) and the value function obtained by the linear-like PI approach using only one inner iteration of the plain PI (right) for the nonlinear system with an unknown optimal value function.

Fig. 8 shows the optimal value function (left) and the value function obtained by the proposed approach (right) using only one inner iteration of the plain-PI indicating the convergence is locally achieved. The obtained results also suggests that the proposed PI algorithm has a potential to construct optimal control even in case an optimal value function is not known.

VII. Conclusions

We have developed a method to determine optimal control strategies for continuous-time nonlinear systems. In particular, we have defined the approximate linear-like PI based on the SDLE to compute an approximate control law. Stabilizability of such an approximate policy-update is proved with Lemma 2, its cost is derived in Lemma 3, while convergence of this control law is provided in Theorem 1. Section V includes the main result in which Definitions 4-6 introduce three slightly different linear-like PIs. Lemmas 4-7 includes the proofs of their convergence properties, while Theorem 2 provides the description of the optimal control law algorithm based on these PIs.

The algorithm has been tested using four different nonlinear systems, including the Van Der Pol oscillator, a nonlinear system with a non quadratic cost, a nonlinear system without a quadratic-like optimal value function and a nonlinear system with unknown optimal control. From the results obtained one can observe the optimality of the proposed approach and the fast local convergence. The results also suggest that the proposed approach has the potential to be used in cases in which the optimal value function is not quadratic-like.

References

[1] D. Bertsekas, Dynamic programming and optimal control. Vol. 1. No. 2. Belmont, MA: Athena scientific, 1995.

[2] M.G. Crandall, and P.L. Lions. "Viscosity solutions of Hamilton-Jacobi equations." Transactions of the American mathematical society 277.1 (1983): 1-42.

[3] R.J. Leake, and R.W. Liu. "Construction of suboptimal control sequences." SIAM Journal on Control 5.1 (1967): 54-63.

[4] A. Wernrud, and A. Rantzer. "On approximate policy iteration for continuous-time systems." Proceedings of the 44th IEEE Conference on Decision and Control, 2005.

[5] R.W. Beard, G.N. Saridis, and J. T. Wen. "Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation." Automatica 33.12 (1997): 2159-2177.

[6] A. Isidori, Nonlinear control systems: an introduction. Springer Berlin Heidelberg, 1985.

[7] H.K. Khalil. Nonlinear control. Vol. 406. New York: Pearson, 2015.

[8] A. Astolfi, D. Karagiannis, and R. Ortega. Nonlinear and adaptive control with applications. Vol. 187. London: Springer, 2008.

[9] R. Sepulchre, M. Jankovic, and P. V. Kokotovic. Constructive nonlinear control. Springer Science and Business Media, 2012.

[10] R. Freeman, and P. V. Kokotovic. Robust nonlinear control design: state-space and Lyapunov techniques. Springer Science and Business Media, 2008.

[11] K.L. Teo, B. Li, C. Yu, and V. Rehbock. "Applied and computational optimal control." Optimization and Its Applications (2021).

[12] A.J. Krener, "The existence of optimal regulators." Proceedings of the 37th IEEE Conference on Decision and Control, 1998.

[13] W.M. McEneaney, "A curse-of-dimensionality-free numerical method for solution of certain HJB PDEs." SIAM journal on Control and Optimization, 2007.

[14] A. Wernli, and G. Cook. "Suboptimal control for the nonlinear quadratic regulator problem." Automatica 11.1 (1975): 75-84.

[15] R. Marino. "An example of a nonlinear regulator." IEEE Transactions on Automatic Control, 1984.

[16] M. Sassano, and A. Astolfi. "Dynamic approximate solutions of the HJ inequality and of the HJB equation for input-affine nonlinear systems." IEEE Transactions on Automatic Control, 2012.

[17] J.D. Pearson. "Approximation methods in optimal control I. Suboptimal control." International Journal of Electronics, 1962.

[18] C.P. Mracek, and J.R. Cloutier. "Control designs for the nonlinear benchmark problem via the state-dependent Riccati equation method." International Journal of robust and nonlinear control 8.4-5 (1998): 401-433.

[19] T. Cimen. "State-dependent Riccati equation (SDRE) control: A survey." IFAC Proceedings Volumes, 2008.

[20] A. Tahirovic, and S. Dzuzdanovic. "A globally stabilizing nonlinear model predictive control framework." 55th Conference on Decision and Control. IEEE, 2016.

[21] A. Tahirovic, and F. Janjos. "A class of SDRE-RRT based kinodynamic motion planners." 2018 American Control Conference. IEEE.

[22] A. Tahirovic, and A. Astolfi. "Optimal control for continuous-time nonlinear systems based on a linear-like policy iteration." 58th Conference on Decision and Control. IEEE, 2019.

[23] D. Vrabie et al. "Adaptive optimal control for continuous-time linear systems based on policy iteration." Automatica 45.2 (2009): 477-484.

[24] R.A Freeman, and J.A. Primbs. "Control Lyapunov functions: New ideas from an old source." Proceedings of 35th IEEE Conference on Decision and Control. IEEE, 1996.

Adnan Tahirović was born in Doboj-Tešanj, Bosnia and Herzegovina. He graduated and obtained his M.Sc. degree from the Faculty of Electrical Engineering, University of Sarajevo, in 2006, from which he received the Golden Plaque best student of generation award. In 2011 he was awarded a Ph.D. degree in Information Technology by the Department of Electronics and Information, Politecnico di Milano, Italy. He was a researcher at NASA JPL, Pasadena, USA, in 2010, and Imperial College London, in 2011. He is currently an Associate Professor at the Department of Automatic Control and Electronics of the Faculty of Electrical Engineering, University of Sarajevo. Since 2018 he also holds a visiting research position at Imperial College London. His research interests include nonlinear control theory, motion planning in mobile robotics, autonomous and multiagent systems, as well as control theory applications in the domain of artificial intelligence and computational neuroscience.

Alessandro Astolfi 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 at 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 of Imperial College London, London (UK), where he is currently Professor of Nonlinear Control Theory and Head of the Control and Power Group. From 1998 to 2003 he was also an Associate Professor at the Dept. of Electronics and Information of the Politecnico of Milano. Since 2005 he has also been a Professor at Dipartimento di Ingegneria Civile e Ingegneria Informatica, University of Rome Tor Vergata. He has been 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), 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, of 30 book chapters and of over 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). He 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 PaperAward(2017). He is a“Distinguished Member”of the IEEE CSS,IEEE Fellow and IFAC Fellow. He served as Associate Editor for Automatica, Systems and Control Letters, the IEEE Trans. on Automatic Control, the International Journal of Control, the European Journal of Control and the Journal of the Franklin Institute; as Area Editor for the Int. J. of Adaptive Control and Signal Processing; as Senior Editor for the IEEE Trans. on Automatic Control; and as Editor-in-Chief for the European Journal of Control. He is currently Editor-in-Chief of the IEEE Trans. on Automatic Control. He served as Chair of the IEEE CSS Conference Editorial Board (2010–2017) and in the IPC of several international conferences. He has been/is a Member of the IEEE Fellow Committee (2016), (2019–2022).