Integral Reinforcement Learning for Continuous-Time Input-Affine Nonlinear Systems with Simultaneous Invariant Explorations

Jae Young Lee, *Student Member, IEEE*, Jin Bae Park*, *Member, IEEE*, and Yoon Ho Choi

***Abstract*—This paper focuses on a class of reinforcement learning (RL) algorithms, named integral RL (I-RL), that solve continuous-time (CT) nonlinear optimal control problems with input-affine system dynamics. First, we extend the concepts of exploration, integral temporal difference, and invariant admissibility to the target CT nonlinear system that is governed by a control policy plus a probing signal called an exploration. Then, we show input-to-state stability (ISS) and invariant admissibility of the closed-loop systems with the policies generated by integral policy iteration (I-PI) or invariantly admissible PI (IA-PI) method. Based on these, three online I-RL algorithms named explorized I-PI and integral Q-learning I, II are proposed, all of which generate the same convergent sequences as I-PI and IA-PI under the required excitation condition on the exploration. All the proposed methods are partially or completely model-free, and can simultaneously explore the state-space in a stable manner during the online learning processes. ISS, invariant admissibility, and convergence properties of the proposed methods are also investigated, and related with these, we show the design principles of the exploration for safe learning. Neural-network-based implementation methods for the proposed schemes are also presented in this paper. Finally, several numerical simulations are carried out to verify the effectiveness of the proposed methods.**

***Index Terms*—reinforcement learning, policy iteration, adaptive optimal control, Q-learning, continuous-time, exploration**

I. Introduction

REINFORCEMENT LEARNING (RL) is a class of learning algorithms that originates from and is inspired by biological animal learning mechanisms, and is designed to learn the best policy by interacting with a given *unknown* environment to maximize their long-term performance [1]–[4]. From the very beginning of the research, RL methods have been extensively studied in the fields of computational intelligence, with special focus on the finite Markov decision process (MDP) [3]. As a result, a variety of RL algorithms in MDP environments have been proposed, including Sarsa, Q-learning, and actor-critic methods, with successful applications [2]–[6] (see [6] for survey). These RL methods were developed based on the two core ideas: *i) temporal difference (TD)* and *ii) exploration/exploitation*, both of which now become the fundamental components in RL [2]–[4].

TD error is one step prediction error, indicating how far the estimated value function is from the true one for the current policy [1]–[4]. Here, the value function, and its action-dependent version called Q-function, implicitly express the long-term performance index for the current policy, and play a central role in modifying the agent's current policy in RL. All RL methods equip at least one update rule, the objective of which is to estimate the value function (or Q-function) by decreasing the associated TD error [7]–[25].

Associated with the exploration/exploitation in RL, there is *an exploitation vs. exploration dilemma* [3], [4]: to achieve an improved response, the RL agents should *exploit* the information they obtain, but at the same time *explore* the whole environments to improve future actions. In RL methods for MDPs, a sufficient number of explorations of each state-action pair is required for the learning of best response, and thus exploitation and exploration should be properly balanced during the learning period [3]–[5]. In this paper, we focus on the exploitation/exploration issues and TD learning methods in RL applied to continuous-time (CT) dynamical systems for adaptive optimal control.

A. HJB Equations, TD Error, and PI in Optimal Control

In the fields of control system engineering, optimal control theories have been developed as one of the fundamental principles in the design of modern control systems [26], [27]. The optimal control policy minimizes a given long-term cost-to-go function, which specifies the desired performance with respect to the system states and control inputs in the long run, implicitly balancing the amount of required control efforts and the desired transient response. Basically, such a minimizing policy can be found using either Pontryagin's minimum principle or dynamic programming. *Both optimal control approaches, however, are intrinsically off-line and require complete knowledge of the system dynamics.*

In optimal control problems, the centerpiece of dynamic programming is the Hamilton-Jacobi-Bellman (HJB) equation [26]–[28] which is termed as the Bellman's optimality equation in the field of computational intelligence (see [2]–[4]). The optimal value function and control policy can be obtained by solving the HJB equation. However, this is a formidable task even in the case of completely known dynamics due to the intractability of the HJB partial differential equation and a problem known as *the curse of dimensionality*.

This research was supported by by Institute of BioMed-IT, Energy-IT and Smart-IT Technology (BEST), a Brain Korea 21 plus program, Yonsei University, and by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (No. NRF-2013R1A1A2012609).

J. Y. Lee and J. B. Park are with Department of Electrical and Electronic Engineering, Yonsei University, 50 Yonsei-ro, Seodaemum-Gu, Seoul, Korea, (e-mail: {jyounglee, jbpark}@yonsei.ac.kr).

Y. H. Choi is with Department of Electronic Engineering, Kyonggi University, Suwon, Kyonggi-Do, Korea, (e-mail: yhchoi@kyonggi.ac.kr).

* Corresponding author

To alleviate these problems, researchers in the fields of both control systems and machine learning have studied the forward-in-time computational methods known as policy iteration (PI) [7]–[17] and value iteration (VI) [17]–[20]. Both methods consist of two processes—*policy evaluation* and *policy improvement*, and the difference between PI and VI lie in policy evaluation (see [1], [3], [21] for a comprehensive understanding). Unlike VI, PI minimizes the associated TD error at each step of policy evaluation to compute the exact value function which can be considered as a Lyapunov function in many cases [1].

Combining the forward-in-time computational methods with neural networks (NNs) in discrete-time (DT) domain, Werbos [19] and years after, Prokhorov and Wunch [20] proposed adaptive dynamic programming (ADP) methods for optimal neuro-control. In these methods, one NN called *critic* plays a role of policy evaluation for approximating the value function (or Q-function), and the other NN called *actor* learns the improved policy as a role of policy improvement. Independently of these works, the inverse optimal neuro-control methods [29], [30] are developed, where the cost function is *a posteriori* determined for the stabilizing feedback. In the inverse optimal approach, one does not need to solve the HJB equation, but cannot freely choose the required cost function unlike the aforementioned ADP methods. The proposed RL methods in this paper are a kind of online PI combined with NNs, which efficiently find, by forward-in-time computation, the CT optimal control solution satisfying the underlying HJB equation, where the cost function is given *a priori*.

B. Exploitation and Exploration in Adaptive Optimal Control

Until recently, the exploitation and exploration issues in control systems have been posed from the perspectives of adaptive/neural control, with the direct connection to the notion of persistent excitation (PE) [31], [32], [33, Section 7.6.2]. Without satisfaction of PE conditions, the learning parameters cannot converge to the true values. On the other hand, PE conditions make the system states and control inputs oscillatory, not convergent to the equilibrium, and even cause unbounded closed-loop responses in many cases, especially when the states escape the stable region of the nonlinear systems [32], [33, Section 8.2]. In this respect, the exploitation vs. exploration dilemma can be reinterpreted from a control theoretic perspective as a dilemma between the satisfaction of PE (efficient exploration) and the satisfactory control performance (exploitation to improve the stability and state convergence).

Without considering the optimality of the target control policy, considerable efforts have been made to design an adaptive controller which efficiently balances the exploitation and exploration for good transient performance by regulating the magnitudes of the exploratory random signals injected through the control input channels [31], [32], [33, Section 7.6.2]. Similar adaptive optimal control schemes for dynamical systems have been also presented in the areas of RL and ADP [13]–[18], [22]–[24], where the so-called probing noises are injected through the control input channels to maintain the PE conditions required to learn the target *optimal* policy and the value function (or Q-function). Among these RL and ADP methods, the iterative Q-learning schemes based on PI [13]–[16] and VI [18] have revealed that the probing noises, called explorations in our previous works [13], [14], play a central role in relaxing the requirements of the system model dynamics. The key point here is that by virtue of the explorations, these learning controllers can be applied to systems with *completely unknown dynamics* to find the optimal policy in an online fashion with guaranteed convergence.

C. CT Nonlinear Adaptive Optimal Control

In CT domain, RL and ADP algorithms to solve the HJB equation were developed without any proof of their stability and convergence at first. After the pioneering works of [11], [12], [21], a class of RL methods named integral RL (I-RL) was presented with stability and convergence studies for solving CT input-affine nonlinear optimal control problems with unknown system drift dynamics (see [1] for a comprehensive survey). These I-RL methods are based on PI and VI in the CT domain which minimize or decrease the associated integral TD (I-TD) error at each step. For the I-RL schemes developed from PI [7]–[9] in particular, which we call integral PI (I-PI) in this paper, the stability and convergence to the optimal solution are proven under the initial admissible policy [12]. The implementation methods based on least squares (LS) and the Galerkin NN approximation were also presented in [12]. However, *all the underlying I-RL methods require exact knowledge of the input-coupling dynamics [1], [21], which restricts the practical use of the algorithms.* Second, unlike RL schemes for DT dynamical systems [16], [18], *there is no way to simultaneously explore the state-space for the satisfaction of PE conditions during online learning*—resetting the state variables to some non-zero points is the only way in these I-RL methods [1], [12], [21] to give additional excitations to the state variables. Lastly, *the given initial admissible region in I-PI is not invariant for the intermediate closed-loop systems, so that the state trajectory may escape the well-defined stable region during the learning phase, especially at the time instant the policy is updated* [10].

For the same CT optimal control problems, these ideas of I-RL are combined with adaptive control methodologies, and the resulting online actor-critic RL methods, called synchronous PI, were recently proposed in [22]–[25]. These methods simultaneously update the actor and critic NNs in an online fashion, and the stability of the RL control system is proven under the assumption of PE. Though exhibiting novel ideas, *they require complete/partial knowledge of the system dynamics [22]–[24] and/or an additional NN to identify the system dynamics [24], [25].* In our previous work [13], we combined I-PI with both CT Q-function and exploration to develop two online advanced I-RL algorithms: explorized PI and integral Q-learning (see also [15] for a similar model-free algorithm). Both algorithms can explore the whole state-space and the latter can be applied to completely unknown dynamics. However, *these algorithms were focused only on solving CT LQR problems and were not extended to general CT nonlinear optimal control frameworks.*

D. Contributions

Based on I-PI in [12] and IA-PI in our recent work [10], this paper proposes the three partially/completely model-free I-RL algorithms, all of which are extensions of the work [13] to CT nonlinear optimal control with input-affine dynamics, and use NNs to approximate the value function and control policy. Here, I-PI and IA-PI provide the basic ideas of I-TD and invariant admissibility (the combined concept of invariance and admissibility), respectively. All the proposed methods can simultaneously and stably explore the state-space during online learning, and can update the given admissible region and the exploration signal to guarantee input-to-state stability (ISS) and invariant admissibility. Moreover, to the best authors' knowledge, the two methods out of the three, named integral Q-learning I and II, respectively, are the first ones that can learn the online optimal solution *in completely unknown dynamics without any use of the additional identifier NN*, which was not achieved by any existing methods [12], [22]–[25]. In the proposed methods, the cost function is *a priori* that can be freely chosen without such restrictions as in inverse optimal approaches [29], [30].

To develop the I-RL methods, this paper extends the concepts of exploration, I-TD, and invariant admissibility to the CT nonlinear input-affine dynamical system that is governed by a control policy and a probing signal called an exploration. Then, we mathematically show i) ISS and invariant admissibility regarding both explorations and policies generated by either IA-PI or I-PI, and ii) the excitation condition on the exploration to uniquely solve the advanced I-TD equation. This unique solution is equal to the solution to the I-TD equation used in I-PI [12] and the associated Lyapunov equation used in PI [9] and IA-PI [10]. From these mathematical results, we propose one partially model-free I-RL method named explorized I-PI, and two model-free I-RL methods named integral Q-learning I and II, respectively. ISS, invariant admissibility, and convergence properties of the proposed I-RL methods, related with the design of explorations, are also given under the uniqueness condition. Finally, NN-based actor-critic LS implementation methods are presented, and we simulate the proposed I-RL methods to verify their performance and to compare them with the others [12], [22], [24].

E. Notations and Mathematical Terminologies

In this paper, the following notations are adopted for a real vector $x \in \mathbb{R}^n$ and any real matrices X and Y.

- $\|x\|$: the Euclidean norm $\sqrt{x^T x}$ of x;
- $\bar{\sigma}(X)$: the maximum singular value of X;
- $\underline{\sigma}(X)$: the minimum singular value of X;
- $X \otimes Y$: the Kronecker product of X and Y;
- $B_0(r)$: an open r-radius ball $\{x : \|x\| < r\}$;
- $\bar{B}_0(r)$: a closed r-radius ball $\{x : \|x\| \le r\}$.

For any N-vectors $x_j \in \mathbb{R}^{n_j}$ $(j = 1, 2, \cdots, N)$, the column stacking operator $\mathrm{col}\{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^{n_1+n_2+\cdots+n_N}$ is defined as $\mathrm{col}\{x_1, x_2, \cdots, x_N\} := [\ x_1^T\ x_2^T\ \cdots\ x_N^T\ ]^T$. We also denote the set of nonnegative integers and real numbers by $\mathbb{Z}_+$ and $\mathbb{R}_+$, respectively. For any two sets Ω_1 and Ω_2 in $\mathbb{R}^n$, "$\Omega_1 \subseteq \Omega_2$" and "$\Omega_1 \subset \Omega_2$" indicate that Ω_1 is a subset and a proper subset of Ω_2, respectively; $\partial\Omega_1$ denotes the boundary of Ω_1. For a given domain $\mathcal{D} \subseteq \mathbb{R}^n$, the set of all continuous and continuously differentiable functions are denoted by $C^0(\mathcal{D})$ and $C^1(\mathcal{D})$, respectively.

Definition. *A function $V : \mathcal{D} \to \mathbb{R}_+$ is said to be positive definite (on $\mathcal{D}$) if it is continuous on $\mathcal{D}$, $V(0) = 0$, and*

$$V(x) > 0, \quad \forall x \in \mathcal{D} \setminus \{0\}.$$

Definition. *A continuous function $\alpha : [0, a) \to [0, \infty)$ is of class $\mathcal{K}$, denoted by $\alpha \in \mathcal{K}$, if it is strictly increasing and $\alpha(0) = 0$. A class $\mathcal{K}$ function α is of class $\mathcal{K}_\infty$, denoted by $\alpha \in \mathcal{K}_\infty$, if $a = \infty$, and $\lim_{r\to\infty} \alpha(r) = \infty$.*

Definition. *A continuous function $\beta : [0, a) \times [0, \infty) \to [0, \infty)$ is of class $\mathcal{KL}$, denoted by $\beta \in \mathcal{KL}$, if $\beta(\cdot, s) \in \mathcal{K}$ for each fixed s, and for each fixed r, $\beta(r, \cdot)$ is decreasing and $\beta(r, s) \to 0$ as $s \to \infty$.*

We denote the gradient of a function $f : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$ by

$$\nabla f(x) := \left[\frac{\partial f(x)}{\partial x_1},\ \frac{\partial f(x)}{\partial x_2},\ \cdots,\ \frac{\partial f(x)}{\partial x_n} \right]^T \in \mathbb{R}^n,$$

where x_j $(1 \le j \le n)$ is the j-th element of $x \in \mathcal{D}$. For a vector-valued function $f(x) = \big[f_1(x),\, f_2(x), \cdots,\, f_m(x)\big]^T \in \mathbb{R}^m$, $\nabla f(x)$ is meant to be a matrix of the first-order derivatives of the form

$$\nabla f(x) := \big[\nabla f_1(x),\, \nabla f_2(x), \cdots,\, \nabla f_m(x)\big] \in \mathbb{R}^{n\times m}.$$

Throughout this paper, t indicates a specific time instant on $[0, \infty)$ and $\tau \in [t, \infty)$ will be used as the time variable after the specified time instant t. In addition, any function $s(\tau)$ of time τ will be denoted as $s(\tau)$, s_τ, or simply s for conciseness.

II. Nonlinear Optimal Control Problems

Consider the following CT input-affine nonlinear system:

$$\dot{x}_\tau = f(x_\tau) + g(x_\tau)u(x_\tau), \quad x(t) = z \in \mathcal{D} \subseteq \mathbb{R}^n \qquad (1)$$

where $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ are the state variable and the control input; z is the state value at given initial time instant $\tau = t$; $\mathcal{D} \subseteq \mathbb{R}^n$ is a set containing a neighborhood of the origin; $f : \mathcal{D} \to \mathbb{R}^n$ with $f(0) = 0$ and $g : \mathcal{D} \to \mathbb{R}^{n\times m}$ are nonlinear functions that are locally Lipschitz on $\mathcal{D}$. For simplicity, we restrict our domain of interest to a neighborhood of the origin, *i.e.*, without loss of generality, we assume that

Assumption 1. *$\mathcal{D} = \bar{B}_0(r_d)$ for some $r_d > 0$.*

The results in this paper can be easily extended to a general domain $\mathcal{D}$ without Assumption 1. For a well-posed problem, we assume $f(x) + g(x)u(x)$ is locally Lipschitz on $\mathcal{D}$ and there is a policy $u = \mu(x)$ that stabilizes the system (1).

Definition 1. *A policy $\mu(x)$ is meant to be a control input function $\mu : \mathcal{D} \to \mathbb{R}^m$ that is continuous on the domain of interest and satisfies $\mu(0) = 0$.*

For a stabilizing policy μ, define the region of attraction as

$$R_A(\mu) := \big\{z \in \mathcal{D} : x_\tau(z; \mu, 0) \to 0 \text{ as } \tau \to \infty\big\},$$

where $x_\tau(z;\mu,0)$ denotes the state trajectory $x(\tau)$ at time $\tau \geq t$ generated by the system (1) with the initial condition $x_t = z \in \mathcal{D}$ and a policy $u = \mu(x)$. Here, the third parameter '0' in $x_\tau(z;\mu,0)$ indicates the zero-exploration and will be clear in Section IV. For simplicity, we write $x_\tau \equiv x_\tau(z;\mu,0)$ if z and μ are well-understood in the context. Using these notations, we precisely define a feasible trajectory and a stabilizing policy on a given region as follows:

Definition 2. *For a given policy $u = \mu(x)$, we say that the state trajectory $x_\tau(\,\cdot\,;\mu,0)$ is feasible on a subset $\Omega \subseteq \mathcal{D}$ if*

$$z \in \Omega \implies x_\tau(z;\mu,0) \in \mathcal{D}, \ \ \forall \tau \geq t. \tag{2}$$

Definition 3. *A policy $u = \mu(x)$ is said to stabilize the system* (1) *on a subset $\Omega \subseteq \mathcal{D}$ if and only if*

1) *$x_\tau(z;\mu,0)$ exists for all $z \in \Omega$ and all $\tau \geq 0$;*
2) *$x_\tau(\,\cdot\,;\mu,0)$ is feasible on Ω;*
3) *the equilibrium '0' of the system $\dot{x} = f + g\mu$ is stable;*
4) *for all $z \in \Omega$, $\lim_{\tau\to\infty} x_\tau(z;\mu,0) = 0$.*

The CT nonlinear optimal control problem considered in this paper consists of the input-affine dynamics (1) and the following performance index (3):

$$J(x_t, u(\cdot)) = \int_t^\infty r(x_\tau, u_\tau)\, d\tau, \tag{3}$$

where $r(x,u) \in \mathbb{R}$ is the cost function defined as $r(x,u) := S(x) + u^T R u > 0$ for a positive definite function $S : \mathcal{D} \to \mathbb{R}_+$ and a positive definite matrix $R \in \mathbb{R}^{m\times m}$. For this performance index, the value function $V^\mu(z)$ for a policy $u = \mu(x)$ and a given (initial) value $x_t = z \in \mathcal{D}$ is defined, if exists, as

$$V^\mu(z) := J(z, u(\cdot))|_{u=\mu(x)}.$$

For the existence of V^μ, the policy μ needs to stabilize the system (1). However, this is not sufficient for the existence, so we introduce the concept of the admissible policy.

Definition 4. *A policy $u = \mu(x)$ is admissible with respect to* (3) *on a subset $\Omega \subseteq \mathcal{D}$, denoted by $\mu \in \mathcal{A}(\Omega)$, if*

1) *$u = \mu(x)$ stabilizes the system* (1) *on Ω,*
2) *$V^\mu(z) < \infty$, for all $z \in \Omega$.*

Note that $\mu \in \mathcal{A}(\Omega)$ implies $\Omega \subseteq R_A(\mu)$ and the existence of $V^\mu(z)\ \forall z \in \Omega$. In this case, V^μ is positive definite on the subset $\Omega \subseteq \mathcal{D}$ since $r(x,\mu(x))$ in (3) is positive definite on $\mathcal{D}$. Moreover, this property can be easily extended to the larger domain $\mathcal{D}$ by assigning a fictitious value to $V^\mu(z)$ for each $z \in \mathcal{D} \setminus \Omega$ such that V^μ is positive definite on $\mathcal{D}$. Therefore, Assumption 1 and [34, Lemma 4.3] imply the existence of $\underline{\alpha}_\mu$, $\bar{\alpha}_\mu \in \mathcal{K}$ satisfying

$$\underline{\alpha}_\mu(\|x\|) \leq V^\mu(x) \leq \bar{\alpha}_\mu(\|x\|) \tag{4}$$

for all $x \in \mathcal{D}$. Similarly, since $S(x)$ is positive definite on $\mathcal{D}$, there exist $\underline{\alpha}_s$, $\bar{\alpha}_s \in \mathcal{K}$ such that

$$\underline{\alpha}_s(\|x\|) \leq S(x) \leq \bar{\alpha}_s(\|x\|) \tag{5}$$

holds for all $x \in \mathcal{D}$. These class $\mathcal{K}$ functions in (4) and (5) will be used in the analysis of the proposed I-RL algorithms.

For admissibility on Ω (or closed-loop stability), the trajectory $x_\tau(\,\cdot\,;\mu,0)$ should be well-defined on Ω in the sense of Definition 2, so that $x_\tau(z;\mu,0)$ starting from any $z \in \Omega$ remains in the well-defined domain $\mathcal{D}$. This is guaranteed if $\Omega \subseteq R_A(\mu)$ and $\mathcal{D}$ contains $R_A(\mu)$ or an invariant estimate of $R_A(\mu)$ containing Ω. However, $R_A(\mu)$ depends on the policy μ, so we can hardly determine such $\mathcal{D}$ independently of μ. To overcome this difficulty and thereby, guarantee the trajectory well-defined, we introduce the concept of invariant admissibility (see our recent work [10] for detailed discussion about this issue).

Definition 5. *A policy $u = \mu(x)$ is invariantly admissible with respect to* (3) *on a subset $\Omega \subseteq \mathcal{D}$, denoted by $\mu \in \mathcal{A}_\mathcal{I}(\Omega)$, if μ is admissible on Ω and*

$$z \in \Omega \implies x_\tau(z;\mu,0) \in \Omega, \ \ \forall \tau \geq t. \tag{6}$$

The invariance (6) in Definition 5 obviously implies the feasibility condition (2). Moreover, if Ω is compact, then the existence of $x_\tau(z;\mu,0)\ \forall z \in \Omega$ and $\forall \tau \geq t$ is guaranteed by (6) and [34, Theorem 3.3]. For this reason, we assume throughout the paper that the invariant admissible set Ω in Definition 5 is compact.

Define the Hamiltonian $H(x,u,p)$ as

$$H(x,u,p) := r(x,u) + p^T(f(x) + g(x)u). \tag{7}$$

Assuming $\mu \in \mathcal{A}(\Omega)$ and V^μ is $C^1(\Omega)$, then it satisfies the following Lyapunov equation for the nonlinear system (1):

$$H(x, \mu(x), \nabla V^\mu(x)) = 0, \ \ \forall x \in \Omega, \tag{8}$$

which is actually the infinitesimal version of (3) and implies

$$\begin{aligned}\dot{V}^\mu(x_\tau) &\equiv (\nabla V^\mu(x_\tau))^T \big(f(x_\tau) + g(x_\tau)\mu(x_\tau)\big) \\ &= -r(x_\tau, \mu(x_\tau)) < 0.\end{aligned} \tag{9}$$

That is, $V^\mu(x)$ is a Lyapunov function for the system (1) [34]. Now, we define the V^μ-induced compact set Ω_d^μ as

$$\Omega_d^\mu := \{x \in \mathcal{D} : V^\mu(x) \leq d\}$$

for some constant $d > 0$, and state two technical lemmas which will be used in the analysis of proposed I-RL methods. The proof of Lemma 1 is in Appendix A.

Lemma 1. *For $\mu \in \mathcal{A}(\Omega)$, if $V^\mu \in \mathcal{C}^1(\Omega)$, the value function V^μ is the unique solution to* (8) *over $\mathcal{C}^1(\Omega)$.*

Lemma 2. *Suppose $V^\mu(x)$ is finite on a compact subset Ω of $\mathcal{D}$. Let $d > 0$ be a constant chosen in the interval $(0, \min_{x\in\partial\Omega} V^\mu(x))$. Then, Ω_d^μ is in the interior of Ω.*

Proof: Assume Ω_d^μ is not in the interior of Ω. Then, there is a point $y \in \Omega_d^\mu$ on the boundary $\partial\Omega$. At this point, $d < \min_{x\in\partial\Omega} V^\mu(x) \leq V^\mu(y)$, but for all $x \in \Omega_d^\mu$, $V^\mu(x) \leq d$, a contradiction '$d < d$'. So, Ω_d^μ is in the interior of Ω. ■

The objective of the I-RL algorithms presented in this paper is to find the best admissible policy μ^* minimizing the performance index (3) and the corresponding optimal value function $V^*(x) := V^{\mu^*}(x)$. Minimizing the Hamiltonian

$H(x, \mu, \nabla V^*)$ among all admissible policies, we can obtain the optimal policy $\mu^*(x)$ as follows:

$$\mu^*(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V^*(x). \tag{10}$$

Furthermore, substituting (10) into (8) and rearranging the equation yield the well-known HJB equation:

$$0 = S(x) + \nabla V^{*T} f(x) - \frac{1}{4} \nabla V^{*T} g(x) R^{-1} g^T(x) \nabla V^*,$$
$$V^*(0) = 0. \tag{11}$$

The existence of $V^* \in \mathcal{C}^1$ satisfying (11) is the necessary and sufficient condition for optimality.

III. PI AND I-RL WITHOUT EXPLORATIONS

The objective of PI [7]–[10] and I-RL (I-PI [12]) methods is to find the solution V^* to the HJB equation (11) and the corresponding optimal policy (10) by iterations. All the I-RL methods proposed in this paper have the same purpose, *i.e.*, finding the optimal solution V^* and μ^*. In this preliminary section, we present and briefly discuss the existing PI and I-RL methods [9], [10], [12] without considering explorations.

Fig. 1 shows the whole process of a PI method, called IA-PI in our recent work [10]. Normally, PI methods consist of policy evaluation for solving the nonlinear Lyapunov equation (8) and policy improvement for updating the policy by the rule

$$\mu_{i+1}(x) = -\frac{1}{2} R^{-1} g^T(x) \cdot \nabla V^{\mu_i}(x) \tag{12}$$

(see [1], [9]). In addition to this, the IA-PI in Fig. 1 updates the next region Ω_{i+1} in the inv. admissible region update step for the invariant admissibility of μ_i and μ_{i+1} on Ω_{i+1}. This step can be omitted if one can find a universal admissible set Ω on which $\mu_i \in \mathcal{A}(\Omega)$ $\forall i \in \mathbb{Z}_+$ and let Ω_i be equal to Ω $\forall i \in \mathbb{Z}_+$. In this case, IA-PI becomes the PI given in [8], [9].

For each $\mu_i \in \mathcal{A}(\Omega_i)$, let Ψ_i be a compact subset such that

1) $\Omega_i \subseteq \Psi_i \subset R_A(\mu_i) \cap \mathcal{D}$;
2) $x_\tau(\cdot, \mu_i, 0)$ is feasible on Ψ_i.

An example of such Ψ_i is Ω_i in IA-PI (see Fig. 1), but we consider the general case where Ψ_i satisfies the above two conditions. By [10, Lemma 2], such Ψ_i for $\mu_i \in \mathcal{A}(\Omega_i)$ guarantees $\mu_i \in \mathcal{A}(\Psi_i)$, and thereby $V^{\mu_i}(x) < \infty$ $\forall x \in \Psi_i$.

Assumption 2. *For each $\mu_i \in \mathcal{A}(\Omega_i)$, V^{μ_i} is $\mathcal{C}^1$ on Ψ_i, and the next region Ω_{i+1} is given by $\Omega_{i+1} = \Omega_{d_i}^{\mu_i}$ for some d_i chosen in the interval $(0, \min_{x \in \partial\Psi_i} V^{\mu_i}(x))$.*

Theorem 1. *Under $\mu_i \in \mathcal{A}(\Omega_i)$ and Assumption 2,*

1) Ω_{i+1} is in the interior of Ψ_i, and $V^{\mu_i} \in \mathcal{C}^1(\Omega_{i+1})$;
2) μ_i and μ_{i+1} are invariantly admissible on Ω_{i+1}.

Proof: $\mu_i \in \mathcal{A}(\Omega_i)$ and [10, Lemma 2] imply that V^{μ_i} is finite on Ψ_i, and we have $V^{\mu_i} \in \mathcal{C}^1(\Psi_i)$ by Assumption 2. So, Lemma 2 with $\Omega = \Psi_i$ implies Ω_{i+1} is in the interior of Ψ_i, which again implies $V^{\mu_i} \in \mathcal{C}^1(\Omega_{i+1})$ since Ω_{i+1} is compact. Moreover, by [10, Theorem 2] with $\Upsilon_i = \Omega_{i+1}$, we have $\mu_i,\ \mu_{i+1} \in \mathcal{A}_\mathcal{I}(\Omega_{i+1})$. ■

Theorem 1 provides a concrete way to construct the next invariantly admissible region Ω_{i+1} at each i-th step, based on

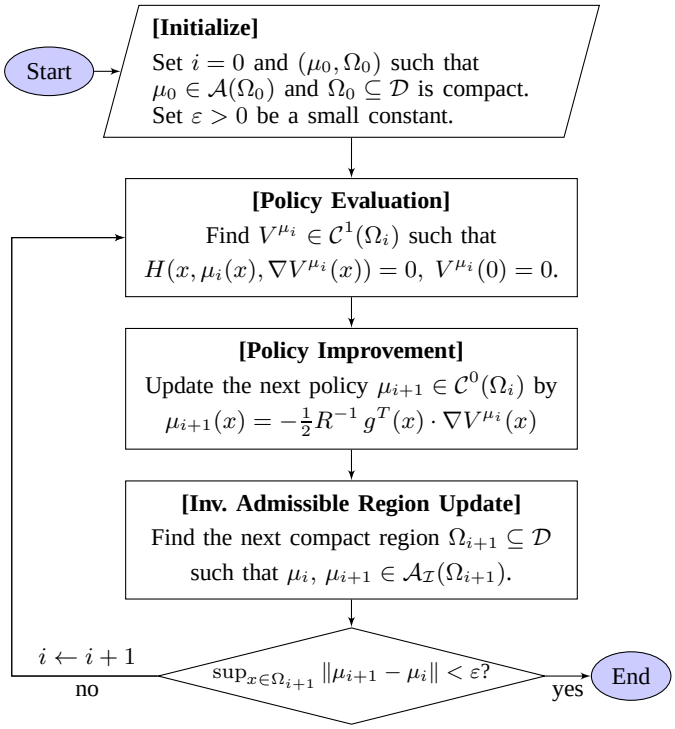


Fig. 1. Description of the IA-PI algorithm [10].

V^{μ_i} and a given admissible region Ψ_i which is a subset of both $R_A(\mu_i)$ and $\mathcal{D}$. However, estimating $R_A(\mu_i)$ or its subset Ψ_i may cause high computational burden, and needs the complete knowledge about the dynamics (f, g). On the other hand, Ψ_i can be chosen as $\Psi_i = \Omega_i$. In this case, Ω_{i+1} becomes rather conservative, but can be determined without such obstacles as shown in the following corollary:

Corollary 1. *Suppose $\mu_i \in \mathcal{A}(\Omega_i)$ and Assumption 2 holds. If Ψ_i is given by $\Psi_i = \Omega_i$, then*

1) Ω_{i+1} is in the interior of Ω_i, and $V^{\mu_i} \in \mathcal{C}^1(\Omega_{i+1})$;
2) μ_i and μ_{i+1} are invariantly admissible on Ω_{i+1}.

Note that since $\mu_{i+1} \in \mathcal{A}_\mathcal{I}(\Omega_{i+1})$ implies $\mu_{i+1} \in \mathcal{A}(\Omega_{i+1})$, *Theorem 1 and Corollary 1 hold for all $i \in \mathbb{Z}_+$ by induction, under $\mu_0 \in \mathcal{A}(\Omega_0)$.* In the case "$\Psi_i = \Omega_i$ for all $\forall i \in \mathbb{Z}_+$," it has been shown in [10] that V^{μ_i} uniformly converges to V^* under certain conditions; this also implies the convergence $\mu_i \to \mu^*$. The convergence in the general case can be also proven in a similar manner under Assumption 2.

Next, integrate (9) from t to $t + T$ to describe I-PI. Then, we obtain the following I-TD equation for a given $\mu \in \mathcal{A}(\Omega)$:

$$V^\mu(x_t) = \int_t^{t+T} r(x_\tau, \mu(x_\tau))\, d\tau + V^\mu(x_{t+T}). \tag{13}$$

This I-TD equation is well-defined for all $x_t \in \Omega$ and for any $T > 0$ if μ is invariantly admissible on Ω since $x_t \in \Omega$ and $\mu \in \mathcal{A}_\mathcal{I}(\Omega)$ guarantees $x_\tau(x_t, \mu, 0)$ remains in the admissible set Ω for all $\tau \in [t, t+T]$. In the same way, IA-PI can be modified by integrating $\dot{V}^{\mu_i}(x_\tau) = -r(x_\tau, \mu_i(x_\tau))$ over the time interval $[t, t+T]$. This modification results in I-PI shown in Algorithm 1, where only the policy evaluation step is described; *the other procedures are exactly same to IA-PI shown in Fig. 1*. When

Algorithm 1. Policy Evaluation of I-PI without Explorations

[Policy Evaluation]

Given $\mu_i \in \mathcal{A}(\Omega_i)$, find $V^{\mu_i} \in \mathcal{C}^1(\Omega_i)$ by solving

$$V^{\mu_i}(x_t) = \int_t^{t+T} r(x_\tau, \mu_i(x_\tau))\ d\tau + V^{\mu_i}(x_{t+T}) \qquad (14)$$

on the domain Ω_i, where $x_\tau \equiv x_\tau(x_t; \mu_i, 0)$.

a universal admissible set Ω is given *a priori* and $\Omega_i = \Omega$ $\forall i \in \mathbb{Z}_+$, then Algorithm 1 becomes the I-PI given in [12].

If both IA-PI and I-PI use the same initial policy μ_0 which is admissible on the same region Ω_0, and generate the same $\{\Omega_{i+1}\}_{i\in\mathbb{Z}_+}$, then they produce the same sequences $\{V^{\mu_i}\}$ and $\{\mu_i\}$. Therefore, I-PI inherits the properties of IA-PI regarding convergence and invariant admissibility discussed above (see also [10], [12]).

Remark 1. *The I-TD* (14) *in Algorithm 1 does not contain any explicit terms about (f, g), and the input coupling term $g(x)$ is only used in policy improvement* (12)*. This makes the algorithm partially model-free, i.e., the system drift dynamics $f(x)$ is not required to be known in I-PI.*

IV. RL Components

To develop the main I-RL methods, we extend the concept of exploration in RL for a finite MDP to nonlinear dynamical systems, and then with detailed mathematical analysis, present advanced I-TD, the refined version of the I-TD (13) with respect to exploration.

A. Exploration in Nonlinear Dynamical Systems

Now, we consider the nonlinear system explored by a known time-varying probing signal e_τ:

$$\dot{x}_\tau = f(x_\tau) + g(x_\tau)[u(x_\tau) + e_\tau], \quad x(t) = z \in \Omega, \qquad (15)$$

where $e : [t, \infty) \to \mathbb{R}^m$ is called an exploration, and assumed piecewise continuous and uniformly bounded; $x_\tau(z; \mu, e)$ denotes the state trajectory $x(\tau)$ at time $\tau \geq t$ generated by the nonlinear system (15) under the given policy $u = \mu(x)$ and exploration e_τ. Unlike in a finite MDP [3] or linear dynamical systems [1], [13], [16], [18], we need the following concept of invariant admissibility of an exploration e.

Definition 6. *For a given policy $\mu \in \mathcal{A}_\mathcal{I}(\Omega)$, an exploration e is said to be invariantly admissible on Ω, denoted by $e \in \mathcal{A}_\mathcal{I}(\Omega; \mu)$ or $(\mu, e) \in \mathcal{A}_\mathcal{I}(\Omega)$, with slight abuse of notation, if*

$$z \in \Omega \implies x_\tau(z; \mu, e) \in \Omega, \ \ \forall \tau \geq t. \qquad (16)$$

Notice that the invariance (16) in Definition 6 is an extension of (6) with respect to an exploration e; if Ω is compact and $(\mu, e) \in \mathcal{A}_\mathcal{I}(\Omega)$, then the existence of $x_\tau(z; \mu, e)$ $\forall z \in \Omega$ and $\forall \tau \geq t$ is guaranteed by (16) and [34, Theorem 3.3], and (16) also implies $x_\tau(\cdot, \mu, e)$ is feasible on Ω. Here, the feasibility of $x_\tau(\cdot, \mu, e)$ can be also defined in a similar manner to Definition 6 by extending Definition 2. Moreover, ISS for the explorized system (15), the stability counterpart of Definition 3, is precisely defined in this paper as follows.

Definition 7. *For a given policy μ and an exploration e, we say that the nonlinear system* (15) *is input-to-state stable on Ω if $x_\tau \equiv x_\tau(z; \mu, e)$ exists $\forall z \in \Omega$ and $\forall \tau \geq t$, x_τ is feasible on Ω, and there exist $\alpha(\cdot) \in \mathcal{K}$ and $\beta(\cdot, \cdot) \in \mathcal{KL}$ such that for any $z \in \Omega$ and all $\tau \geq t$,*

$$\|x_\tau\| \leq \beta\big(\|z\|, \tau - t\big) + \alpha\bigg(\sup_{t\leq s\leq \tau} \|e(s)\|\bigg). \qquad (17)$$

Now, it is stated in the following theorem that ISS and invariant admissibility of e are preserved under the small exploration if the policy is generated by either IA-PI or I-PI.

Theorem 2. *Consider $\{\mu_i\}$ and $\{V^{\mu_i}\}$ generated by IA-PI or I-PI under Assumptions 1 and 2. If the exploration e satisfies*

$$\sup_{t\leq\tau<\infty} \|e(\tau)\| < \sqrt{\frac{\underline{\alpha}_s \circ \bar{\alpha}_{\mu_i}^{-1}(d_i)}{\bar{\sigma}(R)}}, \qquad (18)$$

then $(\mu_{i+1}, e) \in \mathcal{A}_\mathcal{I}(\Omega_{i+1})$ and the nonlinear system (15) *under $u = \mu_{i+1}(x)$ is input-to-state stable on Ω_{i+1}. Moreover, if $\mathcal{D} = \Omega_i = \mathbb{R}^n$ and $\underline{\alpha}_{\mu_i}$, $\underline{\alpha}_s \in \mathcal{K}_\infty$, then ISS holds globally for any $z \in \mathbb{R}^n$ and any bounded exploration e_τ.*

Proof: See Appendix B. ■

B. Advanced I-TD and Design Principles of Exploration

If x_τ is generated by (15) with non-zero exploration, then I-TDs (13) and (14) in policy evaluation of I-PI do not function properly. Meanwhile, if $g(x)$ is not known *a priori*, the next policy μ_{i+1} cannot be updated by policy improvement, either. To solve these two problems, the following e-dependent advanced I-TD is devised from I-TD (13):

$$V(x_t) = \int_t^{t+T} \Big[r(x, \mu(x)) + 2\nu^T(x) R e(\tau)\Big] d\tau + V(x_{t+T}), \qquad (19)$$

where $(\mu, e) \in \mathcal{A}_\mathcal{I}(\Omega)$ is a given policy-exploration pair that is invariantly admissible on a set $\Omega \subseteq \mathcal{D}$; x denotes the trajectory $x_\tau(z; \mu, e)$ for $z = x_t \in \Omega$; $V(x) \in \mathbb{R}$ is positive definite and $\mathcal{C}^1$ on Ω; $\nu(x) \in \mathbb{R}^m$ is a policy to be determined. All the I-RL methods proposed in this paper will be designed based on this advanced I-TD (19). Compared to I-TD (13), the exploration term $\nu^T(x) R e(\tau)$ is added to cancel out the effects of e on I-TD, and to acquire the new policy $\nu(x) = \mu^+(x)$ without knowing $g(x)$ *a priori*. Here, $\mu^+(x)$ is the desired next policy defined in terms of $g(x)$ and $\nabla V(x)$ as

$$\mu^+(x) := -\frac{1}{2} R^{-1} g^T(x) \nabla V(x).$$

For the discussions, we assume without loss of generality that e is T-periodic, *i.e.*, $e_\tau = e_{\tau+T}$ for all $\tau \geq t$.

Theorem 3. *Finding $V \in \mathcal{C}^1(\Omega)$ and a policy ν satisfying* (19) *for all $z = x_t \in \Omega$ is equivalent to solving*

$$H(x, \mu(x), \nabla V(x)) = 2\varphi^T(x) R e_\tau \qquad (20)$$

for all $x \in \Omega$ and $\tau \in [t, t+T)$, where $\varphi(x) := \mu^+(x) - \nu(x)$ is the policy error.

Proof: See Appendix C. ∎

Using Theorem 3, one can easily verify that if $V^\mu \in C^1(\Omega)$,

$$V(x) = V^\mu(x), \quad \nu(x) = \mu^+(x)|_{V=V^\mu} \tag{21}$$

are a solution to the advanced I-TD equation (19) and satisfy

$$H(x, \mu(x), \nabla V(x)) = 0, \quad \varphi(x) = 0, \quad \forall x \in \Omega.$$

However, the solution may not be unique. For example, if $m = 1$ and e_τ is constant, *i.e.*, $e_\tau \equiv c \ \forall \tau \in [t, t+T)$, then Theorem 3 implies ν can be obtained from $V(x)$ and c as

$$\nu(x) = \mu^+(x) + H(x, \mu(x), \nabla V(x))/Rc.$$

This means that for a given $V(x)$, there are infinitely many solutions depending on the constant c unless

$$H(x, \mu(x), \nabla V(x)) \equiv 0.$$

For the case when $g(x)$ is known, $\nu = \mu^+$ can be substituted to (19) to obtain the following simplified advanced I-TD:

$$V(x_t) - V(x_{t+T}) = \int_t^{t+T} \left[r(x, \mu(x)) - \nabla^T V(x) \cdot g(x) e_\tau \right] d\tau. \tag{22}$$

In this case, the solution $V = V^\mu$ to (22) is unique as stated below.

Corollary 2. *Assume that V^μ is $\mathcal{C}^1$ on Ω. If $V \in \mathcal{C}^1(\Omega)$ is the solution to the advanced I-TD* (22)*, then $V = V^\mu$.*

Proof: I-TD (22) is the advanced I-TD (19) with $\varphi(x) = 0$. So, Theorem 3 implies that $V \in \mathcal{C}^1(\Omega)$ satisfying (22) for all $x \in \Omega$ is the solution of the Hamiltonian equation

$$H(x, \mu(x), \nabla V(x)) = 0, \quad \forall x \in \Omega.$$

Then, the application of Lemma 1 concludes $V = V^\mu$. ∎

If $g(x)$ is not known *a priori*, then we cannot substitute $\nu = \mu^+$ to the advanced I-TD (19). In this general case, the uniqueness of (21) depends on the excitation condition. To see this, let $t_j \in [t, t+T]$ $(j = 0, 1, \cdots, L)$ be the time instants satisfying

$$t_0 = t \le t_1 \le t_2 \le \cdots \le t_L = t + T,$$

and assume that e_τ is piecewise constant and determined by

$$e_\tau = c_j, \quad \forall \tau \in [t_j, t_{j+1}), \tag{23}$$

where $\{c_j\}_{j=1}^L$ is a sequence of constant vectors in $\mathbb{R}^m$. We also define the $m \times (l-k)$ matrix $C_{k:l}$ for $1 \le k \le l \le L$ as

$$C_{k:l} := \begin{bmatrix} c_k & c_{k+1} & \cdots & c_l \end{bmatrix}.$$

Then, under the substitution of (23), (20) can be written as

$$H(x, \mu(x), \nabla V(x)) = 2\varphi^T(x) R c_j, \tag{24}$$

for all $x \in \Omega$ and all $j \in \{1, 2, \cdots, L\}$.

Assumption 3a. *There exist κ_1, $\kappa_2 > 0$ such that*

$$\kappa_1 I \le \sum_{j=1}^{L-1} (c_j - c_{j+1})(c_j - c_{j+1})^T \le \kappa_2 I.$$

Theorem 4. *Suppose e_τ is given by* (23) *and $V^\mu \in C^1(\Omega)$. Then, the solution to the advanced I-TD* (19) *is uniquely determined by* (21) *under Assumption 3a.*

Proof: By Theorem 3 and the above discussion, solving (19) for all $x \in \Omega$ is equivalent to finding V and ν satisfying (24) for all $x \in \Omega$ and all $j \in \{1, 2, \cdots, L\}$. From (24), we have $2(c_j - c_{j+1})^T R \varphi(x) = 0$ $(j = 1, 2, \cdots, L-1)$. That is,

$$2(C_{1:L-1} - C_{2:L}) R \varphi(x) = 0. \tag{25}$$

From (25) and Assumption 3a, $\varphi(x) \equiv 0$ is obtained since Assumption 3a is equivalent to

$$\kappa_1 I \le (C_{1:L-1} - C_{2:L})(C_{1:L-1} - C_{2:L})^T \le \kappa_2 I,$$

which implies $\mathrm{rank}\,(C_{1:L-1} - C_{2:L}) = m$. Moreover, the substitution of $\varphi = 0$ into (24) yields $H(x, \mu(x), \nabla V(x)) \equiv 0$. Therefore, the application of Lemma 1 proves $V = V^\mu$, and we obtain $\nu = \mu^+|_{V=V^\mu}$ from $\varphi = 0$. ∎

Under the substitution $V = V^\mu$, (19) can be rewritten as

$$V^\mu(x_t) = \int_t^{t+T} \Big[r(x, \mu(x)) + 2\nu^T(x) R e_\tau \Big] d\tau + V^\mu(x_{t+T}), \tag{26}$$

and (24) is more relaxed to

$$\varphi^T(x) R c_j = 0, \quad \forall x \in \Omega, \quad \forall j \in \{1, 2, \cdots, L\}, \tag{27}$$

due to $H(x, \mu(x), V^\mu(x)) = 0$. In this case, the uniqueness of (21) is guaranteed under the following simple excitation condition:

Assumption 3b. *There exist κ_3, $\kappa_3 > 0$ such that*

$$\kappa_3 I \le \sum_{j=1}^{L} c_j c_j^T \le \kappa_4 I.$$

Theorem 5. *Suppose $V = V^\mu \in C^1(\Omega)$ and e_τ is given by* (23)*. Then, the solution to the advanced I-TD* (19) *is uniquely determined by* (21) *under Assumption 3b.*

Proof: Note that (27) implies $C_{1:L} R \varphi(x) = 0 \ \forall x \in \Omega$, which yields $\varphi(x) \equiv 0$ since $\mathrm{rank}\,(C_{1:L}) = m$ by Assumption 3b. Therefore, by $V = V^\mu$ and $\varphi = 0$, we have $\nu = \mu^+|_{V=V^\mu}$. ∎

Remark 2. *Assumption 3b is equivalent to the following excitation condition for some constants π_3, $\pi_4 > 0$:*

$$\pi_3 I \le \int_t^{t+T} e_\tau e_\tau^T \, d\tau \le \pi_4 I. \tag{28}$$

Similarly, if $t_{j+1} - t_j = T/L$ for all $j \in \{0, 1, 2, \cdots, L-1\}$, then Assumption 3a is equivalent to the existence of π_1, $\pi_2 > 0$ such that

$$\pi_1 I \le \int_t^{t+\frac{L-1}{L}T} (e_\tau - e_{\tau+\frac{T}{L}})(e_\tau - e_{\tau+\frac{T}{L}})^T d\tau \le \pi_2 I. \tag{29}$$

Note that for Assumption 3a, there should exists a subsequence $\{c_{j_k}\}_{k=1}^{m+1}$ whose difference $\{c_{j_k} - c_{j_{k+1}}\}_{k=1}^{m}$ is linearly independent. For this, $L \ge m+1$ is required. On the other hand, the existence of linearly independent subsequence $\{c_{j_k}\}_{k=1}^{m}$ suffices for e_τ to satisfy Assumption 3b. In this

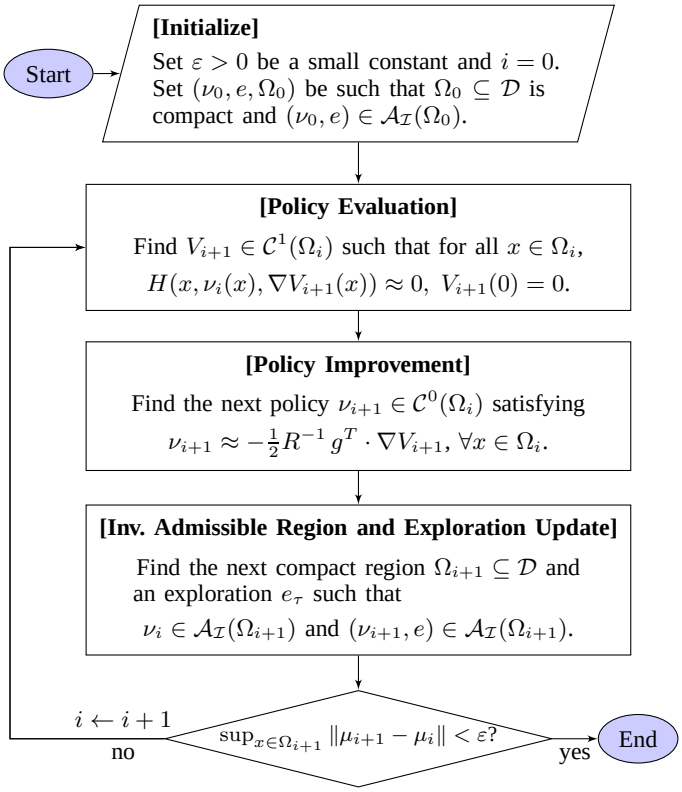


Fig. 2. General description of the proposed online I-RL methods.

case, we need $L \geq m$. Therefore, when V is given by V^μ *a priori*, the construction of the exploration e_τ becomes relatively simple and easy. For instance, if $m = 1$, then two constants $c_1 \neq c_2$ (*e.g.*, $c_1 = 1$ and $c_2 = 0$) are necessary to construct e_τ without violating Assumption 3a. On the contrary, for the case of $V = V^\mu$ and $m = 1$, e_τ satisfying Assumption 3b can be designed using only one constant $c_1 \neq 0$. Remember that 'Assumption 3a' or 'Assumption 3b with $V = V^\mu$' is required to guarantee the uniqueness of the solution (21), as stated in Theorems 4 and 5.

V. Main I-RL Algorithms with Explorations

Motivated by the advanced I-TDs (19), (22), and (26), we propose three partially/completely model-free I-RL algorithms that exploit the exploration e_τ to simultaneously excite the states and learn the next policy without knowing the nonlinear dynamics (f, g). While IA-PI is an off-line method, all the proposed I-RL algorithms can run in online fashion even when the nonlinear system (f, g) is partially/completely unknown and undergoes exploration e_τ. Fig. 2 describes the whole process of the proposed methods in a unified manner, which are similar to but different from IA-PI and I-PI described in Fig. 1 and Algorithm 1, as described below.

1) At each i-th policy evaluation and improvement steps of the proposed methods, the I-RL agent utilizes advanced I-TDs to find V_{i+1} and ν_{i+1} satisfying $V_{i+1} \approx V^{\nu_i}$ and $\nu_{i+1} \approx \nu_{i+1}^+$ on Ω_i, where ν_{i+1}^+ is given by

$$\nu_{i+1}^+ = -\frac{1}{2}R^{-1}g^T(x)\nabla V^{\nu_i}(x).$$

Algorithm 2. Explorized I-PI

Policy Evaluation: Given $(\nu_i, e) \in \mathcal{A}_{\mathcal{I}}(\Omega_i)$ and $z \in \Omega_i$, find $V_{i+1} \in \mathcal{C}^1(\Omega_i)$ such that

$$V_{i+1}(x_t) - V_{i+1}(x_{t+T}) = \int_t^{t+T} \left[r(x, \nu_i(x)) - \nabla^T V_{i+1}(x) \cdot g(x) e_\tau\right] d\tau \quad (30)$$

where $x \equiv x_\tau(z; \nu_i, e)$.

Policy Improvement: Update the next policy ν_{i+1} by

$$\nu_{i+1}(x) = -\frac{1}{2}R^{-1}\, g^T(x) \cdot \nabla V_i(x).$$

Note that all the proposed methods are equal to I-PI and IA-PI in the iteration domain, as long as the generated value functions and policies have no errors. That is, if $V_{i+1} = V^{\nu_i}$ and $\nu_{i+1} = \nu_{i+1}^+$, $\forall i \in \mathbb{Z}_+$ and $\nu_0 = \mu_0$, then, we have $V_{i+1} = V^{\mu_i}$ and $\nu_{i+1} = \mu_{i+1}$, $\forall i \in \mathbb{Z}_+$, where V^{μ_i} and μ_{i+1} are generated by IA-PI or I-PI.

2) While IA-PI and I-PI cannot explore the state-space in online fashion, the proposed methods use invariantly admissible explorations to simultaneously excite the state variables. So, to maintain invariant admissibility of the exploration, the proposed methods (re-)generate e_τ both at the initialization step and after each policy improvement, as shown in Fig. 2.

Note that the three I-RL methods are designed based on the respective advanced I-TDs (19), (22), and (26), which makes differences in policy evaluation and improvement steps. The other parts are exactly same to those presented in Fig. 2, so are omitted in the descriptions of the proposed I-RL methods (Algorithms 2–4).

A. Explorized I-PI

The first one is named as explorized I-PI whose policy evaluation and improvement are described in Algorithm 2. As can be seen from Algorithm 2, explorized I-PI comes from the advanced I-TD (22) and is able to simultaneously excite the states during policy evaluation by using the exploration e_τ. Unlike Algorithm 1, the advanced I-TD (30) contains the explorized term '$\nabla^T V_{i+1}(x) \cdot g(x) e_\tau$' to cancel out the effects of the exploration e. When $e \equiv 0$, explorized I-PI (Algorithm 2) becomes the I-PI described in Algorithm 1 and in Section III, provided that (30) holds for all $z = x_t \in \Omega_i$. In explorized I-PI, e_τ does not need to satisfy the excitation conditions such as those in Assumptions 3a and 3b as neither does the advanced I-TD (22). By Corollary 2, explorized I-PI guarantees uniqueness of the solution $V_i = V^{\nu_i}$ for any given exploration e_τ, and one just need to efficiently explore the state-space using e_τ without considering any excitation conditions *on* e_τ.

B. Integral Q-Learning I, II: Model-free I-RLs

The other two I-RL algorithms proposed in this paper are named integral Q-learning I and II, which are derived from

Algorithm 3. Integral Q-learning I

Policy Evaluation & Improvement: Given $(\nu_i, e) \in \mathcal{A}_{\mathcal{I}}(\Omega_i)$ and $z \in \Omega_i$, find $V_{i+1} \in \mathcal{C}^1(\Omega_i)$ and $\nu_{i+1} \in \mathcal{C}^0(\Omega_i)$ such that

$$V_{i+1}(x_t) - V_{i+1}(x_{t+T}) = \int_t^{t+T} \left[r(x, \nu_i(x)) + 2\nu_{i+1}^T(x) R e_\tau \right] d\tau \qquad (31)$$

where $x \equiv x_\tau(z; \nu_i, e)$.

Algorithm 4. Integral Q-learning II

Policy Evaluation: Given $\nu_i \in \mathcal{A}_{\mathcal{I}}(\Omega_i)$ and $z \in \Omega_i$, find $V_{i+1} \in \mathcal{C}^1(\Omega_i)$ such that

$$V_{i+1}(x_t) = \int_t^{t+T} r(x_\tau, \nu_i(x_\tau))\ d\tau + V_{i+1}(x_{t+T})$$

where $x_\tau \equiv x_\tau(z; \nu_i, 0)$.

Policy Improvement: Given $(\nu_i, e) \in \mathcal{A}_{\mathcal{I}}(\Omega_i)$, $z \in \Omega_i$, and $V_{i+1} \in \mathcal{C}^1(\Omega_i)$, find the next policy $\nu_{i+1} \in \mathcal{C}^0(\Omega_i)$ such that (31) holds, where $x \equiv x_\tau(z; \nu_i, e)$.

the advanced I-TDs (19) and (26), respectively, and can be implemented without knowing the system dynamics (f, g). In both algorithms, the exploration e_τ plays a central role in relaxing the requirement of the knowledge of $g(x)$.

Algorithm 3 describes policy evaluation and improvement of the proposed integral Q-learning I; as mentioned before, the other steps of the algorithms are exactly same to those in Fig. 2. In this method, the I-RL agent finds the solution $V_{i+1} \in \mathcal{C}^1(\Omega_i)$ and $\nu_{i+1} \in \mathcal{C}^0(\Omega_i)$ of the advanced I-TD (31) at the same time. On the contrary, integral Q-learning II illustrated in Algorithm 4 performs policy evaluation and improvement separately. In this second method, policy evaluation uses the zero exploration $e \equiv 0$, and is the same as that of Algorithm 2 under $e \equiv 0$; policy improvement of Algorithm 4 is developed from the advanced I-TD (26) to simultaneously explore the state space, and at the same time, to find the next policy ν_{i+1} satisfying (31) without using the knowledge of (f, g).

For simplicity, we assume in this paper that the exploration e applied to Algorithms 3 and 4 is given by (23) for some constant vectors $\{c_j\}_{j=1}^L$. In this case, for the uniqueness of the solution $V_{i+1} = V^{\nu_i}$ and $\nu_{i+1} = \nu_{i+1}^+$, the vectors $\{c_j\}_{j=1}^L$ should be carefully chosen so that they satisfy Assumption 3a for the first method and Assumption 3b for the second method. In general cases, (28) and (29) can be alternatives to Assumptions 3b and 3a, respectively. Although integral Q-learning II cannot simultaneously explore the state space in policy evaluation, the exploration in policy improvement can be designed in a simpler manner than the exploration in integral Q-learning I. This is because the construction of $\{c_j\}_{j=1}^L$ satisfying Assumption 3b is relatively easier and simpler than that of $\{c_j\}_{j=1}^L$ satisfying Assumption 3a, as mentioned in Section IV.

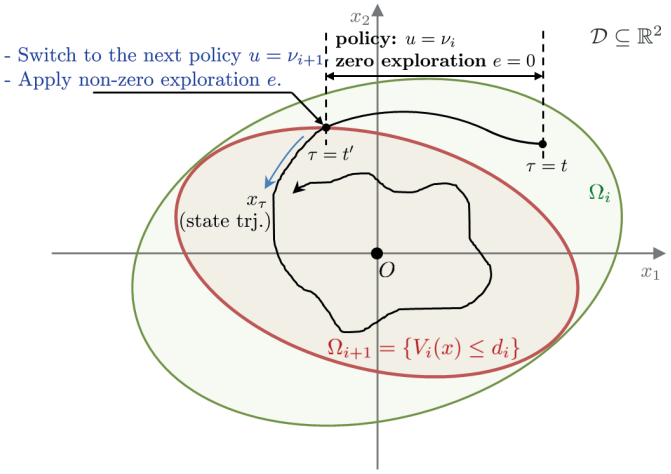


Fig. 3. Switching and exploration scheme when $\Omega_{i+1} \subseteq \Omega_i$ in $\mathbb{R}^2$.

C. Exploration, ISS, and Invariant Admissibility

Regarding the explorations applied to the proposed methods, we have focused on the excitation conditions to uniquely obtain $V_{i+1} \approx V^{\nu_i}$ and $\nu_{i+1} \approx \nu_{i+1}^+$ at each iteration. The required excitation condition for each proposed method can be summarized as follows: 1) **Explorized PI:** None, 2) **Integral Q-learning I:** Assumption 3a, and 3) **Integral Q-learning II:** Assumption 3b. In this subsection, we suppose that at i-th iteration, V_{i+1} and ν_{i+1} has no error, *i.e.*, $V_{i+1} = V^{\nu_i}$ and $\nu_{i+1} = \nu_{i+1}^+$ on Ω_i, and consider the next policy ν_{i+1} and next invariantly admissible region Ω_{i+1}. Let $\underline{\alpha}_{i+1}$ and $\bar{\alpha}_{i+1}$ be of class $\mathcal{K}$ satisfying

$$\underline{\alpha}_{i+1}(\|x\|) \le V_{i+1}(x) \le \bar{\alpha}_{i+1}(\|x\|).$$

In this ideal case, the proposed I-RL methods are equal to IA-PI and I-PI in the iteration domain, so under Assumption 2 with μ_i replaced by ν_i, the policies ν_i and ν_{i+1} are invariantly admissible on Ω_{i+1} by Theorem 1 and $\Omega_{i+1} = \Omega_{d_i}^{\nu_i}$. Moreover, Theorem 2 implies that (ν_{i+1}, e) is invariantly admissible on Ω_{i+1}, and that the system $\dot{x} = f + g(\nu_{i+1} + e)$ is input-to-state stable on Ω_{i+1} if the exploration e is bounded by

$$\sup_{t \le \tau < \infty} \|e(\tau)\| < \sqrt{\underline{\alpha}_s \circ \bar{\alpha}_{i+1}^{-1}(d_i)/\bar{\sigma}(R)}. \qquad (32)$$

If $\mathcal{D} = \mathbb{R}^n$ and $\underline{\alpha}_{i+1}$, $\underline{\alpha}_s \in \mathcal{K}_\infty$, then this ISS holds globally for any bounded exploration e_τ by Theorem 2. In case of that e is constructed from some constant vectors $\{c_j\}_{j=1}^N$ and satisfies (23), the boundedness condition (32) is replaced by

$$\|c_j\| < \sqrt{\underline{\alpha}_s \circ \bar{\alpha}_{i+1}^{-1}(d_i)/\bar{\sigma}(R)}, \quad \forall j \in \{1, 2, \cdots, N\}. \qquad (33)$$

Now, the remaining question is what to do when the state x is outside the invariantly admissible region Ω_{i+1} during online learning at $(i+1)$-th step. Note that using Corollary 1, Ω_{i+1} can be determined as $\Omega_{i+1} = \Omega_{d_i}^{\nu_i}$ under $V_{i+1} = V^{\nu_i}$ and $\nu_i \in \mathcal{A}(\Omega_i)$ by choosing d_i in the interval $(0, \min_{x \in \partial\Omega_i} V_{i+1}(x))$. In this case, we have $\Omega_{i+1} \subseteq \Omega_i$; by Theorem 2, $(\nu_{i+1}, e) \in \mathcal{A}_{\mathcal{I}}(\Omega_{i+1})$ for e satisfying (32). However, it is not guaranteed that (ν_{i+1}, e) is invariantly admissible *on* Ω_i, so e cannot be safely applied to the system $\dot{x} = f + g(\nu_{i+1} + e)$ when the state

x_τ lies in $\Omega_i \setminus \Omega_{i+1}$. In this particular case, the best way to preserve invariant admissibility and ISS is to apply the current policy $u = \nu_i$ and zero exploration $e \equiv 0$ until some finite time $t' \geq t$ at which x_τ enters into Ω_{i+1}, *i.e.*, $x_{t'} \in \Omega_{i+1}$.[1] Then, as illustrated in Fig. 3, the next policy $u = \nu_{i+1}$ and non-zero exploration e satisfying (32) or (33) can be applied thereafter without violating invariant admissibility and ISS on Ω_{i+1}. On the other hand, if one can find Ω_{i+1} that contains Ω_i and satisfies $\nu_i \in \mathcal{A}_\mathcal{I}(\Omega_{i+1})$, then the next policy ν_{i+1} and non-zero exploration e satisfying (32) can be applied any time since x_τ is already in Ω_{i+1}.

Note that in the global case ($\mathcal{D} = \mathbb{R}^n$ and $\underline{\alpha}_{i+1}, \underline{\alpha}_s \in \mathcal{K}_\infty$), the proposed I-RL algorithms can be performed without the initialization and regeneration of Ω_i and e. In the local case, such processes regarding Ω_i and e can be also removed when the exploration e is sufficiently small and x_τ starts from a region Ω near the origin that is small enough to be contained by *any* Ω_{i+1} satisfying $\nu_{i+1} \in \mathcal{A}_\mathcal{I}(\Omega_{i+1})$.

D. NN-Based LS Implementations

The proposed I-RL methods (Algorithms 3–5) can be implemented in the LS sense by using NNs to approximate V_{i+1} and ν_{i+1}. Let $\{\phi_j^c \in \mathcal{C}^1(\mathcal{D})\}_{j=1}^\infty$ and $\{\phi_j^a \in \mathcal{C}^0(\mathcal{D})\}_{j=1}^\infty$ be the sequences of real-valued NN activation functions that are linearly independent and complete on their respective function spaces $\mathcal{C}^1(\mathcal{D})$ and $\mathcal{C}^0(\mathcal{D})$. Here, the superscripts 'a' and 'c' denote actor and critic, respectively. Using these activation functions, $V_{i+1} \in \mathcal{C}^1$ and $\nu_{i+1} \in \mathcal{C}^0$ can be represented as

$$\begin{cases} V_{i+1}(x) = \sum_{j=1}^\infty w_{ij}\phi_j^c(x) \\ \nu_{i+1}(x) = \sum_{j=1}^\infty \mathbf{v}_{ij}\phi_j^a(x) \end{cases} \tag{34}$$

where $w_{ij} \in \mathbb{R}$ and $\mathbf{v}_{ij} \in \mathbb{R}^m$; we consider ($N_c$,$N_a$)-truncation of (34) as NN expressions of V_{i+1} and ν_{i+1}:

$$\begin{cases} \hat{V}_{i+1}(x) = \sum_{j=1}^{N_c} w_{ij}\phi_j^c(x) \equiv \mathbf{w}_i^T\boldsymbol{\phi}_c(x) \\ \hat{\nu}_{i+1}(x) = \sum_{j=1}^{N_a} \mathbf{v}_{ij}\phi_j^a(x) \equiv \mathbf{V}_i^T\boldsymbol{\phi}_a(x), \end{cases}$$

$$\text{where } \begin{cases} \mathbf{w}_i := [\, w_{i1}, w_{i2}, \cdots, w_{iN_c}\,]^T \in \mathbb{R}^{N_c} \\ \mathbf{V}_i := [\, \mathbf{v}_{i1}, \mathbf{v}_{i2}, \cdots, \mathbf{v}_{iN_a}\,]^T \in \mathbb{R}^{N_a \times m} \\ \boldsymbol{\phi}_c(x) := [\, \phi_1^c(x), \phi_2^c(x), \cdots, \phi_{N_c}^c(x)\,]^T \in \mathbb{R}^{N_c} \\ \boldsymbol{\phi}_a(x) := [\, \phi_1^a(x), \phi_2^a(x), \cdots, \phi_{N_a}^a(x)\,]^T \in \mathbb{R}^{N_a}. \end{cases}$$

Using these expressions, (34) can be rewritten as

$$\begin{cases} V_{i+1}(x) = \mathbf{w}_i^T\boldsymbol{\phi}_c(x) + \varepsilon_i^c(x) \\ \nu_{i+1}(x) = \mathbf{V}_i^T\boldsymbol{\phi}_a(x) + \varepsilon_i^a(x), \end{cases} \tag{35}$$

where $\varepsilon_i^c(x)$ and $\varepsilon_i^a(x)$ are NN reconstruction errors. Note that Assumption 1 implies $\mathcal{D}$ is compact. So, there exist $N_c, N_a \in \mathbb{N}$ such that the NN errors ε_i^c and ε_i^a in (35), and $\nabla\varepsilon_i^c$, are all bounded on the compact set $\mathcal{D}$ if $V_{i+1} \in \mathcal{C}^1$ and $\nu_{i+1} \in \mathcal{C}^0$ are finite on $\mathcal{D}$. This boundedness property also holds if the domain is restricted to a compact subset of $\mathcal{D}$ such as $\Omega_{d_i}^{\nu_i}$ in the proposed I-RL algorithms. Also note that since $V_{i+1}(0) = 0$ and $\nu_{i+1}(0) = 0$, we have $\boldsymbol{\phi}_c(0) = 0$ and $\boldsymbol{\phi}_a(0) = 0$ without loss of generality.

Now, consider integral Q-learning I (Algorithm 3) as an implementation example. In this case, substituting (35) into the advanced I-TD (31), we obtain

$$\begin{aligned} \delta_i(x_t, e) =& \big[\boldsymbol{\phi}_c(x_{t+T}) - \boldsymbol{\phi}_c(x_t)\big]^T\mathbf{w}_i \\ &+ \int_t^{t+T} \Big[r(x, \nu_i(x)) + 2\boldsymbol{\phi}_a^T(x)\mathbf{V}_i Re_\tau\Big] d\tau, \end{aligned} \tag{36}$$

where $\delta_i(x_t, e) \in \mathbb{R}$ is the advanced I-TD error given by

$$\delta_i(x_t, e) = \varepsilon_i^c(x_t) - \varepsilon_i^c(x_{t+T}) - 2\int_t^{t+T} (\varepsilon_i^a(x))^T Re_\tau \, d\tau.$$

Define $\mathbf{v}_i \in \mathbb{R}^{N_a m}$ as $\mathbf{v}_i := \mathrm{col}\{\mathbf{v}_{i1}, \mathbf{v}_{i2}, \cdots, \mathbf{v}_{im}\}$. Then, applying $\boldsymbol{\phi}_a^T(x)\mathbf{V}_i Re = \big(Re \otimes \boldsymbol{\phi}_a(x)\big)^T\mathbf{v}_i$ to (36) and then rearranging the equation, we obtain the following expression regarding (31):

$$\delta_i(x_t; e) = \boldsymbol{\psi}^T(x_t; e) \cdot \boldsymbol{\theta}_i + Z(x_t; \nu_i), \tag{37}$$

where $\boldsymbol{\theta}_i = \mathrm{col}\{\mathbf{w}_i, \mathbf{v}_i\}$ is the vector of unknown weights; $\boldsymbol{\psi}(x_t; e)$ and $Z(x_t; \nu_i)$ are given in Table I. The other advanced I-TDs in Algorithms 2 and 4 can be also formulated as (37) with $\boldsymbol{\theta}_i$, $\boldsymbol{\psi}(x_t; e)$, and $Z(x_t; \nu_i)$ given in Table I for each advanced I-TD. In Table I, the advanced I-TD errors $\delta_i(x_t; e)$ for Algorithms 2 and 4 were omitted, but can be easily obtained by the similar procedure. For policy evaluation of Algorithm 4, see [12]; in policy improvement of Algorithm 2, the next neuro-policy $\hat{\nu}_{i+1}$ can be updated by

$$\hat{\nu}_{i+1}(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla\boldsymbol{\phi}_c(x)\mathbf{w}_i \tag{38}$$

using $\hat{V}_{i+1}$, instead of V_{i+1}, as was done in [12].

Let N_θ be the number of elements of $\boldsymbol{\theta}_i$, *e.g.*, $N_\theta = N_c + N_a$ for (31). Then, we have N_θ unknowns in the 1-dimensional equation (37). In the implementations, $\boldsymbol{\theta}_i$ will be uniquely determined in LS sense. Define $\boldsymbol{\psi}[k]$, $\delta_i[k]$, and $Z[k]$ as

$$\begin{cases} \boldsymbol{\psi}[k] := \boldsymbol{\psi}(x_{t+kT}, e), \\ \delta_i[k] := \delta_i(x_{t+kT}, e), \\ Z[k] := Z(x_{t+kT}, \nu_i). \end{cases}$$

Then, referring $x_{t+(k-1)T}$ as a starting point of the advanced I-TDs, the following generalized I-TD error equation can be derived from (37):

$$\delta_i[k] = \boldsymbol{\psi}^T[k] \cdot \boldsymbol{\theta}_i + Z[k], \tag{39}$$

which holds for any $k \in \mathbb{N}$ since x_τ remains in the admissible region Ω_i for all $\tau \geq t$ by $(\nu_i, e) \in \mathcal{A}_\mathcal{I}(\Omega_i)$. Suppose the data $(\boldsymbol{\psi}[k], Z[k])$ for $k = 1, 2, \cdots, N$ are all available, and define the LS error E to be minimized as $E^2 := \frac{1}{2}\sum_{k=1}^N \delta_i^2[k]$. Then, differentiating E^2 in terms of $\boldsymbol{\theta}_i$ with the substitution of (39) yields

$$\frac{\partial E^2}{\partial \boldsymbol{\theta}_i} = \sum_{k=1}^N \frac{\partial \delta_i[k]}{\partial \boldsymbol{\theta}_i}\delta_i[k] = \sum_{k=1}^N \Big[\boldsymbol{\psi}[k]\boldsymbol{\psi}^T[k] \cdot \boldsymbol{\theta}_i + \boldsymbol{\psi}[k]Z[k]\Big].$$

[1] Since $\nu_i \in \mathcal{A}_\mathcal{I}(\Omega_i)$ implies asymptotic Lyapunov's stability, there exists finite time $t' \in [t, \infty)$ such that $x_\tau \equiv x_\tau(z; \nu_i, 0)$ for $z \in \Omega_i$ enters to the smaller set $\Omega_{d_i}^{\nu_i} \subseteq \Omega_i$ at t' under the zero exploration $e_\tau = 0$.

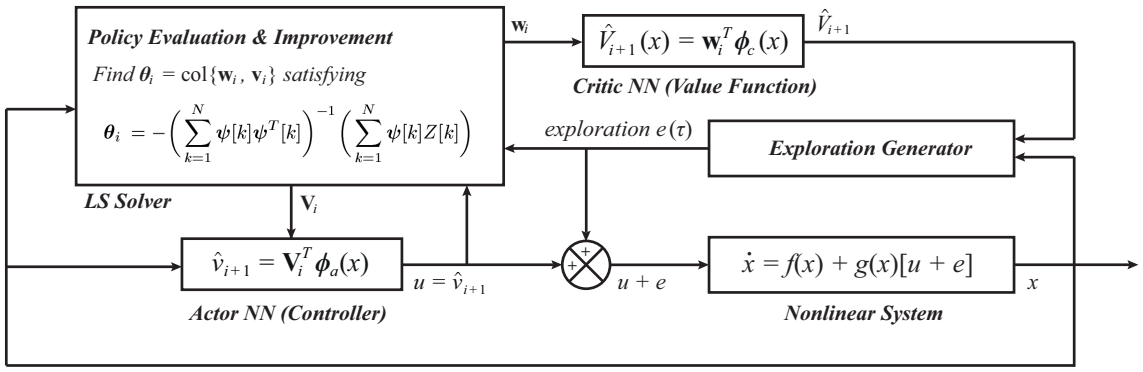


Fig. 4. The whole control scheme scheme with integral Q-learning I (Algorithm 3) and its LS implementation.

TABLE I
FUNCTIONS AND VECTORS OF EACH ADVANCED I-TD ERROR EQUATION (37) FOR NN IMPLEMENTATIONS OF ALGORITHMS 2–4

Algorithm No.	Process Type	$\boldsymbol{\theta}_i$	$\boldsymbol{\psi}(x_t; e)$	$Z(x_t; \nu_i)$
3	Policy Evaluation	$\mathbf{w}_i$	$\boldsymbol{\phi}_c(x_{t+T}) - \boldsymbol{\phi}_c(x_t) - \int_t^{t+T} \nabla^T \boldsymbol{\phi}_c(x) \cdot g(x) e_\tau \, d\tau$	$\int_t^{t+T} r(x, \nu_i(x)) \, d\tau$
4	Policy Evaluation & Improvement	$\text{col}\{\mathbf{w}_i, \mathbf{v}_i\}$	$\text{col}\left\{\boldsymbol{\phi}_c(x_{t+T}) - \boldsymbol{\phi}_c(x_t), \int_t^{t+T} 2\,\boldsymbol{\phi}_a(x) \otimes (Re_\tau) \, d\tau\right\}$	$\int_t^{t+T} r(x, \nu_i(x)) \, d\tau$
5	Policy Improvement	$\mathbf{v}_i$	$\int_t^{t+T} 2\,\boldsymbol{\phi}_a(x) \otimes (Re_\tau) \, d\tau$	$\hat{V}_i(x_{t+T}) - \hat{V}_i(x_t) + \int_t^{t+T} r(x, \nu_i(x)) \, d\tau$

Equating this to zero and rearranging the equation, we obtain the LS solution of the form

$$\boldsymbol{\theta}_{i,LS} = -\left(\sum_{k=1}^{N} \boldsymbol{\psi}[k]\boldsymbol{\psi}^T[k]\right)^{-1}\left(\sum_{k=1}^{N} \boldsymbol{\psi}[k]Z[k]\right). \quad (40)$$

For the existence of the unique LS solution $\boldsymbol{\theta}_{i,LS}$, we need the following excitation condition:

Assumption 4. *There exist κ_5, $\kappa_6 > 0$ such that*

$$\kappa_5 I \le \sum_{k=1}^{N} \boldsymbol{\psi}[k]\boldsymbol{\psi}^T[k] \le \kappa_6 I.$$

Note that the existence of the inverse in (40) is guaranteed by Assumption 4. Similar to Assumptions 3a and 3b, $N \ge L_\theta$ is necessary to satisfy Assumption 4, so at least L_θ-number of data should be collected to perform the LS (40) at each iteration.

The whole control scheme with integral Q-learning I and its LS implementation is demonstrated in Fig. 4. At each iteration, the LS solver collects the data needed to calculate $\boldsymbol{\psi}[k]$, $\delta_i[k]$, and $Z[k]$ for $k = 1, 2, \cdots, N$, and then finds the weight vectors $\mathbf{w}_i$ and $\mathbf{v}_i$ satisfying (40), both of which are transferred to the corresponding actor and critic NNs to update their weights. Here, the actor NN generates the control input; the output of the critic NN $\hat{V}_{i+1}(x)$ is used in the exploration generator module to calculate the bound (32) on the exploration e_τ. In exploration generator, the exploration e_τ is constructed, and modified if necessary, that plays a key role in exciting the signal $\boldsymbol{\psi}(x_t; e)$ in (40), and satisfies i) Assumption 3a (or (29)) and ii) the boundedness condition (32) for ISS and invariant admissibility. The whole control scheme with explorized I-PI or integral Q-learning II can be described in a similar manner by modifying LS solver and actor NN blocks.

E. An LQR Example: The Global Case

In the CT LQR case, the domain $\mathcal{D}$ becomes $\mathbb{R}^n$, $f(x) = Ax$, $g(x) = B$, and $S(x) = x^T S x$ for some matrices A, B, and $S > 0$ with compatible dimensions; the HJB equation (11) becomes the well-known algebraic Riccati equation

$$A^T P^* + P^* A - P^* B R^{-1} B^T P^* + S = 0,$$

for a positive definite matrix $P^* \in \mathbb{R}^{n \times n}$. The optimal value function and policy are given by

$$V^*(x) = x^T P^* x, \;\; \mu^*(x) = -K^* x,$$

where $K^* := R^{-1} B^T P^*$. By standard LQR theory, it is well-known that for any stabilizing linear policy $\mu(x) = -Kx$, $V^\mu(x)$ exists $\forall x \in \mathbb{R}^n$, and can be represented as $V^\mu(x) = x^T P^\mu x$, where $P^\mu \in \mathbb{R}^{n \times n}$ is positive definite. From this observation and the proposed nonlinear integral Q-learning I (Algorithm 3), we obtain the simplified integral Q-learning I for CT LQR, which is shown in Fig. 5 and is a modified version of the original integral Q-learning for CT LQR [13]. In this framework, V_{i+1} and the next policy ν_{i+1} are exactly parameterized as

$$V_{i+1}(x) = x^T P_{i+1} x, \;\; \nu_{i+1}(x) = -K_{i+1} x,$$

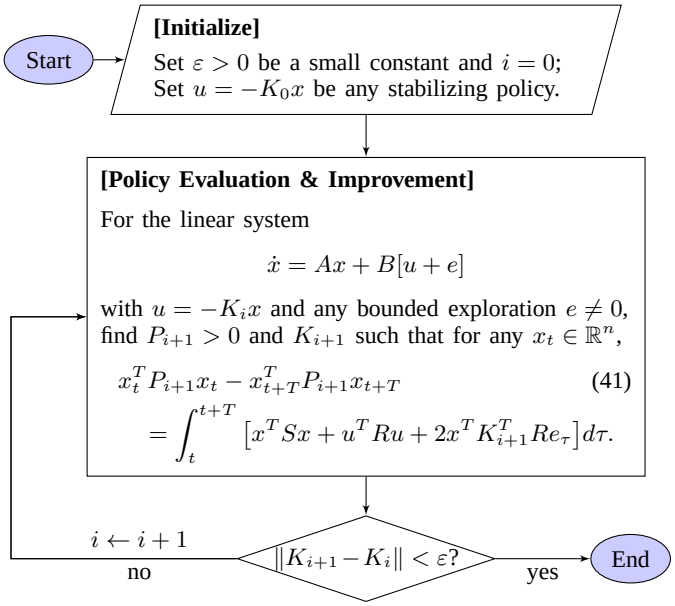


Fig. 5. Simplified integral Q-learning I for CT LQR.

with zero reconstruction errors $\varepsilon_i^c(x) = 0$ and $\varepsilon_i^a(x) = 0$ for all $x \in \mathbb{R}^n$. Hence, if the parameters P_{i+1} and K_{i+1} are uniquely determined by (40) that solves (41) in Fig. 5, then we have $\boldsymbol{\theta}_{i,LS} = \boldsymbol{\theta}_i$. In other words, Assumption 4 is sufficient to guarantee the uniqueness of the target solution $P_{i+1} = P^{\nu_i}$ and $K_{i+1} = R^{-1}B^T P^{\nu_i}$; the other excitation condition on e_τ such as Assumptions 3a or 3b are not required to obtain the unique solution. From this, we can see that under Assumption 4, ISS holds globally by Theorem 2 for any initial condition $z \in \mathbb{R}^n$ and any bounded exploration e since we have $\underline{\alpha}_{\nu_i}(x) = \underline{\sigma}(P^{\nu_i})\|x\|^2$ and $\underline{\alpha}_s(x) = \underline{\sigma}(S)\|x\|^2$, both of which are obviously of class $\mathcal{K}_\infty$. Therefore, in the LQR case, $\Omega_i = \mathbb{R}^n \ \forall i \in \mathbb{Z}_+$, so the invariantly admissible pair (Ω_{i+1}, e) does not need to be updated at each i-th iteration—any bounded non-zero exploration e guaranteeing Assumption 4 for updating (P_{i+1}, K_{i+1}) is sufficient to run the algorithm correctly, as shown in Fig. 5 and explained in this section. The other proposed I-RL algorithms (Algorithms 2 and 4) can be also simplified and analyzed in LQR frameworks in a similar manner to integral Q-learning I, and the analysis results in the same conclusions.

VI. Numerical Simulations

In this section, the proposed I-RL methods are simulated to verify the effectiveness of the proposed I-RL algorithms and the correctness of the relevant theories presented in this paper. In the simulations, we consider the nonlinear system

$$\begin{cases} \dot{x}_1 = -x_1 + x_2 \\ \dot{x}_2 = -(x_1 + x_2)/2 + x_2 h^2(x_1)/2 + h(x_1)u, \end{cases} \tag{42}$$

with a nonlinear $\mathcal{C}^1$-function $h(x_1)$, and the cost $r(x,u)$ with $S(x) = x_1^2 + x_2^2$ and $R = 1$. By the converse HJB approach [35], the optimal solution is given by

$$V^*(x) = x_1^2/2 + x_2^2 \text{ and } \mu^*(x) = -x_2 h(x_1).$$

In this case, $S(x)$ is quadratic, and the nonlinear system (42) can be approximated near the origin by a linear system. Therefore, by the standard LQR theory [26], [27], V^{μ_i} can be approximated by a quadratic function near the origin (see also [9, Remark 3.1.8]); we choose $\mathbf{w}_i$ and $\boldsymbol{\phi}_c(x)$ in the critic NN $\hat{V}_{i+1}(x) = \mathbf{w}_i^T \boldsymbol{\psi}_c(x)$ as

$$\mathbf{w}_i = [\, w_{i1},\ w_{i2},\ w_{i3}\,]^T, \ \ \boldsymbol{\phi}_c(x) = [\, x_1^2,\ x_1 x_2,\ x_2^2\,]^T. \tag{43}$$

To determine the appropriate actor NN structure, we substitute (43), $R = 1$, and $g(x) = [0,\ h(x_1)]^T$ to (38), which results in

$$\hat{\nu}_{i+1}(x) = -\frac{1}{2} w_{i2} \cdot x_1 h(x_1) - w_{i3} \cdot x_2 h(x_1). \tag{44}$$

Now, assume that $h(x_1)$ can be represented as

$$h(x_1) = \vartheta^T \psi(x_1) + \varepsilon_h(x_1), \tag{45}$$

with the weight vector $\vartheta \in \mathbb{R}^M$, the nonlinear regression function $\psi(x_1) \in \mathbb{R}^M$, and the bounded approximation error $\varepsilon_h(x_1)$; M is the number of weights. Then, (44) becomes

$$\hat{\nu}_{i+1}(x) = (\mathbf{v}_i^+)^T \boldsymbol{\rho}(x) + \bar{\varepsilon}_h(x),$$

where $\mathbf{v}_i^+$, $\boldsymbol{\rho}(x) \in \mathbb{R}^{2M}$ are defined as

$$\mathbf{v}_i^+ := -[\, w_{i2}/2,\ w_{i3}\,]^T \otimes \vartheta \text{ and } \boldsymbol{\rho}(x) := x \otimes \psi(x_1), \tag{46}$$

and $\bar{\varepsilon}_h(x)$ is given by $\bar{\varepsilon}_h(x) = -(w_{i2}x_1/2 + w_{i3}x_2)\varepsilon_h(x_1)$, which is obviously bounded in a compact set. From this result, we choose $\mathbf{v}_i$ and $\boldsymbol{\phi}_a(x)$ in the actor NN $\hat{\nu}_{i+1} = \mathbf{v}_i^T \boldsymbol{\phi}_a(x)$ as

$$\mathbf{v}_i = [\, v_{i1},\ v_{i2},\ \cdots,\ v_{i2M}\,]^T \text{ and } \boldsymbol{\phi}_a(x) = \boldsymbol{\rho}(x).$$

Note that the actor NN is used in integral Q-learning I and II to find the next policy $\hat{\nu}_{i+1}$ with $\mathbf{v}_i \approx \mathbf{v}_i^+$ when $h(x_1)$ (or ϑ) is not known. If $h(x_1)$ is perfectly known, then (38) can be used to directly compute the next policy $\hat{\nu}_{i+1}$ as was done in policy improvement of explorized I-PI. At each i-th step, $\mathbf{w}_i$ and/or $\mathbf{v}_i$ will be updated by the LS solution (40) after N data samples $(\boldsymbol{\psi}[k], Z[k])$ $(k = 1, 2, \cdots, N)$ are collected. In the simulations, the objective of the proposed I-RL methods is to find the optimal weight vectors $\mathbf{w}^*$ and/or $\mathbf{v}^*$ given by

$$\mathbf{w}^* = [\, 1/2,\ 0,\ 1\,]^T \text{ and } \mathbf{v}^* = \mathbf{v}_i^+|_{\mathbf{w}_i = \mathbf{w}^*} = -[\, 0^T,\ \ \vartheta^T\,]^T.$$

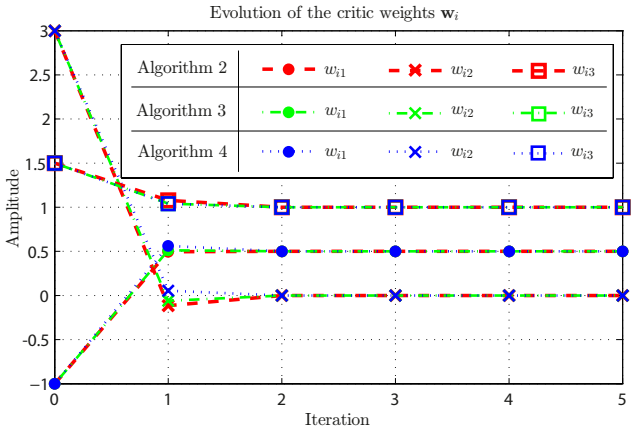
A. Simulation Example 1

For the comparison with I-PI (Algorithm 1) in [12], we first consider the nonlinear system (42) with $h(x_1) = \sin(x_1)$, which can be represented by (45) with $\vartheta = 1$, $\psi(x_1) = \sin x_1$ and $\varepsilon_h(x_1) \equiv 0$. In this case, $\mathbf{v}_i$ and $\boldsymbol{\phi}_a(x)$ are given by

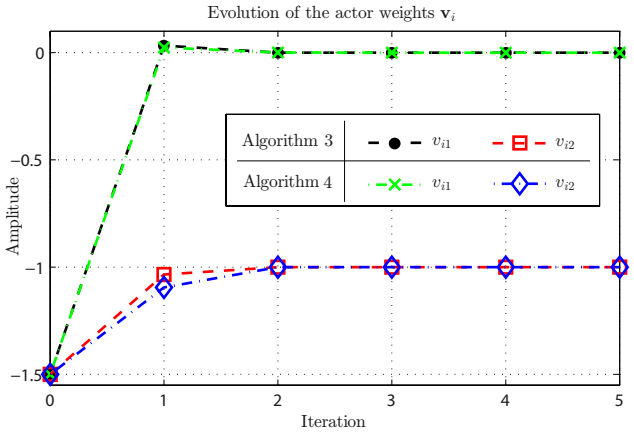
$$\mathbf{v}_i = [\, v_{i1},\ v_{i2}\,]^T \text{ and } \boldsymbol{\phi}_a(x) = [\, x_1 \sin x_1,\ x_2 \sin x_1\,]^T.$$

TABLE II
(Example 1) The Number of Collected Data per Iteration

Process Type & Algorithm No.	N	N_θ
Policy Evaluation of Algorithm 2	30	3
Policy Evaluation & Improvement of Algorithm 3	50	5
Policy Evaluation of Algorithm 4	30	3
Policy Improvement of Algorithm 4	20	2



(a) Critic NN weights $\mathbf{w}_i$



(b) Actor NN weights $\mathbf{v}_i$

Fig. 6. **(Example 1)** Evolution of critic/actor weights for Algorithms 3–5.

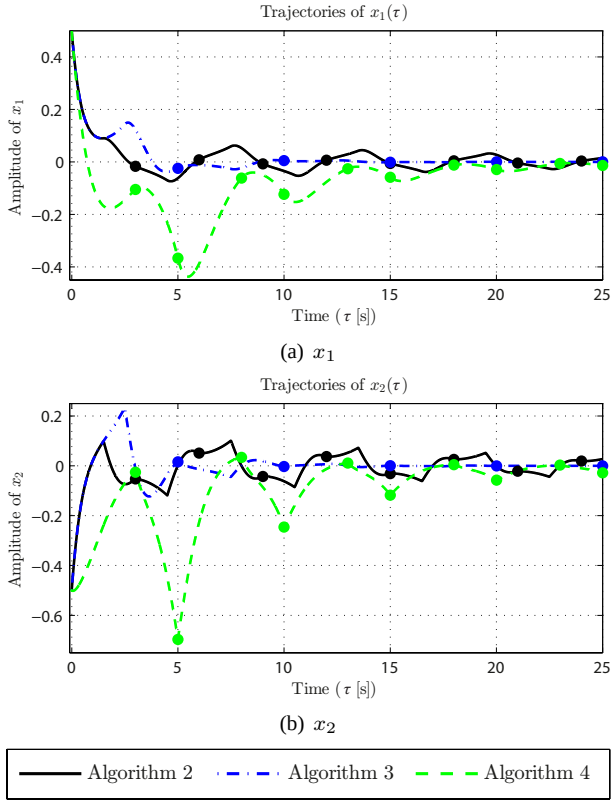


Fig. 7. **(Example 1)** Trajectories of the state variables (a) $x_1(\tau)$ and (b) $x_2(\tau)$ for Algorithms 2–4. The marked points indicate the time instants the critic and/or actor NN weights are updated.

As in [12, Section 6.1], the initial admissible policy is given by $\mu_0(x) = -\frac{3}{2}\sin(x_1)(x_1 + x_2)$; the initial state x_0 at $t = 0$ and the sampling period T are set to $x_0 = [0.5, -0.5]^T$ and $T = 0.1$ [s], respectively; the number of data, N, collected per iteration is determined by '$N = 10 \times N_\theta$' for each process, as demonstrated in Table II. Note that all of these settings correspond to the simulation in [12, Section 6.1], where $N = 30$ was used under $N_\theta = 3$ as well. In the simulations of Algorithms 2 and 3, we used e_τ given by

$$e_\tau = \begin{cases} c & \text{for all } \tau \in [t, t + NT/2), \\ -c & \text{for all } \tau \in [t + NT/2, t + NT), \end{cases} \tag{47}$$

with $c = 2.5$. In policy improvement of Algorithm 4, we used

$$e_\tau = 3.5 \quad \text{for all } \tau \in [t, t + NT). \tag{48}$$

Notice that these explorations (47) and (48) satisfy Assumptions 3a and 3b, respectively, if one considers the extended time interval $[t, t+NT)$, instead of $[t, t+T)$. The update step of (Ω_i, e) is omitted in these simulations for simplicity (see [10] for an example of updating Ω_i in IA-PI), so the same e will be applied for all iteration steps.

Fig. 6 demonstrates the simulation results in the iteration domain—(a) the variations of $\mathbf{w}_i$ generated by the proposed I-RL methods (Algorithms 3–5), and (b) the evolution of $\mathbf{v}_i$ generated by model-free integral Q-learning I and II (Algorithms 4 and 5). As shown in Fig. 6(a), all the critic weights $\mathbf{w}_i$ at each iteration are very close to those in [12, Section 6.1], showing the equivalences of all the proposed I-RL methods and I-PI in the iteration domain under the uniqueness condition. The actor NN weights $\mathbf{v}_i$ in Fig. 6(b) are also close to each other at each iteration. Moreover, as shown in Fig. 6, both weights $\mathbf{w}_i$ and $\mathbf{v}_i$ converge to their optimal values within a few iterations, which is due to *the second-order convergence nature of I-PI and PI in the iteration domain* [13], [17], [21], [36]. Fig. 7 shows the state trajectories, which are all bounded but oscillatory due to the exploration e applied for both the excitation of $\boldsymbol{\psi}[k]$ and online learning in partially/completely unknown dynamics (f, g). In I-PI [12, Section 6.1], $g(x)$ should be known and there is no way to re-excite the signal because of the absence of the exploration e.

B. Simulation Example 2

In this example, integral Q-learning I is applied to the nonlinear system (42) with $h(x_1) = \cos(2x_1) + 2$. This system was also used in [22], [24] to simulate their synchronous actor-critic learning methods. As was done in [22], the weight vector $\mathbf{w}_i$ of the critic NN is initialized to $\mathbf{w}_0 = [1, 1, 1]^T$; the corresponding initial actor NN weight vector is given by $\mathbf{v}_0 = \mathbf{v}_0^+ = -[1/2, 1] \otimes \vartheta$. The states are initialized to zero, *i.e*, $x_0 = [0, 0]^T$; we set $N = 40$ and $T = 25$ [ms], so the LS solution (40) is calculated every 1 [s]. The exploration scheme (47) with $c = 1$ is used to hold Assumption 3a.

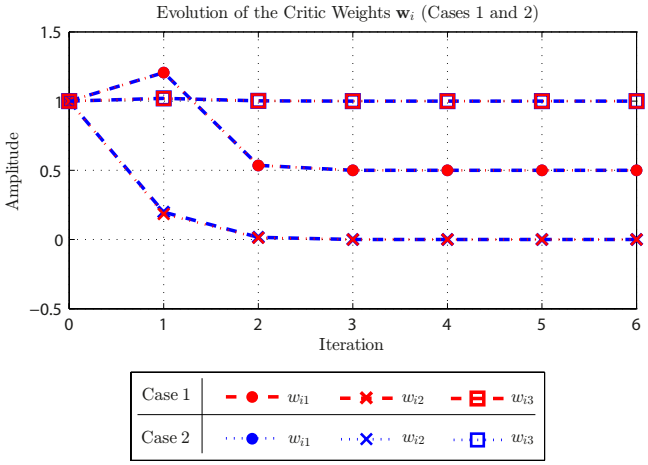


Fig. 8. **(Example 2) Cases 1 and 2:** Evolution of critic weights $\mathbf{w}_i$ generated by integral Q-learning I.

Case 1: first, we assume $h(x_1)$ is linearly parameterized as $h(x_1) = \vartheta^T \psi(x_1)$ with zero approximation error $\varepsilon_h(x) \equiv 0$, where ϑ and $\psi(x_1)$ are given by

$$\vartheta = [\,1,\ \ 2\,]^T \text{ and } \psi(x_1) = [\,\cos(2x_1),\ \ 1\,]^T.$$

From this, $\mathbf{v}_i$ and $\boldsymbol{\phi}_a$ in the actor NN are determined as $\mathbf{v}_i \in \mathbb{R}^4$ and $\boldsymbol{\phi}_a(x) = [\,x_1 \cos(2x_1),\ x_1,\ x_2 \cos(2x_1),\ x_2\,]^T$, and the optimal actor NN weights $\mathbf{v}^*$ are given by

$$\mathbf{v}^* = -[\,0^T,\ \ \vartheta^T\,]^T = [0,\ 0,\ 1,\ 2]^T.$$

The simulation results are shown in Figs. 8 and 9. As shown in Figs. 8 and 9(a), the weights in the critic and actor NNs converge to their respective optimal ones within 3 iterations; Fig. 9(b) illustrates the optimal policy approximation error at $i = 6$, which is very small ($\leq 10^{-10}$), showing the effectiveness of the proposed integral Q-learning I.

Case 2: next, assume $h(x_1)$ in $g(x)$ is completely unknown and consider its expression (45) with $M = 7$ and $\psi(x)$ given by $\psi(x_1) = [1\ x_1\ x_1^2\ x_1^3\ x_1^4\ x_1^5\ x_1^6]^T$. From this and (46), the activation function $\boldsymbol{\phi}_a(x) \in \mathbb{R}^{14}$ of the actor NN can be determined as $\phi_j^a(x) = x_1^j$ for $1 \leq j \leq 7$ and $\phi_j^a(x) = x_1^{j-8} x_2$ for $8 \leq j \leq 14$. The unknown vector ϑ in (45) and in the optimal actor NN weight $\mathbf{v}^*$ can be also obtained as $\vartheta = [3,\ 0,\ -2,\ 0,\ 2/3,\ 0,\ -4/45]^T$ from the Taylor expansion of $h(x_1)$ at $x_1 = 0$:

$$h(x_1) = \cos(2x_1) + 2 = 3 - 2x_1^2 + \frac{2x_1^4}{3} - \frac{4x_1^6}{45} + \mathcal{O}(x_1^8).$$

The simulation results in Case 2 are demonstrated in Figs. 8 and 10. As can be seen from Fig. 8, the critic NN weights at each iteration are almost equal to those in Case 1, which converge to the optimal ones. The final critic NN weights $\mathbf{w}_i$ at $i = 6$ in Case 2 is $\mathbf{w}_6 = [0.5000,\ 0.0000,\ 1.0000]^T$ and Fig. 10(a) shows the corresponding approximation error of the optimal value function. Here, the maximum error is smaller than 10^{-8}, showing the good performance of integral Q-learning I. The optimal and final weights $\mathbf{v}^*$ and $\mathbf{v}_i$ at $i = 6$ of the actor NN are shown in Table III. Though the error $\|\mathbf{v}_{6j} - \mathbf{v}_6^*\|$ for each j is very small, as shown in Table III, the approximation errors of the optimal policy shown in Fig. 10(b) is relatively high compared with those in Fig. 9(b). These errors are due to the approximation of the unknown $h(x_1)$, but can be decreased by incorporating higher-order terms like x_1^9 and $x_2 x_1^8$ into the actor NN.

Discussions: As opposed to the synchronous actor-critic methods [22], [24], where the actor NN with the same structure to the critic NN was introduced for the closed-loop stability, the proposed integral Q-learning methods introduce the actor NN for model-free learning, and as shown in this example, its structure is determined by appropriate procedures when $g(x)$ contains some structural/parametric uncertainties. This makes it possible to learn the online optimal control solution without knowing the nonlinear dynamics (f, g) and without introducing the complex identifier NN structure (see [24] for actor-critic-identifier architecture).

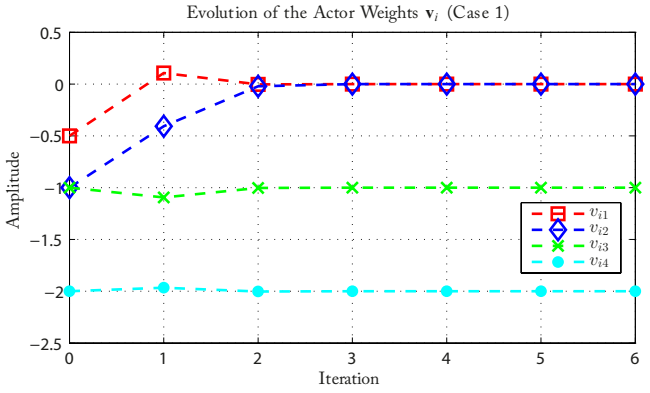
As shown in Figs. 8 and 9(a), the NN weights almost converge at 2 [s], showing that the convergence time is similar to the actor-critic-identifier method in [24], and faster than the model-based actor-critic method in [22]. On the other hand, when $h(x_1)$ is linearly parameterized (Case 1), the approximation error of the optimal control policy is far smaller than the actor-critic methods [22], [24] as shown in Fig. 9(b). These fast, accurate convergence results are mainly due to *the second-order convergence nature of PI and I-PI methods in the iteration domain* [13], [17], [21], [36]. As shown in Fig. 10(a), the approximation error of the optimal value function is also very small even when $h(x_1)$ is completely unknown (Case 2). In this case, however, the approximation error of the optimal control policy is relatively large compared to the actor-critic methods [22], [24]. This is due to the approximation error $\bar{\varepsilon}_h(x)$, which can be made sufficiently small by increasing the number of neurons. Though the proposed I-RL methods show the good convergence properties, they require an initial admissible policy, while the synchronous methods do not [22], [24]. This restriction can be relaxed if the I-RL methods are developed based on VI [21] or generalized PI [17], [21], rather than PI, which is the future work of this paper.

VII. Conclusions

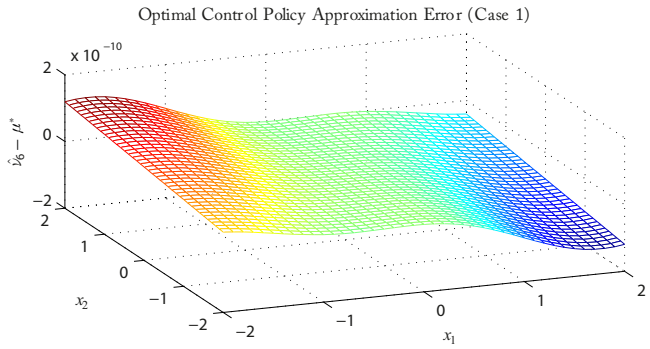
In this paper, we proposed one partially model-free I-RL method named explorized PI and two completely model-free I-RL methods named integral Q-learning I and II, the objective of all of which is to find the online solution to the given CT nonlinear optimal control problem with input-affine dynamics. All the proposed methods are able to simultaneously and stably explore the state-space during the learning phase. To develop the methods, the concepts of exploration, I-TD, and invariant

TABLE III
(Example 2) Case 2: The actor NN weights $\mathbf{v}_6$ and $\mathbf{v}^*$

j	v_{6j}	v_j^*	j	v_{6j}	v_j^*	j	v_{6j}	v_j^*
1	0.0000	0	6	0.0000	0	11	0.0000	0
2	0.0000	0	7	0.0001	0	12	-0.6667	-2/3
3	0.0000	0	8	-3.0000	-3	13	0.0000	0
4	0.0000	0	9	0.0000	0	14	0.0889	4/45
5	0.0000	0	10	2.0000	2			

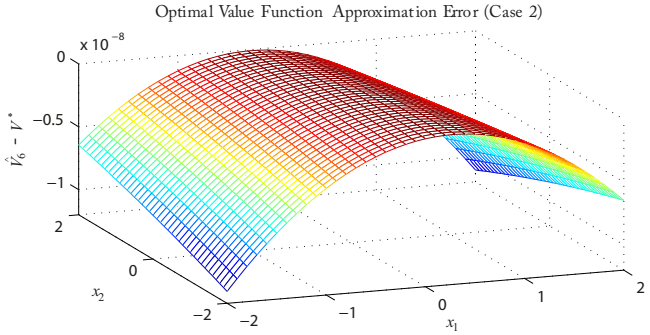


(a) Evolution of actor weights $\mathbf{v}_i$

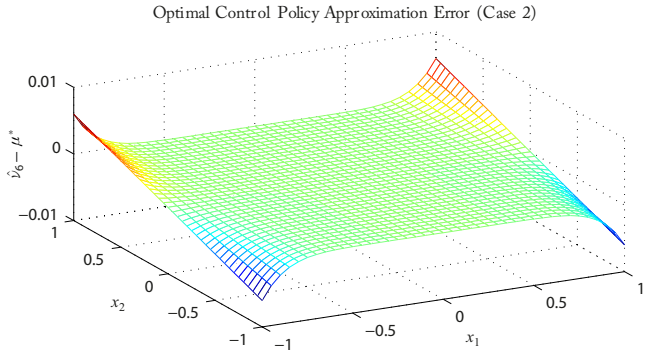


(b) Optimal policy approximation error

Fig. 9. **(Example 2) Case 1:** Simulation results when the policy is exactly parameterized—(a) evolution of the actor NN weights $\mathbf{v}_i$ and (b) optimal control policy approximation error



(a) Optimal value function approximation error



(b) Optimal policy approximation error

Fig. 10. **(Example 2) Case 2:** Approximation errors of the online optimal solution (a) $\hat{V}_6(x)$ and (b) $\hat{\nu}_6(x)$ when $\varepsilon_h(x_1)$ is not identically zero

admissibility were all extended to the CT nonlinear input-affine dynamical system governed by a control policy and an exploration, and then analyzed in details in connection to the proposed I-RL methods, I-PI, and IA-PI. We have also shown that the proposed I-RL methods are all equal to I-PI and IA-PI in the iteration domain, under the given excitation condition on the exploration for the uniqueness of the I-TD solutions. As a result, ISS, invariant admissibility, and convergence properties of the proposed I-RL methods were all given as well, and discussed in details, under the well-designed explorations satisfying the required excitation and boundedness conditions. The NN-based implementation methods of the proposed methods were also presented, and their performance was verified by numerical simulations and compared with the other methods.

REFERENCES

[1] 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, 2009.

[2] J. Si, A. G. Barto, W. B. Powell, and D. Wunsch, *Handbook of learning and approximate dynamic programming*. Wiley-IEEE Press, 2004.

[3] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge Univ Press, 1998.

[4] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. Wiley-Interscience, 2007.

[5] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, 1992.

[6] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.

[7] G. N. Saridis and C. S. G. Lee, "An approximation theory of optimal control for trainable manipulators," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 3, pp. 152–159, 1979.

[8] R. W. Beard, G. N. Saridis, and J. T. Wen, "Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation," *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.

[9] R. W. Beard, *Improving the closed-loop performance of nonlinear systems*. PhD thesis, Rensselaer Polytechnic Institute, 1995.

[10] J. Y. Lee, J. B. Park, and Y. H. Choi, "Invariantly admissible policy iteration for a class of nonlinear optimal control problems," *Submitted to Syst. Contol Lett. (available at http://arxiv.org/abs/1402.4187)*, 2014.

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

[12] D. Vrabie and F. L. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Netw.*, vol. 22, no. 3, pp. 237–246, 2009.

[13] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica*, vol. 48, no. 11, pp. 2850–2859, 2012.

[14] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning with explorations for continuous-time nonlinear systems," in *Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1042–1047, 2012.

[15] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, no. 10, pp. 2699–2704, 2012.

[16] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *American Control Conference (ACC)*, vol. 3, pp. 3475–3479, 1994.

[17] J. Y. Lee, J. B. Park, and Y. H. Choi, "On integral generalized policy iteration for continuous-time linear quadratic regulations," *Automatica*, vol. 50, no. 2, pp. 475–489, 2014.

[18] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H_∞ control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[19] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches*, pp. 493–525, 1992.

[20] D. V. Prokhorov and D. C. Wunsch II, "Adaptive critic designs," *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 997–1007, 1997.
[21] D. Vrabie, *Online Adaptive Optimal Control For Continuous-time Systems*. Ph. D. Thesis, TX, USA: University of Texas at Arlington, 2010.
[22] 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.
[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*, 2013.
[24] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, 2013.
[25] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, 2013.
[26] F. L. Lewis and V. L. Syrmos, *Optimal control*. John Wiley, 1995.
[27] D. E. Kirk, *Optimal control theory: an introduction*. Dover Pubns, 2004.
[28] R. Bellman, *Dynamic Programming*. NJ: Princeton Univ., 1957.
[29] F. O.-Tellez, E. N. Sanchez, and A. G. Loukianov, "Discrete-time neural inverse optimal control for nonlinear systems via passivation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1327–1339, 2012.
[30] L. J. Ricalde and E. N. Sanchez, "Inverse optimal neural control of a class of nonlinear systems with constrained inputs for trajectory tracking," *Optim. Control Appl. Methods*, vol. 33, no. 2, pp. 176–198, 2012.
[31] M. Fu and B. Barmish, "Adaptive stabilization of linear systems via switching control," *IEEE Trans. Autom. Control*, vol. 31, no. 12, pp. 1097–1103, 1986.
[32] E. B. Kosmatopoulos, "Control of unknown nonlinear systems with efficient transient performance using concurrent exploitation and exploration," *IEEE Trans. Neural Netw.*, vol. 21, no. 8, pp. 1245–1261, 2010.
[33] P. A. Ioannou and J. Sun, *Stable and robust adaptive control*. 1995.
[34] H. K. Khalil, *Nonlinear systems*. Prentice Hall, 2002.
[35] V. Nevistić and J. A. Primbs, "Constrained nonlinear optimal control: a converse HJB approach," *Technical report 96-021*, 1996.
[36] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Trans. Autom. Cont.*, vol. 13, no. 1, pp. 114–115, 1968.

APPENDIX A: PROOF OF LEMMA 1

The proof will be done by contradiction. For $\mu(x) \in \mathcal{A}(\Omega)$, assume that there exists another function $V \in \mathcal{C}^1(\Omega)$ that is positive definite and satisfies the Lyapunov equation

$$H(x, \mu(x), \nabla V(x)) = 0 \quad \forall x \in \Omega, \quad V(0) = 0. \tag{49}$$

From (49), $r(x,u) > 0$, and the definition of H, we have $(\nabla V(x))^T(f(x)+g(x)\mu(x)) < 0, \forall x \in \Omega \setminus \{0\}$, which again implies $\nabla V(x) \neq 0$ and $f(x) + g(x)\mu(x) \neq 0 \ \forall x \in \Omega \setminus \{0\}$. Subtracting (49) from (8) yields

$$\begin{aligned} H(x,&\mu(x), \nabla V) - H(x, \mu(x), \nabla V^\mu) \\ &= [\nabla V(x) - \nabla V^\mu(x)]^T \cdot (f(x) + g(x)\mu(x)) = 0, \end{aligned}$$

which holds $\forall x \in \Omega$. Therefore, we obtain $V(x) = V^\mu(x) + c$ for a constant c, $\forall x \in \Omega \setminus \{0\}$ since $f(x) + g(x)\mu(x) \neq 0$ $\forall x \in \Omega \setminus \{0\}$. Here, $V(0) = V^\mu(0) = 0$ results in $c = 0$ and thereby, $V(x) = V^\mu(x)$ is obtained for all $x \in \Omega$, a contradiction. Therefore, the value function V^μ is the unique solution of (8) over $\mathcal{C}^1(\Omega)$, the completion of the proof.

APPENDIX B: PROOF OF THEOREM 2

Since $\mu_0 \in \mathcal{A}(\Omega_0)$, Assumption 2 and Theorem 1 imply that for any $i \in \mathbb{Z}_+$, Ω_{i+1} is in the interior of $\mathcal{D}$, $V^{\mu_i} \in \mathcal{C}^1(\Omega_{i+1})$, and μ_i and μ_{i+1} are invariantly admissible on Ω_{i+1}. So, μ_i and V^{μ_i} satisfy (8) with $\mu = \mu_i$, *i.e.*, for all $x \in \Omega_{i+1}$,

$$(\nabla V^{\mu_i}(x))^T(f(x) + g(x)\mu_i(x)) = -r(x, \mu_i(x)). \tag{50}$$

Then, differentiating $V^{\mu_i}(x)$ along the trajectory $x(z; \mu_{i+1}, e)$ and substituting (12) and (50) yield

$$\begin{aligned} \dot{V}^{\mu_i}(x) &= (\nabla V^{\mu_i}(x))^T(f(x) + g(x)[\mu_{i+1}(x) + e]) \\ &= -r(x, \mu_i) - 2\mu_{i+1}^T Re - 2\mu_{i+1}^T R(\mu_{i+1} - \mu_i). \end{aligned}$$

Applying Young's inequality $2x^TRy \leq x^TRx + y^TRy$ for $x, y \in \mathbb{R}^m$, we obtain

$$\dot{V}^{\mu_i}(x) \leq -S(x) + e^TRe, \tag{51}$$

which can be further expanded using $\underline{\alpha}_s(\|x\|) \leq S(x)$ and $V^{\mu_i}(x) \leq \bar{\alpha}_{\mu_i}(\|x\|)$ (see (4) and (5)) as

$$\dot{V}^{\mu_i} \leq -(1-\theta)S(x) - \theta\hat{\alpha}\big(V^{\mu_i}(x)\big) + \bar{\sigma}(R) \cdot \left(\sup_{t \leq \tau < \infty} \|e_\tau\|^2 \right),$$

where $\hat{\alpha} := \underline{\alpha}_s \circ \bar{\alpha}_{\mu_i}^{-1}$ is of class $\mathcal{K}$ and defined on the interval $[0, \bar{\alpha}_{\mu_i}(r_d)]$ by Assumption 1 and [34, Lemma 4.2]; $\theta \in (0,1)$ is a constant satisfying

$$\sup_{t \leq \tau < \infty} \|e_\tau\|^2 < \theta \cdot \hat{\alpha}\big(d_i\big)/\bar{\sigma}(R), \tag{52}$$

Since we assume the exploration e satisfies (18), such θ always exists in $(0,1)$. Therefore, we have

$$\dot{V}^{\mu_i}(x) \leq -(1-\theta)S(x), \tag{53}$$

for all $x \in \Omega_{i+1}$ satisfying $V^{\mu_i}(x) \geq r_i$, where r_i is given by

$$r_i \equiv \hat{\alpha}^{-1}\left(\bar{\sigma}(R) \cdot \frac{\big(\sup_{t \leq \tau < \infty} \|e_\tau\|\big)^2}{\theta} \right). \tag{54}$$

Now, substituting (52) into (54) and rearranging the equation yields $r_i < d_i$. Hence, noting that $\Omega_{i+1} = \Omega_{d_i}^{\mu_i}$ by Assumption 2, we can conclude that

$$\Omega_{r_i}^{\mu_i} = \{x \in \mathcal{D} : V_i^\mu(x) \leq r_i\}$$

is in the interior of Ω_{i+1}, and (53) holds for all $x \in \Omega_{i+1} \setminus \Omega_{r_i}^{\mu_i}$. This implies $\dot{V}^{\mu_i}$ is negative definite on the boundary $\partial\Omega_{i+1}$, so $x_\tau(z; \mu_{i+1}, e)$ starting in $z \in \Omega_{i+1}$ stays in Ω_{i+1} for all τ. That is, $e \in \mathcal{A}_\mathcal{I}(\Omega_{i+1}; \mu_{i+1})$.

Next, Assumption 1 and Theorem 1 imply that Ω_{i+1} is in the interior of $B_0(r_d)$, so we have $\Omega_{r_i}^{\mu_i} \subset \Omega_{i+1} \subset B_0(r_d)$. Applying (4) and (5) to (53) to prove ISS, we obtain

$$\begin{aligned} \dot{V}^{\mu_i}(x_\tau) &\leq -(1-\theta)\hat{\alpha}(V^{\mu_i}(x_\tau)) && (55) \\ &\leq -(1-\theta)\hat{\alpha}(r_i) \equiv -k < 0 && (56) \end{aligned}$$

for all $x_\tau \in \Omega_{i+1} \setminus \Omega_{r_i}^{\mu_i}$, Hence, (56) and $e \in \mathcal{A}_\mathcal{I}(\Omega_{i+1}; \mu_{i+1})$ imply that for any $z \in \Omega_{i+1} \setminus \Omega_{r_i}^{\mu_i}$, there is $t' > t$ such that

$$\begin{cases} x_\tau(z, \mu_{i+1}, e) \in \Omega_{i+1} \setminus \Omega_{r_i}^{\mu_i} \text{ for all } \tau \in [t, t+t'), \\ x_\tau(z, \mu_{i+1}, e) \in \Omega_{r_i}^{\mu_i} \text{ for all } \tau \geq t + t'. \end{cases}$$

Assume $\hat{\alpha}$ is locally Lipshitz without loss of generality[2] and let v_τ be the solution to the scalar differential equation $\dot{v}_\tau =$

[2]See the proof of [34, Theorem 4.9].

$-(1-\theta)\hat{\alpha}(v_\tau)$ under the initial condition $v(t) = V^{\mu_i}(z)$. Then, [34, Lemma 3.4 and Lemma 4.4] and (55) show that there is $\beta_v \in \mathcal{KL}$, defined on $[0,\, \bar{\alpha}_{\mu_i}(r_d)]$, such that

$$V^{\mu_i}(x(\tau)) \le v(\tau) = \beta_v\big(V^{\mu_i}(z),\, \tau - t\big),$$

for any initial condition $z \in \Omega_{i+1} \setminus \Omega_{r_i}^{\mu_i}$ and all $\tau \in [t + t')$. Therefore, using (4) yields the following inequality:

$$\begin{aligned}\|x_\tau\| &\le \underline{\alpha}_{\mu_i}^{-1}\big(V^{\mu_i}(x_\tau)\big) \le \underline{\alpha}_{\mu_i}^{-1} \circ \beta_v\big(V^{\mu_i}(z),\, \tau - t\big) \\ &\le \underline{\alpha}_{\mu_i}^{-1} \circ \beta_v\big(\bar{\alpha}_{\mu_i}(\|z\|),\, \tau - t\big) \equiv \beta(\|z\|,\, \tau - t\big), \quad (57)\end{aligned}$$

where $\beta(y, s) \equiv \underline{\alpha}_{\mu_i}^{-1} \circ \beta_v\big(\bar{\alpha}_{\mu_i}(y),\, s\big)$ is of class $\mathcal{KL}$ by [34, Lemma 4.2]. On the other hand, for all $x_\tau \in \Omega_{r_i}^{\mu_i}$, we have $V^{\mu_i}(x_\tau) \le r_i$, and from (4) and (54),

$$\|x_\tau\| \le \alpha\bigg(\sup_{t \le s < \infty} \|e(s)\|\bigg), \quad (58)$$

where $\alpha(y) \equiv \underline{\alpha}_s^{-1}\big(\bar{\sigma}(R)y^2/\theta\big)$ is of class $\mathcal{K}$ [34, Lemma 4.2]. Finally, (57) and (58) imply that for all $z \in \Omega_{i+1}$ and all $\tau \ge t$, the trajectory $x_\tau(z; \mu_{i+1}, e)$ satisfies the inequality

$$\|x_\tau\| \le \beta(\|z\|,\, \tau - t\big) + \alpha\bigg(\sup_{t \le s < \infty} \|e(s)\|\bigg), \quad (59)$$

under (18). Here, instead of $[t, \infty)$, the supremum on the right hand side can be chosen over $[t, \tau]$ since x_τ depends only on $e(s)$ for $t \le s \le \tau$. This completes the proof of the local ISS. For the global case $\mathcal{D} = \Omega_i = \mathbb{R}^n$, $\underline{\alpha}_{\mu_i}$, $\bar{\alpha}_{\mu_i}$, $\underline{\alpha}_s$, and $\bar{\alpha}_s$ are all defined on $[0, \infty)$, and $\underline{\alpha}_{\mu_i} \in \mathcal{K}_\infty$ implies $\bar{\alpha}_{\mu_i} \in \mathcal{K}_\infty$ by (4). Therefore, d_i and the upper bound in (18) can be arbitrarily large, so ISS holds for arbitrary exploration e. Furthermore, the initial condition z can be also arbitrarily chosen since $\Omega_{r_i}^{\mu_i}$ with r_i ($< d_i$) defined in (54), and thereby, Ω_{i+1} ($\supset \Omega_{r_i}^{\mu_i}$) can be extended to $\mathbb{R}^n$ by increasing d_i (or r_i) to ∞ to include any given initial state $z \in \mathbb{R}^n$, the completion of the proof.

Appendix C: Proof of Theorem 3

Note that $(\mu, e) \in \mathcal{A}_{\mathcal{I}}(\Omega)$ implies $x_\tau(z; \mu, e)$ lies entirely in Ω, for all $\tau \ge t$. So, $V \in \mathcal{C}^1(\Omega)$ satisfies

$$V(x_{t+T}) - V(x_t) = \int_t^{t+T} \dot{V}(x_\tau)\, d\tau, \quad (60)$$

for any initial value $x_t = z \in \Omega$, where the time derivative $\dot{V}(x_\tau)$ is given by

$$\dot{V}(x_\tau) = \nabla^T V(x_\tau) \cdot (f(x_\tau) + g(x_\tau)[\mu(x_\tau) + e_\tau]). \quad (61)$$

Defining $\mathcal{H}(x, e) := H(x, \mu(x), \nabla V(x)) - 2\varphi^T(x)Re$ and substituting (60) and (61) into the I-TD (19), we obtain

$$\int_t^{t+T} \mathcal{H}\big(x_\tau(z; \mu, e), e_\tau\big)\, d\tau = 0. \quad (62)$$

Therefore, finding the solution of the advanced I-TD (19) for all $x_t = z \in \Omega$ is equivalent to solving (62) $\forall z \in \Omega$. Since $x_\tau(z; \mu, e) \in \Omega$ for all $\tau \ge t + T$, following the same steps with starting time $t + MT$, instead of t, yields

$$\int_{t+MT}^{t+(M+1)T} \mathcal{H}\big(x_\tau(z; \mu, e), e_\tau\big)\, d\tau = 0, \;\; \forall M \in \mathbb{Z}_+.$$

Then, summing up the integrals for all $M \in \mathbb{Z}_+$, we obtain

$$h(t; z) \equiv \int_t^\infty \mathcal{H}\big(x_\tau(z; \mu, e), e_\tau\big)\, d\tau = 0.$$

That is, $h(t; z) = 0$ for all $t \ge 0$ and all $z \in \Omega$. So, we have $\dot{h}(t; z) = -\mathcal{H}\big(x_\tau(z; \mu, e), e_\tau\big)|_{\tau=t} = 0$, and thereby,

$$\mathcal{H}\big(z, e_t\big) = 0, \;\; \forall t \ge 0 \text{ and } \forall z \in \Omega. \quad (63)$$

Since e is T-periodic, *i.e.*, $e_\tau = e_{\tau+T}$ for all $\tau \ge t$, (63) is reduced to

$$\mathcal{H}\big(z, e_\tau\big) = 0, \;\; \forall \tau \in [t, t + T) \text{ and } \forall z \in \Omega,$$

which is equivalent to (20). The proof of the opposite direction can be easily done by first integrating (20) and then substituting (60) and (61).

Jae Young Lee received his B.S. degree in Information & Control Engineering in 2006 from Kwangwoon University, Seoul, Korea. Since Sep. 2006, he has been working as a research assistant in the Control Engineering Laboratory, Yonsei University, Seoul, Korea, where he is currently pursuing his Ph.D. degree in Electrical and Electronic Engineering. His major research interests include approximate dynamic programming/reinforcement learning, optimal/adaptive control, nonlinear control theories, neural networks, and applications to unmanned vehicles, multi-agent systems, robotics, and power systems.

Jin Bae Park received the B.S. degree in electrical engineering from Yonsei University, Seoul, Korea, and the M.S. and Ph.D. degrees in electrical engineering from Kansas State University, Manhattan, KS, USA, in 1977, 1985, and 1990, respectively. Since 1992, he has been with the Department of Electrical and Electronic Engineering, Yonsei University, where he is currently a Professor. His major research interests include robust control and filtering, nonlinear control, intelligent mobile robot, fuzzy logic control, neural networks, adaptive dynamic programming, chaos theory, and genetic algorithms. Dr. Park served as the Editor-in-Chief (2006-2010) for the International Journal of Control, Automation, and Systems and the President (2013) for the Institute of Control, Robot, and Systems Engineers (ICROS). He is currently serving as the Senior Vice President for Yonsei University.

Yoon Ho Choi received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from Yonsei University, Seoul, Korea, in 1980, 1982, and 1991, respectively. Since 1993, he has been with Department of Electronic Engineering, Kyonggi University, Suwon, Korea, where he is currently a Professor. He was with Department of Electrical Engineering, The Ohio State University, where he was a Visiting Scholar (2000-2002, 2009-2010). His research interests include nonlinear control, intelligent control, multi-legged and mobile robots, networked control systems, and ADP based control. Prof. Choi was the Director (2003-2004, 2007-2008) for the Institute of Control, Robotics and Systems (ICROS). He is serving as the Vice-President for the ICROS (2012-present).