Policy Iterations for Reinforcement Learning Problems in Continuous Time and Space — Fundamental Theory and Methods ⋆

Jaeyoung Lee [a,*], Richard S. Sutton [b]

[a] *Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, N2L 3G1.*

[b] *Department of Computing Science, University of Alberta, Edmonton, AB, Canada, T6G 2E8.*

Abstract

Policy iteration (PI) is a recursive process of policy evaluation and improvement for solving an optimal decision-making/control problem, or in other words, a reinforcement learning (RL) problem. PI has also served as the fundamental to develop RL methods. In this paper, we propose two PI methods, called differential PI (DPI) and integral PI (IPI), and their variants to solve a general RL problem in continuous time and space (CTS), with the environment modeled by a system of ordinary differential equations (ODEs). The proposed methods inherit the current ideas of PI in classical RL and optimal control and theoretically support the existing RL algorithms in CTS: TD-learning and value-gradient-based (VGB) greedy policy update. We also provide case studies including 1) discounted RL and 2) optimal control tasks. Fundamental mathematical properties — admissibility, uniqueness of the solution to the Bellman equation, monotone improvement, convergence, and optimality of the solution to the Hamilton-Jacobi-Bellman equation (HJBE) — are all investigated in-depth and improved from the existing theory, along with the general and case studies. Finally, the proposed ones are simulated with an inverted-pendulum model and their model-based and partially model-free implementations to support the theory and further investigate them.

Key words: policy iteration, reinforcement learning, optimization under uncertainties, continuous time and space, iterative schemes, adaptive systems

1 Introduction

Policy iteration (PI) is a class of approximate dynamic programming (ADP) for recursively solving an optimal decision-making/control problem by alternating between *policy evaluation* to obtain the value function (VF) w.r.t. the current policy (a.k.a. the current control law in control theory) and *policy improvement* to improve the policy by optimizing it using the obtained VF (Sutton and Barto, 2018; Lewis and Vrabie, 2009). PI was first proposed by Howard (1960) in a stochastic environment known as the Markov decision process (MDP) and has served as a fundamental principle to develop RL methods, especially in an environment modeled or approximated by an MDP in discrete time and space. Finite-time convergence of PIs towards the optimal solution has been proven in a finite MDP, and the forward-in-time computation of PI like the other ADP methods alleviates the problem known as the curse of dimensionality (Powell, 2007). A discount factor $\gamma \in [0, 1]$ is normally introduced to both PI and RL to suppress the future reward and thereby have a finite return. Sutton and Barto (2018) give a comprehensive overview of PI and RL algorithms with their practical applications and recent success.

On the other hand, the dynamics of a real physical task is in the majority of cases modeled as a system of (ordinary) differential equations (ODEs) inevitably in continuous time and space (CTS). PI has also been studied in such a continuous domain mainly under the framework of deterministic optimal control, where the optimal solution is characterized by the partial differential Hamilton-Jacobi-Bellman (HJB) equation (HJBE). However, an HJBE is extremely difficult or hopeless to be solved analytically, except for a very few exceptional cases. PI in this field is often referred to as successive approximation of the HJBE (to recursively solve it!), and the main difference among them lies in their policy evaluation — the earlier versions of PI solve the associated *differential* Bellman equation (BE) (a.k.a. Lyapunov or Hamiltonian equation) to obtain the VF for the target pol-

⋆ The authors gratefully acknowledge the support of Alberta Innovates–Technology Futures, the Alberta Machine Intelligence Institute, DeepMind, the Natural Sciences and Engineering Research Council of Canada, and the Japanese Science and Technology agency (JST) ERATO project JPMJER1603: HASUO Metamathematics for Systems Design.

* Corresponding author. Tel.: +1 587 597 8677.

Email addresses: jaeyoung.lee@uwaterloo.ca (Jaeyoung Lee), rsutton@ualberta.ca (Richard S. Sutton).

icy (e.g., Leake and Liu, 1967; Kleinman, 1968; Saridis and Lee, 1979; Beard, Saridis, and Wen, 1997; Abu-Khalaf and Lewis, 2005 to name a few). Murray, Cox, Lendaris, and Saeks (2002) proposed a trajectory-based policy evaluation that can be viewed as a deterministic Monte-Carlo prediction (Sutton and Barto, 2018). Motivated by those two approaches above, Vrabie and Lewis (2009) proposed a partially model-free PI scheme [1] called integral PI (IPI), which is more relevant to RL in that the associated BE is of a temporal difference (TD) form — see Lewis and Vrabie (2009) for a comprehensive overview. Fundamental mathematical properties of those PIs, i.e., convergence, admissibility, and monotone improvement of the policies, are investigated in the literature above. As a result, it has been shown that the policies generated by PI methods are always monotonically improved and admissible; the sequence of VFs generated by PI methods in CTS is shown to converge to the optimal solution, quadratically in the LQR case (Kleinman, 1968). These fundamental properties are also discussed, improved, and generalized in this paper in a general setting that includes both RL and optimal control problems in CTS.

On the other hand, the aforementioned PI methods in CTS were all designed via Lyapunov's stability theory (Khalil, 2002) to ensure that the generated policies all asymptotically stabilizes the dynamics and yield finite returns (at least on a bounded region around an equilibrium state), provided that so is the initial policy. Here, the dynamics under the initial policy needs to be asymptotically stable to run the PI methods, which is, however, quite contradictory for IPI — it is partially model-free, but it is hard or even impossible to find such a stabilizing policy *without knowing the dynamics*. Besides, compared with the RL problems in CTS, e.g., those in (Doya, 2000; Mehta and Meyn, 2009; Frémaux, Sprekeler, and Gerstner, 2013), this stability-based approach restricts the range of the discount factor γ and the class of the dynamics and the cost (i.e., reward) as follows.

(1) When discounted, the discount factor $\gamma \in (0, 1)$ must be larger than some threshold so as to hold the asymptotic stability of the target optimal policy (Gaitsgory, Grüne, and Thatcher, 2015; Modares, Lewis, and Jiang, 2016). If not, there is no point in considering stability: PI finally converges to that (possibly) *non-stabilizing* optimal solution, even if the PI is convergent and the initial policy is *stabilizing*. Furthermore, the threshold on γ depends on the dynamics (and the cost), and thus it cannot be calculated without knowing the dynamics, a contradiction to the use of any (partially) model-free methods such as IPI. Due to these restrictions on γ, the PI methods mentioned above for nonlinear optimal control focused on the problems without discount factor, rather than discounted ones.

(2) In the case of optimal regulations, (i) the dynamics is assumed to have at least one equilibrium state; [2] (ii) the goal is to stabilize the system optimally for that equilibrium state, although bifurcation or multiple isolated equilibrium states to be considered may exist; (iii) for such optimal stabilization, the cost is crafted to be positive (semi-)definite — when the equilibrium state of interest is transformed to zero without loss of generality (Khalil, 2002). Similar restrictions exist in optimal tracking problems that can be transformed into equivalent optimal regulation problems (e.g., see Modares and Lewis, 2014).

In this paper, we consider a general RL framework in CTS, where reasonably minimal assumptions were imposed — 1) the global existence and uniqueness of the state trajectories, 2) (whenever necessary) continuity, differentiability, and/or existence of maximum(s) of functions, and 3) no assumption on the discount factor $\gamma \in (0, 1]$ — to include a broad class of problems. The RL problem in this paper not only contains those in the RL literature (e.g., Doya, 2000; Mehta and Meyn, 2009; Frémaux et al., 2013) in CTS but also considers the cases beyond stability framework (at least theoretically), where state trajectories can be still bounded or even diverge (Proposition 2.2; §5.4; §§G.2 and G.3 in Appendices (Lee and Sutton, 2020a)). It also includes input-constrained and unconstrained problems presented in both RL and optimal control literature as its special cases.

Independent of the research on PI, several RL methods have come to be proposed in CTS based on RL ideas in the discrete domain. Advantage updating was proposed by Baird III (1993) and then reformulated by Doya (2000) under the environment represented by a system of ODEs; see also Tallec, Blier, and Ollivier (2019)'s recent extension of advantage updating using deep neural networks. Doya (2000) also extended TD(λ) to the CTS domain and then combined it with his proposed policy improvement methods such as the value-gradient-based (VGB) greedy policy update. See also Frémaux et al. (2013)'s extension of Doya (2000)'s continuous actor-critic with spiking neural networks. Mehta and Meyn (2009) proposed Q-learning in CTS based on stochastic approximation. Unlike in MDP, however, these RL methods were rarely relevant to the PI methods in CTS due to the gap between optimal control and RL — the proposed PI methods bridge this gap with a direct connection to TD learning in CTS and VGB greedy policy update (Doya, 2000; Frémaux et al., 2013). The investigations of the ADP for the other RL methods remain as a future work or see our preliminary result (Lee and Sutton, 2017).

1.1 Main Contributions

In this paper, the main goal is to build up a theory on PI in a general RL framework, from the ideas of PI in classical RL and optimal control, when the time domain and the state-action space are all continuous and a system of ODEs models

[1] The term "partially model-free" in this paper means that the algorithm can be implemented using some partial knowledge (i.e., the input-coupling terms) of the dynamics.

[2] For an example of a dynamics with no equilibrium state, see (Haddad and Chellaboina, 2008, Example 2.2).

the environment. As a result, a series of PI methods are proposed that theoretically support the existing RL methods in CTS: TD learning and VGB greedy policy update. Our main contributions are summarized as follows.

(1) Motivated by the PI methods in optimal control, we propose a model-based PI named differential PI (DPI) and a partially model-free PI called IPI, in our general RL framework. The proposed schemes do not necessarily require an initial stabilizing policy to run and can be considered a sort of fundamental PI methods in CTS.

(2) By case studies that contain both discounted RL and optimal control frameworks, the proposed PI methods and theory for them are simplified, improved, and specialized, with strong connections to RL and optimal control in CTS.

(3) Fundamental mathematical properties for PI (and ADP) — admissibility, uniqueness of the solution to the BE, monotone improvement, convergence, and optimality of the solution to the HJBE — are all investigated in-depth along with the general and case studies. Optimal control case studies also examine the stability properties of PI. As a result, the existing properties for PI in optimal control are improved and rigorously generalized.

Simulation results for an inverted-pendulum model are also provided with the model-based and partially model-free implementations to support the theory and further investigate the proposed methods under an admissible (but not necessarily stabilizing) initial policy, with the strong connections to 'bang-bang control' and 'RL with simple binary reward,' both of which are beyond the scope of our theory. Here, the RL problem in this paper is formulated stability-freely (which is well-defined under the minimal assumptions), so that the (initial) admissible policy is not necessarily stabilizing in the theory and the proposed PI methods to solve it.

1.2 Organizations

This paper is organized as follows. In §2, our general RL problem in CTS is formulated along with mathematical backgrounds, notations, and statements related to BEs, policy improvement, and the HJBE. In §3, we present and discuss the two main PI methods (i.e., DPI and IPI) and their variants, with strong connections to the existing RL methods in CTS. We show in §4 the fundamental properties of the proposed PI methods: admissibility, uniqueness of the solution to the BE, monotone improvement, convergence, and optimality of the solution to the HJBE. Those properties in §4 and the Assumptions made in §§2 and 4 are simplified, improved, and relaxed in §5 with the following case studies: 1) concave Hamiltonian formulations (§5.1); 2) discounted RL with bounded VF/reward (§5.2); 3) RL problem with local Lipschitzness (§5.3); 4) nonlinear optimal control (§5.4). In §6, we discuss and provide the simulation results of the main PI methods. Finally, conclusions follow in §7.

Due to space limitations, appendices are published separately (Lee and Sutton, 2020a). It provides related works (§A), a summary of notations and terminologies (§B), details on the theory and implementations (§§C–E, and H), a pathological example (§F), additional case studies (§G), and all the proofs (§I). Throughout the paper, any section starting with an alphabet as above will indicate a section in the appendix (Lee and Sutton, 2020a).

1.3 Notations and Terminologies

The following notations and terminologies will be used throughout the paper (see §B for a complete list of notations and terminologies, including those not listed below). In any mathematical statement, iff stands for "if and only if" and s.t. for "such that." "$\doteq$" indicates the equality relationship that is true *by definition*.

(Sets, vectors, and matrices). $\mathbb{N}$ and $\mathbb{R}$ are the sets of all natural and real numbers, respectively. $\mathbb{R}^{n\times m}$ is the set of all n-by-m real matrices. A^T is the transpose of $A \in \mathbb{R}^{n\times m}$. $\mathbb{R}^n \doteq \mathbb{R}^{n\times 1}$ denotes the n-dimensional Euclidean space. $\|x\|$ is the Euclidean norm of $x \in \mathbb{R}^n$, i.e., $\|x\| \doteq (x^\mathsf{T}x)^{1/2}$.

(Euclidean topology). Let $\Omega \subseteq \mathbb{R}^n$. Ω is said to be compact iff it is closed and bounded. Ω° denotes the interior of Ω; $\partial\Omega$ is the boundary of Ω. If Ω is open, then $\Omega \cup \partial\Omega$ (resp. Ω) is called an n-dimensional manifold with (resp. without) boundary. A manifold contains no isolated point.

(Functions, sequences, and convergence). A function $f : \Omega \to \mathbb{R}^m$ is said to be C^1, denoted by $f \in \mathrm{C}^1$, iff all of its first-order partial derivatives exist and are continuous over the interior Ω°; $\nabla f : \Omega^\circ \to \mathbb{R}^{m\times n}$ denotes the gradient of f. A sequence of functions $\langle f_i\rangle_{i=1}^\infty$, abbreviated by $\langle f_i\rangle$ or f_i, is said to converge locally uniformly iff for each $x \in \Omega$, there is a neighborhood of x on which $\langle f_i\rangle$ converges uniformly. For any two functions $f_1, f_2 : \mathbb{R}^n \to [-\infty, \infty)$, we write $f_1 \leqslant f_2$ iff $f_1(x) \leq f_2(x)$ for all $x \in \mathbb{R}^n$.

2 Preliminaries

Let $\mathcal{X} \doteq \mathbb{R}^l$ be the state space and $\mathbb{T} \doteq [0, \infty)$ be the time space. An m-dimensional manifold $\mathcal{U} \subseteq \mathbb{R}^m$ with or without boundary is called an action space. We also denote $\mathcal{X}^\mathsf{T} \doteq \mathbb{R}^{1\times l}$ for notational convenience. The environment in this paper is described in CTS by a system of ODEs:

$$\dot{X}_t = f(X_t, U_t), \quad U_t \in \mathcal{U} \tag{1}$$

where $t \in \mathbb{T}$ is time instant, $\mathcal{U} \subseteq \mathbb{R}^m$ is an action space, and the dynamics $f : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ is a continuous function; $X_t, \dot{X}_t \in \mathcal{X}$ denote the state vector and its time derivative, at time t, respectively; the action trajectory $t \mapsto U_t$ is a continuous function from $\mathbb{T}$ to $\mathcal{U}$. We assume that $t = 0$ is the initial time without loss of generality [3] and that

[3] If the initial time t_0 is not zero, then proceed with the time variable $t' = t - t_0$, which satisfies $t' = 0$ at the initial time $t = t_0$.

Assumption. *The state trajectory $t \mapsto X_t$ is uniquely defined over the entire time interval $\mathbb{T}$.* [4]

A policy π refers to a continuous function $\pi : \mathcal{X} \to \mathcal{U}$ that determines the state trajectory $t \mapsto X_t$ by $U_t = \pi(X_t)$ for all $t \in \mathbb{T}$. For notational efficiency, we employ the $\mathbb{G}$-notation $\mathbb{G}^x_\pi[\,Y]$, which means the value Y when $X_0 = x$ and $U_t = \pi(X_t)$ for all $t \in \mathbb{T}$. Here, $\mathbb{G}$ stands for "Generator," and $\mathbb{G}^x_\pi$ can be thought of as the corresponding notation of the expectation $\mathbb{E}_\pi[\,\cdot\,|S_0 = x]$ in the RL literature (Sutton and Barto, 2018), without playing any stochastic role. Note that the limits and integrals are exchangeable with $\mathbb{G}^x_\pi[\,\cdot\,]$ in order (whenever those limits and integrals are defined for any $X_0 \in \mathcal{X}$ and any action trajectory $t \mapsto U_t$). For example, for any continuous function $v : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{G}^x_\pi\bigg[\int v(X_t)\,dt\bigg] = \int \mathbb{G}^x_\pi\big[v(X_t)\big]\,dt = \int v\big(\mathbb{G}^x_\pi[X_t]\big)\,dt,$$

where the three mean the same: $\int v(X_t)\,dt$ when $X_0 = x$ and $U_t = \pi(X_t)\ \forall t \in \mathbb{T}$. Also note: $\mathbb{G}^x_\pi[U_t] = \mathbb{G}^x_\pi[\pi(X_t)]$.

Finally, the time-derivative $\dot v : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ of a C^1 function $v : \mathcal{X} \to \mathbb{R}$ is given by $\dot v(X_t, U_t) = \nabla v(X_t) f(X_t, U_t)$ — applying the chain rule and (1). Here, $X_t \in \mathcal{X}$ and $U_t \in \mathcal{U}$ are free variables, and $\dot v$ is continuous since so is f.

2.1 RL problem in Continuous Time and Space

The RL problem considered in this paper is to find the best policy π_* that maximizes the infinite horizon value function (VF) $v_\pi : \mathcal{X} \to [-\infty, \infty)$ defined as

$$v_\pi(x) \doteq \mathbb{G}^x_\pi\bigg[\int_0^\infty \gamma^t \cdot R_t\,dt\bigg], \tag{2}$$

where the reward R_t is determined by a continuous reward function $r : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ as $R_t = r(X_t, U_t)$; $\gamma \in (0, 1]$ is the discount factor. Throughout the paper, the attenuation rate

$$\alpha \doteq -\ln \gamma \geq 0$$

will be used interchangeably for simplicity. For a policy π, we denote $f_\pi(x) \doteq f(x, \pi(x))$ and $r_\pi(x) \doteq r(x, \pi(x))$, which are continuous as so are f, r, and π by definitions.

Assumption. *A maximum of the reward function r:*

$$r_{\max} \doteq \max\big\{r(x,u) : (x,u) \in \mathcal{X} \times \mathcal{U}\big\}$$

exists and for $\gamma = 1$, $r_{\max} = 0$. [5]

Note that the integrand $t \mapsto \gamma^t R_t$ is continuous since so are $t \mapsto X_t$, $t \mapsto U_t$, and r. So, by the above assumption on r, the time integral and thereby v_π in (2) are well-defined in the Lebesque sense (Folland, 1999, Chapter 2) over $[-\infty, \infty)$ and as shown below, uniformly upper-bounded.

Lemma 2.1 *There exists a constant $\bar v \in \mathbb{R}$ s.t. $v_\pi \leqslant \bar v$ for any policy π; $\bar v = 0$ for $\gamma = 1$ and otherwise, $\bar v = r_{\max}/\alpha$.*

By Lemma 2.1, the VF is always less than some constant, but it is still possible that $v_\pi(x) = -\infty$ for some $x \in \mathcal{X}$. In this paper, the finite VFs are characterized by the notion of admissibility given below.

Definition. *A policy π (or its VF v_π) is said to be admissible, denoted by $\pi \in \Pi_{\mathsf{a}}$ (or $v_\pi \in \mathcal{V}_{\mathsf{a}}$), iff $v_\pi(x)$ is finite for all $x \in \mathcal{X}$. Here, Π_{a} and $\mathcal{V}_{\mathsf{a}}$ denote the sets of all admissible policies and admissible VFs, respectively.*

To make our RL problem feasible, we assume:

Assumption. *There exists at least one admissible policy, and every admissible VF is C^1.* (3)

The following proposition gives a criterion for admissibility and boundedness.

Proposition 2.2 *A policy π is admissible if there exist a function $\rho : \mathcal{X} \to \mathbb{R}$ and a constant $\kappa < \alpha$, both possibly depending on the policy π, such that*

$$\forall x \in \mathcal{X}\colon\ e^{\kappa t} \cdot \rho(x) \leq \mathbb{G}^x_\pi[R_t]\ \text{ for all } t \in \mathbb{T}. \tag{4}$$

Moreover, v_π is bounded if so is ρ.

Remark. *The criterion* (4) *means that the reward R_t under π does not diverge to $-\infty$ exponentially with the rate α or higher. For $\gamma = 1$ (i.e., $\alpha = 0$), it means exponential convergence $R_t \to 0$. The condition* (4) *is fairly general and so satisfied by the examples in §§5.2, G.2, and G.3.*

2.2 Bellman Equations with Boundary Condition

Define the Hamiltonian function $h : \mathcal{X} \times \mathcal{U} \times \mathcal{X}^{\mathsf{T}} \to \mathbb{R}$ as

$$h(x, u, p) \doteq r(x, u) + p\, f(x, u) \tag{5}$$

(which is continuous as so are f and r) and the γ-discounted cumulative reward $\mathfrak{R}_\eta$ up to a given time horizon $\eta > 0$ as

$$\mathfrak{R}_\eta \doteq \int_0^\eta \gamma^t \cdot R_t\,dt$$

as a short-hand notation. The following lemma then shows the equivalence of the Bellman-like (in)equalities.

Lemma 2.3 *Let $\sim$ be a binary relation on $\mathbb{R}$ that belongs to $\{=, \leq, \geq\}$ and $v : \mathcal{X} \to \mathbb{R}$ be C^1. Then, for any policy π,*

$$v(x) \sim \mathbb{G}^x_\pi\big[\,\mathfrak{R}_\eta + \gamma^\eta \cdot v(X_\eta)\,\big] \tag{6}$$

holds for all $x \in \mathcal{X}$ and all horizon $\eta > 0$ iff

$$\alpha \cdot v(x) \sim h(x, \pi(x), \nabla v(x)) \qquad \forall x \in \mathcal{X}. \tag{7}$$

[4] Not imposed on our problem for generality but strongly related to this global existence of the unique state trajectory $t \mapsto X_t$ is Lipschitz continuity of f and f_π (Khalil, 2002, Theorems 3.1 and 3.2); see also §5.3 for related discussions to this.

[5] If $r_{\max} \neq 0$ and $\gamma = 1$, then, proceed with the reward function $r'(x, u) \doteq r(x, u) - r_{\max}$ whose maximum is now zero.

By splitting the time-integral in (2) at $\eta > 0$, we can easily see that the VF v_π satisfies the Bellman equation (BE):

$$v_\pi(x) = \mathbb{G}_\pi^x\big[\mathfrak{R}_\eta + \gamma^\eta \cdot v_\pi(X_\eta)\big] \tag{8}$$

that holds for any $x \in \mathcal{X}$ and $\eta > 0$. Assuming $v_\pi \in \mathcal{V}_{\mathsf{a}}$ and using (8), we obtain its boundary condition at $\eta = \infty$.

Proposition 2.4 *Suppose that π is admissible. Then,*

$$\lim_{t\to\infty} \mathbb{G}_\pi^x\big[\gamma^t \cdot v_\pi(X_t)\big] = 0 \qquad \forall x \in \mathcal{X}.$$

By the application of Lemma 2.3 to the BE (8) under (3), the following *differential BE* holds whenever $\pi \in \Pi_{\mathsf{a}}$:

$$\alpha \cdot v_\pi(x) = h(x, \pi(x), \nabla v_\pi(x)), \tag{9}$$

where the function $x \mapsto h(x, \pi(x), \nabla v_\pi(x))$ is continuous since so are the associated functions h, π, and ∇v_π. Whenever necessary, we call (8) the *integral BE* to distinguish it from the *differential BE* (9).

In what follows, we state that the boundary condition (12), *the counterpart of that in Proposition 2.4,* is actually necessary and sufficient for a solution v of the BE (10) or (11) to be equal to the corresponding VF v_π and ensure $\pi \in \Pi_{\mathsf{a}}$.

Theorem 2.5 (Policy Evaluation)
Fix the horizon $\eta > 0$ and suppose there exists a function $v : \mathcal{X} \to \mathbb{R}$ s.t. either of the followings holds for a policy π:

(1) v satisfies the integral BE:

$$v(x) = \mathbb{G}_\pi^x\big[\mathfrak{R}_\eta + \gamma^\eta \cdot v(X_\eta)\big] \qquad \forall x \in \mathcal{X}; \tag{10}$$

(2) v is C^1 and satisfies the differential BE:

$$\alpha \cdot v(x) = h(x, \pi(x), \nabla v(x)) \qquad \forall x \in \mathcal{X}. \tag{11}$$

Then, π is admissible and $v = v_\pi$ iff

$$\lim_{k\to\infty} \mathbb{G}_\pi^x\big[\,\gamma^{k\cdot\eta} \cdot v(X_{k\cdot\eta})\,\big] = 0 \qquad \forall x \in \mathcal{X}. \tag{12}$$

For sufficiency, the boundary condition (12) can be replaced with inequality on v and π under certain conditions on v, as shown in §C. This sufficient condition is particularly related to the optimal control framework in §5.4 but applicable to any case in this paper as an alternative to (12) (see §C).

2.3 Policy Improvement

Define a partial order among policies: $\pi \preccurlyeq \pi'$ iff $v_\pi \leqslant v_{\pi'}$. Then, we say that a policy π' is improved over π iff $\pi \preccurlyeq \pi'$. In CTS, the Bellman inequality (13) for $v = v_\pi$ ensures this policy improvement over an admissible policy π as shown below. The inequality becomes the BE (9) when $\pi = \pi'$.

Lemma 2.6 *If $v \in \mathrm{C}^1$ is upper-bounded (by zero if $\gamma = 1$) and satisfies for a policy π'*

$$\alpha \cdot v(x) \le h(x, \pi'(x), \nabla v(x)) \qquad \forall x \in \mathcal{X}, \tag{13}$$

then π' is admissible and $v \leqslant v_{\pi'}$.

In what follows, for the existence of a maximally improving policy, we assume on the Hamiltonian function h:

Assumption. *There exists a function $u_* : \mathcal{X} \times \mathcal{X}^\mathsf{T} \to \mathcal{U}$ such that u_* is continuous and*

$$u_*(x,p) \in \arg\max_{u\in\mathcal{U}} h(x,u,p) \quad \forall (x,p) \in \mathcal{X} \times \mathcal{X}^\mathsf{T}. \tag{14}$$

Here, (14) simply means that for each (x,p), the function $u \mapsto h(x,u,p)$ has its maximum at $u_*(x,p) \in \mathcal{U}$. Then, for any admissible policy π, there exists a continuous function $\pi' : \mathcal{X} \to \mathcal{U}$ such that

$$\pi'(x) \in \arg\max_{u\in\mathcal{U}} h(x, u, \nabla v_\pi(x)) \quad \forall x \in \mathcal{X}. \tag{15}$$

We call such a continuous function π' a *maximal policy* (over $\pi \in \Pi_{\mathsf{a}}$). Given u_*, we can directly obtain a maximal policy π' by

$$\pi'(x) = u_*(x, \nabla v_\pi(x)). \tag{16}$$

In general, there may exist multiple maximal policies, but if u_* in (14) is *unique*, then π' satisfying (15) is *uniquely* given by (16). For non-affine optimal control problems, Leake and Liu (1967) and Bian, Jiang, and Jiang (2014) imposed assumptions similar to the above Assumption on u_* plus its uniqueness. Here, the existence of u_* is ensured if $\mathcal{U}$ is compact; u_* is unique if the function $u \mapsto h(x,u,p)$ is strictly concave and C^1 for each (x,p) — see §D for more studies; for such case examples, see §5.1; Cases 1 and 2 in §6.

Theorem 2.7 (Policy Improvement)
Suppose π is admissible. Then, the policy π' given by (15) *is also admissible and satisfies $\pi \preccurlyeq \pi'$.*

2.4 Hamilton-Jacobi-Bellman Equation (HJBE)

Under the Assumptions made so far, the optimal solution of the RL problem can be characterized via the HJBE (17):

$$\alpha \cdot v_*(x) = \max_{u\in\mathcal{U}} h(x, u, \nabla v_*(x)) \qquad \forall x \in \mathcal{X} \tag{17}$$

and the associated policy $\pi_* : \mathcal{X} \to \mathcal{U}$ such that

$$\pi_*(x) \in \arg\max_{u\in\mathcal{U}} h(x, u, \nabla v_*(x)) \qquad \forall x \in \mathcal{X}, \tag{18}$$

both of which are the key to prove the convergence of PIs towards the optimal solution v_* (and π_*) in §4. Note that once a C^1 solution $v_* : \mathcal{X} \to \mathbb{R}$ to the HJBE (17) exists, then so does a continuous function (i.e., a policy) π_* satisfying (18) by the Assumption on the existence of a continuous function u_* satisfying (14) and is given by

$$\pi_*(x) = u_*(x, \nabla v_*(x)). \tag{19}$$

In what follows, we show that satisfying the HJBE (17) and (18) is necessary for (v_*, π_*) to be optimal over the entire admissible space.

Theorem 2.8 *If there exists an optimal policy π_* whose VF v_* satisfies $v \leqslant v_*$ for any $v \in \mathcal{V}_{\mathsf{a}}$, then v_* and π_* satisfy the HJBE* (17) *and* (18)*, respectively.*

There may exist another optimal policy π'_* than π_*, but their VFs are always the same by $\pi_* \preccurlyeq \pi'_*$ and $\pi'_* \preccurlyeq \pi_*$ and equal to a solution v_* to the HJBE (17) by Theorem 2.8. In this paper, if exist, π_* denotes any one of the optimal policies, and v_* is the unique common VF for them which we call the optimal VF. In general, they denote a solution v_* to the HJBE (17) and an associated HJB policy π_* s.t. (18) holds (or if specified, an associated function π_* satisfying (18)).

Remark 2.9 *The reward function r has to be appropriately designed in such a way that the function $u \mapsto h(x,u,p)$ for each (x,p) at least has a maximum (so that* (14) *holds for some u_*). Otherwise, the maximal policy π' in* (15) *and/or the solution v_* to the HJBE* (17) *(and accordingly, π_* in* (18)*) may not exist since neither do the maxima in those equations; such a pathological example is given in §F for a simple non-affine dynamics f. In §5.1.2, we revisit this issue and propose a technique applicable to a class of non-affine RL problems to ensure the existence and continuity of u_*.*

The optimality of the HJB solution (v_*, π_*) is investigated more in §E, e.g., the sufficient conditions and case studies, in connection with the PIs presented in the next section.

3 Policy Iterations

Now, we are ready to state the two main PI schemes, DPI and IPI. Here, the former is a model-based approach, and the latter is a partially model-free PI. Their simplified (partially model-free) versions discretized in time will be also discussed after that. Until §6, we present and discuss those PI schemes in an ideal sense without introducing any function approximator, such as neural network, and any discretization in the state space. [6]

3.1 Differential Policy Iteration (DPI)

Our first PI, named differential policy iteration (DPI), is a model-based PI scheme extended from optimal control to our RL problem (e.g., see Leake and Liu, 1967; Beard et al., 1997; Abu-Khalaf and Lewis, 2005). Algorithm 1 describes the whole procedure of DPI — it starts with an initial admissible policy π_0 (line 1) and performs policy evaluation and improvement until v_i and/or π_i converges (lines 2–5). In *policy evaluation* (line 3), the agent solves the differential BE (20) to obtain the VF $v_i = v_{\pi_{i-1}}$ for the last policy π_{i-1}. Then, v_i is used in *policy improvement* (line 4) to obtain the next policy π_i by maximizing the associated Hamiltonian function in (21). Here, if $v_i = v_*$, then $\pi_i = \pi_*$ by (18) and (21).

Basically, DPI is model-based (see the definition (5) of h)

Algorithm 1: Differential Policy Iteration (DPI)

1 **Initialize:** π_0, an initial admissible policy; $i \leftarrow 1$, iteration index;
2 **repeat**
3 **Policy Evaluation:** given π_{i-1}, find a C^1 function $v_i : \mathcal{X} \to \mathbb{R}$ satisfying the differential BE:
$$\alpha \cdot v_i(x) = h(x, \pi_{i-1}(x), \nabla v_i(x)) \quad \forall x \in \mathcal{X}; \quad (20)$$
4 **Policy Improvement:** find a policy π_i such that
$$\pi_i(x) \in \arg\max_{u \in \mathcal{U}} h(x, u, \nabla v_i(x)) \ \forall x \in \mathcal{X}; \quad (21)$$
5 $i \leftarrow i + 1$;
until *convergence is met.*

and does not rely on any state trajectory data. On the other hand, its policy evaluation is closely related to TD learning methods in CTS (Doya, 2000; Frémaux et al., 2013). To see this, note that (20) can be expressed w.r.t. (X_t, U_t) as $\mathbb{G}^x_{\pi_{i-1}}[\delta_t(v_i)] = 0$ for all $x \in \mathcal{X}$ and $t \in \mathbb{T}$, where δ_t denotes the TD error defined as

$$\delta_t(v) \doteq R_t + \dot{v}(X_t, U_t) - \alpha \cdot v(X_t)$$

for any C^1 function $v : \mathcal{X} \to \mathbb{R}$. Frémaux et al. (2013) used δ_t as the TD error in their model-free actor-critic and approximated v and the model-dependent part $\dot{v}$ in δ_t with a spiking neural network. δ_t is also the TD error in TD(0) in CTS (Doya, 2000), where $\dot{v}(X_t, U_t)$ is approximated by $(v(X_t) - v(X_{t-\Delta t}))/\Delta t$ in backward time, for a sufficiently small time step Δt chosen within the time interval $(0, \alpha^{-1})$; under the above approximation of $\dot{v}$, $\delta_t(v)$ can be expressed in a similar form to the TD error in discrete-time as

$$\delta_t(v) \approx R_t + \hat{\gamma}_{\mathsf{d}} \cdot V(X_t) - V(X_{t-\Delta t}) \quad (22)$$

for $V \doteq v/\Delta t$ and $\hat{\gamma}_{\mathsf{d}} \doteq 1 - \alpha \Delta t \approx e^{-\alpha \Delta t}$ $(= \gamma^{\Delta t})$. Here, the discount factor $\hat{\gamma}_{\mathsf{d}}$ belongs to $(0,1)$ if so is γ, thanks to $\Delta t \in (0, \alpha^{-1})$, and $\hat{\gamma}_{\mathsf{d}} = 1$ whenever $\gamma = 1$. In summary, the policy evaluation of DPI is a process to solve the ideal BE corresponding to the existing TD learning methods in CTS (Doya, 2000; Frémaux et al., 2013).

3.2 Integral Policy Iteration (IPI)

Algorithm 2 describes the second PI, integral policy iteration (IPI), whose difference from DPI is that (20) and (21) in the policy evaluation and improvement are replaced by (23) and (24), respectively. The other steps are the same as DPI, except that the time horizon $\eta > 0$ is initialized (line 1) before the main loop.

In *policy evaluation* (line 3), IPI solves the integral BE (23) for a given fixed horizon $\eta > 0$ without using the explicit knowledge of the dynamics f of the system (1) — there are no explicit terms of f in (23), and the information on the dynamics f is implicitly captured by the state trajectory

[6] When we implement any of the PI schemes, both are obviously required (except linear quadratic regulation (LQR) cases) since the structure of the VF is veiled and it is impossible to perform the policy evaluation and improvement for an (uncountably) infinite number of points in the continuous state space $\mathcal{X}$ (see also §6 for implementation examples, with §H for details).

Algorithm 2: Integral Policy Iteration (IPI)

1 **Initialize:** $\begin{cases} \pi_0, \text{ an initial admissible policy;} \\ \eta > 0, \text{ time horizon;} \\ i \leftarrow 1, \text{ iteration index;} \end{cases}$

2 **repeat**

3 **Policy Evaluation:** given π_{i-1}, find a C^1 function $v_i : \mathcal{X} \to \mathbb{R}$ satisfying the integral BE:

$$v_i(x) = \mathbb{G}^x_{\pi_{i-1}}\big[\mathfrak{R}_\eta + \gamma^\eta \cdot v_i(X_\eta)\big] \ \ \forall x \in \mathcal{X}; \tag{23}$$

4 **Policy Improvement:** find a policy π_i such that

$$\pi_i(x) \in \arg\max_{u \in \mathcal{U}} \big[r(x,u) + \nabla v_i(x) f_{\mathsf{c}}(x,u) \big] \ \ \forall x \in \mathcal{X}; \tag{24}$$

5 $i \leftarrow i + 1$;

until *convergence is met.*

data $\{X_t : 0 \le t \le \eta\}$ generated under π_{i-1} at each ith step for a number of initial states $X_0 \in \mathcal{X}$. Note that by Theorem 2.5, solving the integral BE (23) for a fixed $\eta > 0$ and its differential version (20) in DPI are equivalent (as long as v_i satisfies the boundary condition (29) in §4).

In *policy improvement* (line 4), we consider the decomposition (25) of the dynamics f:

$$f(x,u) = f_{\mathsf{d}}(x) + f_{\mathsf{c}}(x,u), \tag{25}$$

where $f_{\mathsf{d}} : \mathcal{X} \to \mathcal{X}$ called a drift dynamics is independent of the action u and assumed *unknown*, and $f_{\mathsf{c}} : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ is the corresponding input-coupling dynamics assumed *known a priori*;[7] both f_{d} and f_{c} are assumed continuous. Since the term $\nabla v_\pi(x) f_{\mathsf{d}}(x)$ does not contribute to the maximization with respect to u, policy improvement (15) can be rewritten under the decomposition (25) as

$$\pi'(x) \in \arg\max_{u \in \mathcal{U}} \big[r(x,u) + \nabla v_\pi(x) f_{\mathsf{c}}(x,u) \big] \quad \forall x \in \mathcal{X} \tag{26}$$

by which the policy improvement (line 4) of Algorithm 2 is directly obtained. Note that the policy improvement (24) in Algorithm 2 and (26) are partially model-free, i.e., the maximizations do not depend on the unknown drift dynamics f_{d}.

The policy evaluation and improvement of IPI are completely and partially model-free, respectively. Thus the whole procedure of Algorithm 2 is partially model-free, i.e., it can be done even when a drift dynamics f_{d} is completely unknown. In addition to this partially model-free nature, the horizon $\eta > 0$ in IPI can be any value — it can be large or small — as long as the cumulative reward $\mathfrak{R}_\eta$ has no significant error when approximated in practice. In this sense, the time horizon η plays a similar role as the number n in the n-step TD predictions in discrete-time (Sutton and Barto, 2018). Indeed, if $\eta = n\Delta t$ for some $n \in \mathbb{N}$ and a sufficiently small $\Delta t > 0$, then by the forward-in-time approximation $\mathfrak{R}_\eta \approx G_n \cdot \Delta t$, where

$$G_n \doteq R_0 + \gamma_{\mathsf{d}} \cdot R_{\Delta t} + \gamma_{\mathsf{d}}^2 \cdot R_{2\Delta t} + \cdots + \gamma_{\mathsf{d}}^{n-1} \cdot R_{(n-1)\Delta t}$$

and $\gamma_{\mathsf{d}} \doteq \gamma^{\Delta t} \in (0,1]$, the integral BE (23) is expressed as

$$V_i(x) \approx \mathbb{G}^x_{\pi_{i-1}}\big[G_n + \gamma_{\mathsf{d}}^n \cdot V_i(X_\eta)\big], \tag{27}$$

where $V_i \doteq v_i/\Delta t$. We can also apply a higher-order approximation of $\mathfrak{R}_\eta$ — for instance, under the trapezoidal approximation, we have

$$V_i(x) \approx \mathbb{G}^x_{\pi_{i-1}}\Big[G_n + \tfrac{1}{2} \cdot (\gamma_{\mathsf{d}}^n \cdot R_\eta - R_0) + \gamma_{\mathsf{d}}^n \cdot V_i(X_\eta)\Big],$$

which uses the end-point reward R_η while (27) does not. Note that the TD error (22) is not easy to generalize for such multi-step TD predictions. When $n = 1$, on the other hand, the n-step BE (27) becomes

$$V_i(x) \approx \mathbb{G}^x_{\pi_{i-1}}\big[R_0 + \gamma_{\mathsf{d}} \cdot V_i(X_{\Delta t})\big] \ \ \forall x \in \mathcal{X}, \tag{28}$$

a similar TD expression to the BE in discrete-time (Sutton and Barto, 2018) and the TD error (22) in CTS.

3.3 Variants with Time Discretizations

As discussed in §§3.1 and 3.2 above, the BEs in DPI and IPI can be discretized in time in order to

(1) approximate $\dot v_i = \nabla v_i \cdot f$ in DPI, model-freely;
(2) calculate the cumulative reward $\mathfrak{R}_\eta$ in IPI;
(3) yield TD formulas similar to the BEs in discrete-time.

For instance, for a sufficiently small Δt, the discretized BE corresponding to DPI and TD(0) in CTS (Doya, 2000) is:

$$V_\pi(x) \approx \mathbb{G}^x_\pi\big[R_{\Delta t} + \hat\gamma_{\mathsf{d}} \cdot V_\pi(X_{\Delta t})\big] \qquad \forall x \in \mathcal{X},$$

where $V_\pi \doteq v_\pi/\Delta t$. The discretized BE for IPI is obviously of the form (28) for $n = 1$ and (27) for $n > 1$ (or one of the BEs with a higher-order approximation of $\mathfrak{R}_\eta$); if the integral BE (8) is discretized with a trapezoidal approximation for $n = 1$, then we also have

$$V_\pi(x) \approx \mathbb{G}^x_\pi\Big[\tfrac{1}{2} \cdot (R_0 + \gamma_{\mathsf{d}} \cdot R_{\Delta t}) + \gamma_{\mathsf{d}} \cdot V_\pi(X_{\Delta t})\Big] \ \ \forall x \in \mathcal{X}.$$

Combining any one of those BEs, discretized in time, with the following policy improvement:

$$\pi'(x) \in \arg\max_{u \in \mathcal{U}} \big[r(x,u) + \Delta t \cdot \nabla V_\pi(x) f_{\mathsf{c}}(x,u)\big] \ \forall x \in \mathcal{X},$$

where $\Delta t \cdot \nabla V_\pi$ replaces ∇v_π in (26), we can further obtain a partially model-free variant of the proposed PI methods. For example, a one-step IPI variant ($n = 1$) is shown in §5.2 (when the reward or initial VF is bounded). These variants are practically important since they contain neither $\dot v_i$ nor $\dot V_i$ (both of which depend on the full-dynamics f) nor the cumulative reward $\mathfrak{R}_\eta$ (which has been approximated out in the variants of IPI). As these variants are approximate versions of DPI and IPI, they also approximately satisfy the

[7] There are an infinite number of ways of choosing f_{d} and f_{c}; one typical choice is $f_{\mathsf{d}}(x) = f(x,0)$ and $f_{\mathsf{c}}(x,u) = f(x,u) - f_{\mathsf{d}}(x)$.

same properties as DPI and IPI shown in the subsequent sections.

4 Fundamental Properties of Policy Iterations

This section shows the fundamental properties of DPI and IPI — admissibility, the uniqueness of the solution to each policy evaluation, monotone improvement, and convergence (towards an HJB solution). We also discuss the optimality of the HJB solution (§§4.2 and E.1) based on the convergence properties of PIs. In any mathematical statements, $\langle v_i \rangle$ and $\langle \pi_i \rangle$ denote the sequences of the solutions to the BEs and the policies, both generated by Algorithm 1 or 2 under:

Boundary Condition. *If π_{i-1} is admissible, then*

$$\lim_{t \to \infty} \mathbb{G}^x_{\pi_{i-1}} \big[\gamma^t \cdot v_i(X_t) \big] = 0 \quad \forall x \in \mathcal{X}. \tag{29}$$

Theorem 4.1 *π_{i-1} is admissible and $v_i = v_{\pi_{i-1}}$ $\forall i \in \mathbb{N}$. Moreover, the policies are monotonically improved, that is,*

$$\pi_0 \preccurlyeq \pi_1 \preccurlyeq \cdots \preccurlyeq \pi_{i-1} \preccurlyeq \pi_i \preccurlyeq \cdots .$$

Theorem 4.2 (Convergence) *Denote $\hat{v}_*(x) \doteq \sup_{i \in \mathbb{N}} v_i(x)$. Then, $\hat{v}_*$ is lower semicontinuous; $v_i \to \hat{v}_*$* **a.** *pointwise;* **b.** *uniformly on $\Omega \subset \mathcal{X}$ if Ω is compact and $\hat{v}_*$ is continuous over Ω;* **c.** *locally uniformly if $\hat{v}_*$ is continuous.*

In what follows, $\hat{v}_*$ always denotes the converging function $\hat{v}_*(x) \doteq \sup_{i \in \mathbb{N}} v_i(x) = \lim_{i \to \infty} v_i(x)$ in Theorem 4.2.

4.1 Convergence towards v_* and π_*

Now, we provide convergence $v_i \to v_*$ to a solution v_* to the HJBE (17). One core technique is to use the PI operator $\mathcal{T} : \mathcal{V}_{\mathsf{a}} \to \mathcal{V}_{\mathsf{a}}$ defined on the space $\mathcal{V}_{\mathsf{a}}$ of admissible VFs as

$$\begin{cases} \mathcal{T} v_{\pi_{i-1}} \doteq v_{\pi_i} & \text{for any } i \in \mathbb{N}; \\ \mathcal{T} v_\pi \doteq v_{\pi'} & \text{for any other } v_\pi \in \mathcal{V}_{\mathsf{a}}, \end{cases}$$

where π' is a maximal policy over the given policy $\pi \in \Pi_{\mathsf{a}}$. Let $\mathcal{T}^N$ be the Nth recursion of $\mathcal{T}$ defined as $\mathcal{T}^0 v \doteq v$ and $\mathcal{T}^N v \doteq \mathcal{T}^{N-1}[\mathcal{T} v]$ for $v \in \mathcal{V}_{\mathsf{a}}$. Then, the VF sequence $\langle v_i \rangle$ satisfies $\mathcal{T}^N v_1 = v_{N+1}$ for all $N \in \mathbb{N}$.

In what follows, we denote v^* a (unique) fixed point of $\mathcal{T}$.

Proposition 4.3 *If v^* is a fixed point of $\mathcal{T}$, then $v^* = v_*$, i.e., v^* is a solution to the HJBE* (17)*.*

By Proposition 4.3, convergence $v_i \to v^*$ implies that $\langle v_i \rangle$ converges towards a solution v_* to the HJBE (17). In what follows, we first show the convergence $v_i \to v^*$ under:

Assumption 4.4 *$\mathcal{T}$ has a unique fixed point v^*.*

Theorem 4.5 *Under Assumption 4.4, there exists a metric $d : \mathcal{V}_{\mathsf{a}} \times \mathcal{V}_{\mathsf{a}} \to [0, \infty)$ such that $\mathcal{T}$ is a contraction (and thus continuous) under d and $v_i \to v^*$ in the metric d.*

Theorem 4.5 shows the convergence $v_i \to v^*$ in a metric d under which $\mathcal{T}$ is continuous. However, there is no information about which metric it is. In what follows, we focus on locally uniform convergence, in connection to Theorem 4.2. Let d_Ω be a pseudometric on $\mathcal{V}_{\mathsf{a}}$ defined for $\Omega \subseteq \mathcal{X}$ as

$$d_\Omega(v, w) \doteq \sup \big\{ \big| v(x) - w(x) \big| : x \in \Omega \big\} \text{ for } v,\ w \in \mathcal{V}_{\mathsf{a}}.$$

Then, uniform convergence $v_i \to v^*$ on Ω becomes equivalent to convergence $v_i \to v^*$ in the pseudometric d_Ω.

Theorem 4.6 *Suppose $\hat{v}_* \in \mathcal{V}_{\mathsf{a}}$ and for each compact subset Ω of $\mathcal{X}$, $\mathcal{T}$ is continuous under d_Ω. If Assumption 4.4 is true, then $v_i \to v^*$ locally uniformly and $v^* = \hat{v}_*$.*

The convergence condition in Theorem 4.6 comes from Leake and Liu (1967)'s approach that is now extended to our RL framework. The next theorem is motivated by the convergence results of PIs for optimal control of input-affine dynamics (Saridis and Lee, 1979; Beard et al., 1997; Murray et al., 2002; Abu-Khalaf and Lewis, 2005; Vrabie and Lewis, 2009) and provides the conditions for stronger convergence towards v_* and π_*.

Assumption 4.7 *For each $x \in \mathcal{X}$, the* argmax*-correspondence $p \mapsto \arg\max_{u \in \mathcal{U}} h(x, u, p)$ has a closed graph. That is, for each $x \in \mathcal{X}$ and any sequence $\langle p_k \rangle$ in $\mathcal{X}^\mathsf{T}$ converging to p_*,*

$$\begin{cases} u_k \in \arg\max\limits_{u \in \mathcal{U}} h(x, u, p_k) \\ \lim_{k \to \infty} u_k = u_* \in \mathcal{U} \end{cases} \Longrightarrow u_* \in \arg\max_{u \in \mathcal{U}} h(x, u, p_*).$$

Assumption 4.8 $\begin{cases} \textbf{a.}\ \langle \nabla v_i \rangle \text{ converges locally uniformly;} \\ \textbf{b.}\ \langle \pi_i \rangle \text{ converges pointwise.} \end{cases}$

Theorem 4.9 *Under Assumptions 4.7 and 4.8, $\hat{v}_*$ is a solution v_* to the HJBE* (17) *such that $v_* \in \mathrm{C}^1$ and*

(1) $v_i \to v_*$, $\nabla v_i \to \nabla v_*$ both locally uniformly;

(2) $\pi_i \to \pi_*$ pointwise, for a function π_* satisfying (18)*.*

Remark 4.10 *If the* argmax*-set is a singleton (so the maximal function u_* satisfying* (14) *is unique), then Assumption 4.7 is equivalent to the continuity of $p \mapsto u_*(x, p)$ for each $x \in \mathcal{X}$ and thus implied by the continuity of u_* (assumed in §2.3!) — for such examples, see §§5.1.1 and G.3. In this particular case, π_* in Theorem 4.9 is uniquely given by* (19) *and thus continuous (i.e., π_* is a policy).*

In summary, we have established the following convergence properties:

- (C1) convergence $v_i \to v_*$ in a metric;
- (C2) locally uniform convergence $v_i \to v_*$;
- (C3) locally uniform convergence $\nabla v_i \to \nabla v_*$, and pointwise convergence $\pi_i \to \pi_*$,

under certain conditions and the minimal assumptions made in this section and §2.

(Weak/Strong Convergence) Theorem 4.5 ensures weak convergence (C1) under Assumption 4.4 only; Theorem 4.6 gives strong convergence (C2) but with additional conditions — continuity of $\mathcal{T}$ in the uniform pseudometric d_Ω and convergence $\hat{v}_* \ (= \lim_{i\to\infty} v_i) \in \mathcal{V}_{\mathsf{a}}$. We note that

(1) the unique fixed point v^* therein and in Assumption 4.4 is a solution v_* to the HJBE (17) (Proposition 4.3);

(2) whenever (C2) is true, both v^* and v_* are characterized by Theorem 4.2 as $v^*(x) = v_*(x) = \sup_{i\in\mathbb{N}} v_i(x)$.

(Stronger Convergence) By Theorem 4.9, if a PI converges in a way described therein, then under Assumption 4.7, it does with stronger convergence properties (C2) and (C3) for $v_* = \hat{v}_* \in \mathrm{C}^1$, wherein the limit function $\hat{v}_* \ (= \lim_{i\to\infty} v_i)$ becomes a solution v_* to the HJBE (17). In this case,

(1) $\mathcal{T}$ is never used, hence no assumption is imposed on $\mathcal{T}$;

(2) the limit function π_* is not guaranteed to be a policy due to its possible discontinuity;

(3) the concave Hamiltonian formulation in §5.1 ensures $\pi_i \to \pi_*$ *locally uniformly* for a *policy* π_*, with both Assumptions 4.7 and 4.8b *relaxed* (e.g., Theorem 5.1).

4.2 Optimality: Sufficient Conditions

For each type of convergence above, we provide a sufficient condition for v_* in the HJBE (17) to be optimal in the sense that for any given initial admissible policy π_0, $v_i \to v_*$ in the respective manner with monotonicity $v_i \leqslant v_{i+1}\ \forall i \in \mathbb{N}$. For the optimality of v_* with the stronger convergence, (C2) and (C3), we additionally assume that:

Assumption 4.11 *The solution v_* to the HJBE* (17)*, if exists, is unique over* C^1 *and upper-bounded (by zero if $\gamma = 1$).*

Due to space limitation, those sufficient conditions for optimality and related discussions are presented in §E.1.

5 Case Studies

With strong connections to RL and optimal control in CTS, this section studies the special cases of the general RL problem formulated in §2. In those case studies, the proposed PI methods and theory for them are simplified and improved as summarized in Table 1. The blanks in Table 1 are filled with "Assumed" or, in simplified policy improvement sections, "No." The connections to stability theory in optimal control are also made in this section. The optimality of the HJB solution (v_*, π_*) for each case is studied and summarized in §E.2; more case studies are given in §G.

For simplicity, we let $f^x(u) \doteq f(x,u)$ and $r^x(u) \doteq r(x,u)$ for $x \in \mathcal{X}$. Both f^x and r^x are continuous for each x since so are f and r. The mathematical terminologies employed in this section are given in §B, with a summary of notations.

5.1 Concave Hamiltonian Formulations

Here, we study the special settings of the reward function r, which make the function $u \mapsto h(x,u,p)$ strictly concave and C^1 (after some input-transformation in the cases of non-affine dynamics). In these cases, policy improvement maximizations (14), (15), and (18) become convex optimizations whose solutions exist and are given in closed-forms. We will see that this dramatically simplifies the policy improvement itself and strengthen the convergence properties. Although we focus on certain classes of dynamics — the input-affine and then a class of non-affine ones — the idea is extendible to a general nonlinear system of the form (1) (see §G.1 for such an extension).

5.1.1 Case I: Input-affine Dynamics

First, consider the following case: for each $x \in \mathcal{X}$,

(1) f^x *is affine*, i.e., the input-coupling term $f_{\mathsf{c}}(x,u)$ in the decomposition (25) is linear in u, so that the dynamics f can be represented for a matrix-valued continuous function $F_{\mathsf{c}} : \mathcal{X} \to \mathbb{R}^{l\times m}$ as

$$f(x,u) = f_{\mathsf{d}}(x) + F_{\mathsf{c}}(x)u; \tag{30}$$

(2) r^x *is strictly concave and represented by*

$$r(x,u) = \mathfrak{r}(x) - \mathfrak{c}(u) \tag{31}$$

for a continuous function $\mathfrak{r} : \mathcal{X} \to \mathbb{R}$ and a *strictly convex* C^1 function $\mathfrak{c} : \mathcal{U} \to \mathbb{R}$ whose gradient $\nabla\mathfrak{c}$ is *surjective*, i.e., $\nabla\mathfrak{c}(\mathcal{U}^{\mathrm{o}}) = \mathbb{R}^{1\times m}$. Here, $\mathcal{U}^{\mathrm{o}}$ is the interior of $\mathcal{U}$; both $\mathfrak{r}$ and $-\mathfrak{c}$ are assumed to have their respective maximums.

This framework includes those in (Rekasius, 1964; Beard et al., 1997; Doya, 2000; Abu-Khalaf and Lewis, 2005; Vrabie and Lewis, 2009; Lee, Park, and Choi, 2015) as special cases; it still contains a broad class of dynamics such as Newtonian dynamics (e.g., robot manipulator and vehicle models). In this case, the function $u \mapsto h(x,u,p)$ is strictly concave and C^1 (see the definition (5)). Hence, as mentioned in §2.3 (see §D for the behind theory), the unique maximal function $u_* \equiv u_*(x,p)$ satisfying (14) corresponds to the unique regular point $\bar{u} \in \mathcal{U}^{\circ}$ s.t. $-\nabla\mathfrak{c}(\bar{u}) + p\,F_{\mathsf{c}}(x) = 0$, where the gradient $\nabla\mathfrak{c}^{\mathsf{T}} : \mathcal{U}^{\mathrm{o}} \to \mathbb{R}^m$ is strictly monotone and bijective on its domain $\mathcal{U}^{\mathrm{o}}$ (see §I.3). Rearranging it w.r.t. $\bar{u}$, we obtain the closed-form solution u_* of (14):

$$u_*(x,p) = \sigma\big(F_{\mathsf{c}}^{\mathsf{T}}(x)\,p^{\mathsf{T}}\big), \tag{32}$$

where $\sigma : \mathbb{R}^m \to \mathcal{U}^{\mathrm{o}}$, defined as the inverse of $\nabla\mathfrak{c}^{\mathsf{T}}$, i.e., $\sigma \doteq (\nabla\mathfrak{c}^{\mathsf{T}})^{-1}$, is also strictly monotone and continuous (see §I.3); thus, u_* is continuous. Using this *unique* expression (32), we obtain the unique closed-form solution (16) of the policy improvement maximization (15) (or (26)) as

$$\pi'(x) = \sigma\big(F_{\mathsf{c}}^{\mathsf{T}}(x)\nabla v_\pi^{\mathsf{T}}(x)\big) \tag{33}$$

a.k.a. the value-gradient-based (VGB) greedy policy update (Doya, 2000). This simplifies the policy improvement of DPI

Table 1
Summary of Case Studies: Relaxations and Simplifications of the Assumptions and Policy Improvement

Problem Formulation	Concave Hamiltonian	Discounted RL with bounded		RL with local Lipschitzness[b]	Nonlinear optimal control[b]	LQR
		VF[a]	state trj.			
Section	5.1 / G.1	5.2	G.2	5.3	5.4	G.3
Global existence and uniqueness of the state trj.			True, conditionally[c]			True
Existence of an admissible policy, i.e., $\Pi_{\mathsf{a}} \neq \varnothing$		True				
C^1-regularity (3) and continuity of admissible VFs		Continuous, conditionally[b]				
Assumptions 4.4 and 4.11 (w.r.t. $\mathcal{T}$ and the HJBE)						
Existence of a continuous maximal function u_* and Assumption 4.7	True					
Boundary conditions (12) and (29)		True, conditionally[d]	True		True, conditionally[e]	
Conditions for $v_i \to v_*$, $\nabla v_i \to \nabla v_*$, $\pi_i \to \pi_*$	Relaxed[f]					
Simplified policy improvement	Yes					Yes

[a] Once the initial VF v_{π_0} in the PI methods is bounded, so is v_{π_i} for all $i \in \mathbb{N}$; a stronger case is when the reward function r is bounded.
[b] f and/or f_π is assumed locally Lipschitz.
[c] True if f_π is locally Lipschitz in §G.2 and in addition, in §§5.3 and 5.4, if $\pi \in \Pi_{\mathsf{a}}$ (see the modified definitions of Π_{a} therein).
[d] True if v and v_i are bounded — this makes sense only when the target VF is bounded.
[e] True if i) the system (1) under π is globally asymptotically stable or ii) $\exists \kappa > 0$ s.t. $r_\pi \leqslant \kappa \cdot v$ holds (see also §C).
[f] Assumptions 4.7 and 4.8 are reduced to Assumption 4.8b (see Theorems 4.9 and 5.1).

and IPI (and their variants) shown in §3 as

Policy Improvement: update the next policy π_i by

$$\pi_i(x) = \sigma\big(F_{\mathsf{c}}^{\mathsf{T}}(x) \nabla v_i^{\mathsf{T}}(x)\big). \tag{34}$$

Similarly, the HJB policy π_* satisfying (18) is also *uniquely* given by (19) and (32), i.e., $\pi_*(x) = \sigma\big(F_{\mathsf{c}}^{\mathsf{T}}(x) \nabla v_*^{\mathsf{T}}(x)\big)$, under (30) and (31). Moreover, Theorem 4.9 can be simplified and strengthened by relaxing the assumptions on the policies and policy improvement as follows.

Theorem 5.1 *Under* (30)*,* (31)*, and Assumption 4.8a, $\hat{v}_*$ is a solution v_* to the HJBE* (17) *such that $v_* \in \mathrm{C}^1$ and $v_i \to v_*, \nabla v_i \to \nabla v_*$, and $\pi_i \to \pi_*$, all locally uniformly.*

Remark 5.2 *Assumption 4.8a is necessary for convergence in Theorem 5.1 and, in fact so are similar uniform convergence assumptions on $\langle \nabla v_i \rangle$ for convergence given in the existing literature on PIs for optimal control (e.g., Saridis and Lee, 1979; Beard et al., 1997; Murray, Cox, and Saeks, 2003; Abu-Khalaf and Lewis, 2005; Bian et al., 2014 to name a few). This is due to the fact that even the uniform convergence of v_i (e.g., Theorem 4.2c) implies nothing about the convergence of its gradient ∇v_i; it cannot even ensure the differentiability of the limit function $\hat{v}_*$ (Rudin, 1964; Thomson, Bruckner, and Bruckner, 2001). Here, Assumption 4.8a or any type of uniform convergence of $\langle \nabla v_i \rangle$ is by no means trivial to prove, and thus its relaxation remains as a future work (even in the optimal control frameworks in the existing literature, which are similar to that in §5.4 under* (30)–(31)*, to the best authors' knowledge).*

One way to effectively take the input constraints into considerations is to construct the action space $\mathcal{U}$ as

$$\mathcal{U} = \big\{u \in \mathbb{R}^m : |u_j| \leq u_{\mathsf{max},j},\ 1 \leq j \leq m\big\},$$

where $u_j \in \mathbb{R}$ is the j-th element of u, and $u_{\mathsf{max},j} \in (0, \infty]$ is the corresponding physical constraint. In this case, $\mathfrak{c}$ in (31) can be chosen as

$$\mathfrak{c}(u) = \lim_{v \to u} \int_0^v (s^{\mathsf{T}})^{-1}(\mathfrak{u}) \cdot \Gamma \, d\mathfrak{u} \tag{35}$$

for a positive definite matrix $\Gamma \in \mathbb{R}^{m \times m}$ and a continuous function $s : \mathbb{R}^m \to \mathcal{U}^{\mathrm{o}}$ that is strictly monotone, odd, and bijective and makes $\mathfrak{c}(u)$ in (35) finite at any point u on the boundary $\partial\mathcal{U}$;[8] This formulation gives the closed-form expression $\sigma(\mathfrak{u}) = (\nabla \mathfrak{c}^{\mathsf{T}})^{-1}(\mathfrak{u}) = s(\Gamma^{-1}\mathfrak{u})$ and includes the sigmoidal examples (Cases 1 and 2) in §6 as special cases — see also (Doya, 2000; Abu-Khalaf and Lewis, 2005) for similar sigmoidal examples. Another well-known example is the unconstrained problem:

$$\mathcal{U} = \mathbb{R}^m \ \ (u_{\mathsf{max},j} = \infty \text{ for each } j) \text{ and } s(\mathfrak{u}) = \mathfrak{u}/2, \tag{36}$$

by which (35) becomes $\mathfrak{c}(u) = u^{\mathsf{T}} \Gamma u$; the LQR case in §G.3 with $E = 0$ shows such an example.

Remark 5.3 *Once r^x is strictly concave for each $x \in \mathcal{X}$, the reward function r can be always represented as*

$$r(x, u) = \mathfrak{r}(x) - \mathfrak{c}(x, u), \tag{37}$$

where $\mathfrak{r}$ and $\mathfrak{c}$ are continuous and have a maximum and a minimum, respectively; for each $x \in \mathcal{X}$, $\mathfrak{c}^x \doteq \mathfrak{c}(x, \cdot)$ is strictly convex. In this general case, if $\mathfrak{c}^x$ for each $x \in \mathcal{X}$

[8] $\partial\mathcal{U} = \{u \in \mathbb{R}^m : u_j = u_{\mathsf{max},j} \text{ for some } j = 1, 2, \cdots, m\}$.

satisfies the same properties as $\mathfrak{c}$ *in* (31)*, then the unique maximal function* u_* *and the maximal policy* π' *over* $\pi \in \Pi_{\mathsf{a}}$ *can be obtained in the same way to* (32) *and* (33) *as*

$$\begin{cases} u_*(x,p) = \sigma^x\big(F_{\mathsf{c}}^{\mathsf{T}}(x)\, p^{\mathsf{T}}\big) \\ \pi'(x) = \sigma^x\big(F_{\mathsf{c}}^{\mathsf{T}}(x) \nabla v_\pi^{\mathsf{T}}(x)\big) \end{cases} \tag{38}$$

for $\sigma^x \doteq ((\nabla \mathfrak{c}^x)^{\mathsf{T}})^{-1}$*. In addition, if* $(x,\mathfrak{u}) \mapsto \sigma^x(\mathfrak{u})$ *is continuous, then Theorem 5.1 (specifically, Lemma I.7 in §I.3) can be generalized with* σ *replaced by* σ^x*. Some examples of such* σ^x *are as follows.*

(1) Γ *in* (35) *is a continuous function over* $\mathcal{X}$*. In this case,* σ^x *is given by* $\sigma^x(\mathfrak{u}) = s(\Gamma^{-1}(x)\cdot \mathfrak{u})$*.*

(2) In the LQR setting (§G.3), $\sigma^x(\mathfrak{u}) = \Gamma^{-1}(\mathfrak{u}/2 - E^{\mathsf{T}}x)$ *and whenever* $E = 0$*,* $\sigma^x(\mathfrak{u}) = \sigma(\mathfrak{u}) = \Gamma^{-1}\mathfrak{u}/2$*.*

5.1.2 Case II: a Class of Non-affine Dynamics

If f^x is not affine, then the choice of the reward function r is critical. Provided in §F is such an example, where a choice of r in the form of (31) and (35) fails to give closed-form solutions to policy improvement and the HJBE (17); such a choice of r results in a pathological Hamiltonian h in the unconstrained case (see §F for details).

Such a pathological case and difficulty, on the other hand, can be avoided for the non-affine dynamics f of the form:

$$f(x,u) = f_{\mathsf{d}}(x) + F_{\mathsf{c}}(x)\varphi(u), \tag{39}$$

where $\varphi : \mathcal{U} \to \mathcal{A} \subseteq \mathbb{R}^m$ is a continuous function from the action space $\mathcal{U}$ to another action space $\mathcal{A}$ and has its inverse $\varphi^{-1} : \mathcal{A}^{\mathrm{o}} \to \mathcal{U}^{\mathrm{o}}$ between the interiors. Note that (39) corresponds to the decomposition (25) with the input-coupling part $f_{\mathsf{c}}(x,u) = F_{\mathsf{c}}(x)\varphi(u)$ and includes the input-affine dynamics (30) as a special case $\varphi(u) = u$ and $\mathcal{A} = \mathcal{U}$.

Motivated by Kiumarsi, Kang, and Lewis (2016), we propose to set the reward function r under (39) as

$$r(x,u) = \mathfrak{r}(x) - \mathfrak{c}(\varphi(u)), \tag{40}$$

where $\mathfrak{r} : \mathcal{X} \to \mathbb{R}$ and $\mathfrak{c} : \mathcal{A} \to \mathbb{R}$ are functions that satisfy the properties of $\mathfrak{r}$ and $\mathfrak{c}$ in (31) but w.r.t. the action space $\mathcal{A}$ in place of $\mathcal{U}$. Under (39) and (40), the proposed PIs have the following properties, extended from §5.1.1 (e.g., Theorem 5.1), although the argmax-set "$\arg\max_{u\in\mathcal{U}} h(x,u,p)$" in this case may not be a singleton (another maximizer may exist on the boundary $\partial\mathcal{U}$). For notational simplicity, we denote $\tilde{\sigma}(\mathfrak{u}) \doteq \varphi^{-1}[\sigma(\mathfrak{u})]$ in the theorem below.

Theorem 5.4 *Under* (39) *and* (40)*,* **a.** *a maximal policy* π' *over* $\pi \in \Pi_{\mathsf{a}}$ *is given by* $\pi'(x) = \tilde{\sigma}\big(F_{\mathsf{c}}^{\mathsf{T}}(x)\nabla v_\pi^{\mathsf{T}}(x)\big)$ *explicitly;* **b.** *if the policies are updated in policy improvement by*

$$\pi_i(x) = \tilde{\sigma}\big(F_{\mathsf{c}}^{\mathsf{T}}(x)\nabla v_i^{\mathsf{T}}(x)\big), \tag{41}$$

then under Assumption 4.8a, $\hat{v}_*$ *is a solution* v_* *to the HJBE* (17) *s.t.* $v_* \in \mathrm{C}^1$ *and* $v_i \to v_*$*,* $\nabla v_i \to \nabla v_*$*, and* $\pi_i \to \pi_*$*, all locally uniformly, where* $\pi_*(x) = \tilde{\sigma}\big(F_{\mathsf{c}}^{\mathsf{T}}(x)\nabla v_*^{\mathsf{T}}(x)\big)$*.*

Similarly to Remark 5.3, the results are extendible to the general case where φ and/or $\mathfrak{c}$ depend on the state $x \in \mathcal{X}$.

5.2 Discounted RL with Bounded VF

Boundedness of a VF is stronger than admissibility. Likewise, when discounted, a bounded VF has stronger properties than admissible ones. One example is continuity stated in the next proposition; the extension to the general cases ($\gamma = 1$ and/or $v_\pi \in \mathcal{V}_{\mathsf{a}}$) is by no means trivial.

Proposition 5.5 *Suppose that* f_π *is locally Lipschitz and that* $\gamma \in (0,1)$*. Then,* v_π *is continuous if* v_π *is bounded.*

Continuity is a necessary condition to be C^1. In the RL problem formulation in §2, we have assumed the C^1-regularity (3) and thereby continuity on every admissible VF, but no proof was provided regarding them; Proposition 5.5 above bridges this gap when the VF is discounted and bounded. In this case, the boundary condition (12) is also true.

Proposition 5.6 *If* $v : \mathcal{X} \to \mathbb{R}$ *is bounded and* $\gamma \in (0,1)$*, then* v *satisfies the boundary condition* (12) *for any policy* π*.*

Moreover, when the VF is discounted and bounded, the BE (10) (resp. (11)) has the unique solution $v = v_\pi$ over all bounded (resp. bounded C^1) functions, and the boundedness is preserved under the policy improvement operation.

Corollary 5.7 *Let* $\gamma \in (0,1)$ *and* π *be a policy. Then,*

(1) if there exists a bounded function v *satisfying the integral BE* (10) *or with* $v \in \mathrm{C}^1$*, the differential BE* (11)*, then* v_π *is bounded (hence, admissible) and* $v = v_\pi$*.*

(2) if v_π *is bounded (hence, admissible), then so is* $v_{\pi'}$ *and we have* $\pi \preccurlyeq \pi'$*, where* π' *is a maximal policy over* π*.*

Algorithm 3: Variants of IPI and DPI with Bounded v_{π_0}

Initialize: π_0, an initial policy s.t. v_{π_0} is bounded; $\Delta t > 0$, a small time step ($0 < \Delta t \ll 1$); $i \leftarrow 1$;

repeat (under $\gamma \in (0,1)$)

  Policy Evaluation: given policy π_{i-1}, find a *bounded* C^1 function $V_i : \mathcal{X} \to \mathbb{R}$ such that for all $x \in \mathcal{X}$,

  (IPI Variant): $V_i(x) \approx \mathbb{G}^x_{\pi_{i-1}}\big[R_0 + \gamma_{\mathsf{d}} \cdot V_i(X_{\Delta t})\big]$;

  (DPI Variant): for $\alpha_{\mathsf{d}} \doteq \alpha \Delta t\ (= -\ln \gamma_{\mathsf{d}})$,

$$\alpha_{\mathsf{d}} \cdot V_i(x) = h\big(x, \pi_{i-1}(x), \Delta t \cdot \nabla V_i(x)\big);$$

  Policy Improvement: find a policy π_i s.t. for all $x \in \mathcal{X}$,

$$\pi_i(x) \in \arg\max_{u\in\mathcal{U}} \big[\, r(x,u) + \Delta t \cdot \nabla V_i(x) f_{\mathsf{c}}(x,u)\big];$$

  $i \leftarrow i + 1$;

until *convergence is met.*

In fact, if the reward function r is bounded, then so is the VF for *any* given policy (so long as the state trajectory $t \mapsto X_t$ exists); hence the above results become stronger as follows.

Assumption 5.8 *r is bounded and $\gamma \in (0, 1)$.*

Corollary 5.9 *Under Assumption 5.8, the followings hold for any given policy π and any maximal policy π' over π:*

(1) v_π and $v_{\pi'}$ are bounded (hence, admissible); $\pi \preccurlyeq \pi'$;

(2) v_π is continuous if f_π is locally Lipschitz;

(3) if a bounded function v satisfies the integral BE (10) *or with $v \in \mathrm{C}^1$, the differential BE* (11)*, then $v = v_\pi$.*

For a given policy π, the VF properties in Corollary 5.9 are also true when r_π (but not necessarily r) is bounded (see the proof of Corollary 5.9 in §I.3). In this case (and the general cases where v_π is bounded somehow), Proposition 5.5, Corollary 5.7, and mathematical induction show that $\mathcal{T}^N v_\pi$ for any N satisfies the properties of the VFs in Corollary 5.9. In other words, if for the initial policy π_0,

Assumption. *r_{π_0} (or the VF v_{π_0}) is bounded and $\gamma \in (0, 1)$*

which is weaker than Assumption 5.8, then the sequences $\langle v_i \rangle$ and $\langle \pi_i \rangle$ generated by DPI or IPI satisfy: for any $i \in \mathbb{N}$,

(1) $v_i = v_{\pi_{i-1}}$,

(2) $v_{\pi_{i-1}}$ is bounded and $\pi_{i-1} \preccurlyeq \pi_i$,

(3) $v_{\pi_{i-1}}$ is continuous if $f_{\pi_{i-1}}$ is locally Lipschitz,

under the boundedness of each v_i to ensure the boundary condition (29) to be true by Proposition 5.6.

Algorithm 3 shows the respective variants of IPI and DPI when $\gamma \in (0, 1)$ and the VF v_{π_0} w.r.t. the initial policy π_0 is bounded. Here, the boundedness of v_{π_0} can be made by that of r or r_{π_0}. In the policy evaluation, the variants of IPI and DPI solve, for $V_i \doteq v_i/\Delta t$, the discretized BE (28) and the differential BE (20), respectively; the other steps of both variants are same and derived from their originals (Algorithms 1 and 2) by replacing η and v_i with the small given time step Δt and $\Delta t \cdot V_i$, respectively. Implementation examples of both variants in Algorithm 3 are given and discussed in §6 with several types of bounded reward functions r and a function approximator for V_i. The other types of variants (e.g., IPI with the n-step prediction (27)) can be also obtained by replacing the BE in policy evaluation with one of the other BEs in §3 (e.g., (27)). Since these variants all assume both $\gamma \in (0, 1)$ and the boundedness of the initial VF v_{π_0}, it is sufficient to find a *bounded* C^1 function V_i in each policy evaluation (line 3) to have the properties above regarding $\langle v_i \rangle$ and $\langle \pi_i \rangle$, without assuming the boundary condition (29) on v_i ($= \Delta t \cdot V_i$).

5.3 RL with Local Lipschitzness

Let $\begin{cases} \Pi_{\mathsf{Lip}} \doteq \text{the set of all } \textit{locally Lipschitz} \text{ policies}, \\ \mathrm{C}^1_{\mathsf{Lip}} \doteq \{v \in \mathrm{C}^1 : \nabla v \text{ is locally Lipschitz}\}. \end{cases}$

In §§5.3 and 5.4, we consider the RL problems, where

Assumption. *The dynamics f and the maximal function u_* in* (14) *are locally Lipschitz,*

and always use the notations π' and π_* to denote the maximal and HJB policies *given by* (16) *and* (19), respectively.

The Assumption implies continuity of f and u_* and ensures:

(1) π' and π_* are locally Lipschitz (i.e., $\pi', \pi_* \in \Pi_{\mathsf{Lip}}$) so long as v_π in (16) and v_* in (19) are $\mathrm{C}^1_{\mathsf{Lip}}$, respectively;

(2) the dynamics f_π under $\pi \in \Pi_{\mathsf{Lip}}$ is locally Lipschitz, and thereby the state trajectory $t \mapsto \mathbb{G}^x_\pi[X_t]$ for each $x \in \mathcal{X}$ is uniquely defined over the maximal existence interval $[0, t_{\mathsf{max}}(x; \pi)) \subseteq \mathbb{T}$ (Khalil, 2002, Theorem 3.1).

Here, $t_{\mathsf{max}}(x; \pi) \in (0, \infty]$ is defined for and depends on both initial state x and $\pi \in \Pi_{\mathsf{Lip}}$; whenever $t_{\mathsf{max}}(x; \pi) < \infty$,

$$\mathbb{G}^x_\pi(\|X_t\|) \to \infty \text{ as } t \to t_{\mathsf{max}}(x; \pi)$$

if the limit exists (e.g., see Khalil, 2002, Example 3.3). To circumvent this finite-time explosion issue, we set $v_\pi(x)$ to "$-\infty$" whenever $t_{\mathsf{max}}(x; \pi)$ is finite, that is, redefine v_π as

$$v_\pi(x) \doteq \begin{cases} \mathbb{G}^x_\pi\left[\int_0^\infty \gamma^t \cdot R_t \, dt\right] & \text{if } t_{\mathsf{max}}(x; \pi) = \infty, \\ -\infty, \quad \text{otherwise.} \end{cases} \tag{42}$$

Here, existence and uniqueness of the state trajectories were not assumed; $t_{\mathsf{max}}(\,\cdot\,; \pi)$ and thus v_π in (42) are well-defined as long as $\pi \in \Pi_{\mathsf{Lip}}$. Hence, with slight abuse of notation, we restrict the admissible sets Π_{a} and $\mathcal{V}_{\mathsf{a}}$ by redefining them as

$$\Pi_{\mathsf{a}} \doteq \{\pi \in \Pi_{\mathsf{Lip}} : v_\pi(x) \text{ is finite for all } x \in \mathcal{X}\},$$
$$\mathcal{V}_{\mathsf{a}} \doteq \{v_\pi : \pi \in \Pi_{\mathsf{a}}\}.$$

Note that for each $x \in \mathcal{X}$, the value $v_\pi(x)$ is finite and the state trajectory $t \mapsto \mathbb{G}^x_\pi[X_t]$ is defined uniquely over $\mathbb{T}$ if $\pi \in \Pi_{\mathsf{a}}$ or $v_\pi \in \mathcal{V}_{\mathsf{a}}$. This global existence of the unique state trajectories was assumed in the general RL problem formulated in §2 but now is encapsulated by admissibility.

In what follows, we provide the policy improvement theorem extended from Theorem 2.7, without assuming any existence and uniqueness of the state trajectories, but under

Assumption. $\mathcal{V}_{\mathsf{a}} \subset \mathrm{C}^1_{\mathsf{Lip}}$

to ensure the maximal policy $\pi' \in \Pi_{\mathsf{Lip}}$ whenever $v_\pi \in \mathcal{V}_{\mathsf{a}}$.

Theorem 5.10 (Policy Improvement) *If there exist a compact subset $\Omega \subset \mathcal{X}$ and $\mathcal{K}_\infty$ functions ρ_1, ρ_2 such that for a policy $\pi \in \Pi_{\mathsf{a}}$,*

$$\rho_1(\|x\|_\Omega) \le \overline{v} - v_\pi(x) \le \rho_2(\|x\|_\Omega) \quad \forall x \in \mathcal{X},$$

where $\|x\|_\Omega \doteq \inf_{y \in \Omega} \|x - y\|$, then $\pi' \in \Pi_{\mathsf{a}}$ and $\pi \preccurlyeq \pi'$.

5.4 Nonlinear Optimal Control

The objective of optimal control is to stabilize the system (1) w.r.t. a given equilibrium point $(x_\mathsf{e}, u_\mathsf{e})$ while minimizing a given cost functional. Here, any point in $\mathcal{X} \times \mathcal{U}$ such that $\dot{x}_\mathsf{e} = f(x_\mathsf{e}, u_\mathsf{e}) \equiv 0$ is called an equilibrium point $(x_\mathsf{e}, u_\mathsf{e})$; it can be transformed to $(0,0)$ and thus let $(x_\mathsf{e}, u_\mathsf{e}) = (0,0)$ without loss of generality (Khalil, 2002) and assume that $f(0,0) = 0$. Note that if a policy π satisfies $\pi(0) = 0$, then we have $0 = f_\pi(0)$, i.e., $x_\mathsf{e} = 0$ is an equilibrium point of the system (1) under π.

The optimal control framework in this subsection is a particular case of the locally Lipschitz RL problem in §5.3 above. Hence, we impose the same assumptions on it: the local Lipschitzness of f and u_*, with π' and π_* denoting the respective policies *given by* (16) *and* (19), the inclusion $\mathcal{V}_\mathsf{a} \subset \mathrm{C}^1_\mathsf{Lip}$, and the extended definition (42) of the VF v_π for $\pi \in \Pi_\mathsf{Lip}$.

On the other hand, we define a class of policies Π_0 as

$$\Pi_0 \doteq \{\pi \in \Pi_\mathsf{Lip} : \pi(0) = 0\}$$

and, with slight abuse of notation, redefine Π_a and $\mathcal{V}_\mathsf{a}$ by

$$\Pi_\mathsf{a} \doteq \{\pi \in \Pi_0 : v_\pi(x) \text{ is finite for all } x \in \mathcal{X}\}$$

and $\mathcal{V}_\mathsf{a} \doteq \{v_\pi : \pi \in \Pi_\mathsf{a}\}$. Here, we have merely added the condition $\pi(0) = 0$ into the definitions of Π_a and $\mathcal{V}_\mathsf{a}$ in §5.3. With these notations, $x_\mathsf{e} = 0$ comes to be an equilibrium point of the system (1) under $\pi \in \Pi_0$ ($\supset \Pi_\mathsf{a}$).

Similarly to §5.3, this subsection does not assume existence and uniqueness of the state trajectories; $\pi \in \Pi_\mathsf{a}$ ensures global existence of the unique state trajectories under π. In addition, the boundary conditions (12) and (29) are not assumed but either proven or replaced by those corresponding to Theorem C.1 in §C (e.g., see Theorem 5.16).

Whenever necessary, we use the cost functions $c \doteq -r$ and $c_\pi \doteq -r_\pi$, the cost VF $J_\pi \doteq -v_\pi$, and $J_* \doteq -v_*$, rather than $-r$, $-r_\pi$, $-v_\pi$, and $-v_*$, respectively, for simplicity and consistency to optimal control; the cost at time $t \in \mathbb{T}$ is denoted by $C_t \doteq c(X_t, U_t) = -R_t$.

We consider a positive definite cost function c, i.e., assume

$$c(x,u) > 0 \quad \forall (x,u) \neq (0,0), \text{ and } c(0,0) = 0. \tag{43}$$

Then, by (43) and the definition, the value $J_\pi(x)$ is always restricted to $[0,\infty]$ and, similarly to (42), $J_\pi(x) = \infty$ whenever $t_\mathsf{max}(x;\pi) < \infty$; otherwise, $J_\pi(x) = \mathbb{G}^x_\pi\big[\int_0^\infty \gamma^t C_t \, dt\big]$.

Lemma 5.11 *c_π for $\pi \in \Pi_0$ is positive definite.*

Lemma 5.12 *Let $\pi \in \Pi_\mathsf{a}$. Then,* **a.** *J_π is positive definite;* **b.** *$x \mapsto \dot{J}_\pi(x, \pi(x))$ is negative semidefinite iff*

$$\alpha J_\pi \leqslant c_\pi \tag{44}$$

and **c.** *$x \mapsto \dot{J}_\pi(x, \pi(x))$ is negative definite iff*

$$\alpha J_\pi(x) < c_\pi(x) \qquad \forall x \in \mathcal{X} \setminus \{0\}. \tag{45}$$

In what follows, we assume that c_π for any $\pi \in \Pi_0$ does not radially converge to zero, i.e.,

$$c_\pi(x) \not\to 0 \text{ whenever } \|x\| \to \infty. \tag{46}$$

Given the conditions in Lemma 5.12, J_π is, in fact, a Lyapunov function for the dynamics $\dot{X}_t = f_\pi(X_t)$ (Khalil, 2002), as shown in the Lyapunov stability theorem below.

Theorem 5.13 *The equilibrium point $x_\mathsf{e} = 0$ of dynamics f_π under $\pi \in \Pi_\mathsf{a}$ is stable if* (44) *holds, asymptotically stable if* (45) *is true, and globally asymptotically stable if $\gamma = 1$ or, in addition to* (45)*, J_π is radially unbounded.*

Remark 5.14 *Whenever $\gamma = 1$,* (45) *is true since $\alpha = 0$ and c_π is positive definite by Lemma 5.11. Hence, admissibility directly implies global asymptotic stability by Theorem 5.13 (here, the radial unboundedness of J_π is not assumed!).*

Next, we show that global asymptotic stability ensures the uniqueness of the solution to the BEs.

Theorem 5.15 (Policy Evaluation)
Let $x_\mathsf{e} = 0$ under $\pi \in \Pi_0$ be globally asymptotically stable. If there exists a function $v : \mathcal{X} \to \mathbb{R}$ s.t. v is continuous at 0, $v(0) = 0$, and the BE (10) *or, with $v \in \mathrm{C}^1$,* (11) *holds, then $\pi \in \Pi_\mathsf{a}$ and $v = v_\pi$.*

The uniqueness of the solution to the BE can be also given under another condition than stability. When discounted, it is more general than the stability conditions (44) and (45) as well as contains the cases where $x_\mathsf{e} = 0$ is not necessarily stable, and the state trajectories could even diverge. For the statement and further discussions, we define

$$\Pi \doteq \{\pi \in \Pi_0 : t_\mathsf{max}(x;\pi) = \infty \text{ for all } x \in \mathcal{X}\}$$

which satisfies the chain of inclusions: $\Pi_0 \subset \Pi \subset \Pi_\mathsf{a}$.

Theorem 5.16 (Policy Evaluation) *Suppose that there exists $v \in \mathrm{C}^1$ satisfying the BE* (10) *or* (11) *for a policy $\pi \in \Pi$. Then, $\pi \in \Pi_\mathsf{a}$ and $v = v_\pi$ if $J \doteq -v$ is positive definite and $\kappa J \leqslant c_\pi$ for some $\kappa > 0$.*

The policy improvement theorem in §5.3, i.e., Theorem 5.10, can be also extended as follows.

Theorem 5.17 (Policy Improvement) *Let $\pi \in \Pi_\mathsf{a}$ and J_π be radially unbounded. Then, $\pi' \in \Pi_\mathsf{a}$ and $J_{\pi'} \leqslant J_\pi$.*

From the theory and discussions above, we propose the following three conditions for the PI methods in the optimal control framework: for all $i \in \mathbb{N}$ and $J_i \doteq -v_i$,

(1) $\pi_0 \in \Pi_\mathsf{a}$,
(2) $J_i \in \mathrm{C}^1_\mathsf{Lip}$ is positive definite and radially unbounded,
(3) there exists $\kappa_i > 0$ such that $\alpha\kappa_i \cdot J_i \leqslant c_{\pi_{i-1}}$. (47)

Those three conditions are devised in order to run the PI methods without assuming the existence of unique state trajectories and the boundary condition (29). In the second one,

the positive definiteness of J_i is required for policy evaluation, and its radial unboundedness is for policy improvement. The third one (47) is always true for $\gamma = 1$, but when discounted ($\alpha > 0$), it is just a different representation of the inequality in Theorem 5.16 and thus required to be true for policy evaluation. Note that when $\kappa_i \in (0, 1)$, (47) is weaker than both of the stability conditions $\alpha J_i \leqslant c_{\pi_{i-1}}$ and

$$\alpha J_i(x) < c_{\pi_{i-1}}(x) \qquad \forall x \neq \mathcal{X} \setminus \{0\} \tag{48}$$

corresponding to (44) and (45), respectively.

Theorem 5.18 *Under the three conditions above, $\pi_{i-1} \in \Pi_{\mathsf{a}}$ and $J_i = J_{\pi_{i-1}} \geqslant J_{\pi_i}$ $\forall i \in \mathbb{N}$. Moreover, $x_{\mathsf{e}} = 0$ is globally asymptotically stable under π_{i-1} if* (48) *is true (or if $\gamma = 1$).*

Without assuming the boundary condition (29) and the existence of unique state trajectories, the other properties in §4 can be also extended under the above three conditions, by following the same proofs in §4, but with Theorem 4.1 therein replaced by Theorem 5.18.

Once the optimal cost VF J_* is known *a priori*, the radial unboundedness of each J_i can be replaced by that of J_*. In this case, every J_π for $\pi \in \Pi_{\mathsf{a}}$ is radially unbounded by optimality $0 \leqslant J_* \leqslant J_\pi$ $\forall \pi \in \Pi_{\mathsf{a}}$. Hence, J_i becomes radially unbounded if $J_i \in \mathcal{V}_{\mathsf{a}}$. We note that in the middle of the proof of Theorem 5.18 (see §I.3), $J_i \in \mathcal{V}_{\mathsf{a}}$ is proven before the radial unboundedness of J_i is applied.

Limitations also exist. First, it is difficult to check the radial unboundedness and (47); J_* is unknown until we have found at the end. Secondly, the results cannot be applied to the locally admissible cases, where the VF (equivalently J_π) is finite only around the equilibrium point $x_{\mathsf{e}} = 0$ locally, not globally over $\mathcal{X}$. Lastly, not easy to verify in general is the local Lipschitzness assumptions on u_* and ∇J_π which are necessary for $\pi' \in \Pi_{\mathsf{Lip}}$. An example that is free from these limitations is the LQR (see §G.3 for a case study).

Remark 5.19 *This article is the first to define admissibility without asymptotic stability to the best authors' knowledge. This concept can be broadly applied, e.g., to the discounted LQR cases in §G.3 where the system may not be stable under an admissible policy due to $\gamma \in (0, 1)$. In fact, when $\gamma = 1$, admissibility of a policy π (i.e., $\pi \in \Pi_{\mathsf{a}}$) implies global asymptotic stability under π, with J_π served as a Lyapuonv function, as discussed in Remark 5.14. This reveals that asymptotic stability can be excluded from the definition of admissibility, even in the existing optimal control frameworks (as long as the VF is C^1). We also believe that our concept of admissibility can be generalized even when v_π (or equivalently, J_π) is locally finite around the equilibrium $x_{\mathsf{e}} = 0$ (i.e., locally admissible), not globally.*

Remark. *If the dynamics f is non-affine, then the cost function c has to be properly designed (e.g., by the techniques introduced in §§5.1.2 and G.1) to avoid the pathology in the Hamiltonian discussed in Remark 2.9 and §5.1.2. Note that such pathology can happen (see §F) even when c is positive definite (and quadratic when unconstrained), which is a typical choice in optimal control.*

6 Inverted-Pendulum Simulation Examples

To support the theory and further investigate the proposed PI methods, we simulate the variants of DPI and IPI shown in Algorithm 3 applied to an inverted-pendulum model:

$$\ddot{\vartheta}_t = -0.01\dot{\vartheta}_t + 9.8 \sin \vartheta_t - U_t \cos \vartheta_t,$$

where $\vartheta_t \in \mathbb{R}$ and $U_t \in \mathcal{U}$ are the angular position of and the external torque input to the pendulum at time t, respectively; the action space is given by $\mathcal{U} = [-u_{\mathsf{max}}, u_{\mathsf{max}}] \subset \mathbb{R}$, with the torque limit $u_{\mathsf{max}} = 5$ [N·m]. Letting $X_t \doteq [\, \vartheta_t \;\; \dot{\vartheta}_t \,]^{\mathsf{T}}$, then the dynamics can be expressed as (1) and (30) with

$$f_{\mathsf{d}}(x) = \begin{bmatrix} x_2 \\ 9.8 \, \sin x_1 - 0.01 x_2 \end{bmatrix} \text{ and } F_{\mathsf{c}}(x) = \begin{bmatrix} 0 \\ -\cos x_1 \end{bmatrix},$$

where $x = [\, x_1 \;\; x_2 \,]^{\mathsf{T}} \in \mathcal{X}$ $(= \mathbb{R}^2)$. In the simulations, we employ the zero initial policy $\pi_0(x) \equiv 0$, with the discount factor $\gamma = 0.1$ and the time step $\Delta t = 10$ [ms].

The solution V_i of the policy evaluation at each ith iteration is represented by a linear function approximator V as

$$V_i(x) \approx V(x; \theta_i) \doteq \theta_i^{\mathsf{T}} \phi(x), \tag{49}$$

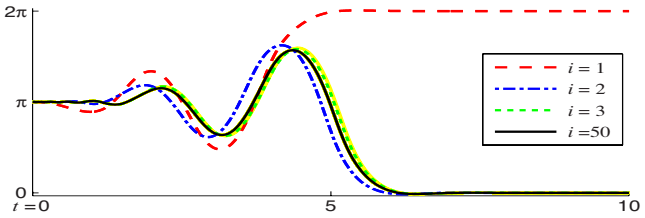
for its weights $\theta_i \in \mathbb{R}^L$ and features $\phi : \mathcal{X} \to \mathbb{R}^L$, with $L = 121$. Each policy evaluation determines θ_i as the least-squares solution θ_i^* minimizing the Bellman errors over the set of initial states uniformly distributed as the $(N \times M)$-grid points over the region $\Omega = [-\pi, \pi] \times [-6, 6] \subset \mathcal{X}$. Here, N and M are the total numbers of the grids in the x_1- and x_2-directions, respectively; we choose $N = 20$ and $M = 21$, so the total 420 number of grid points in Ω are used as initial states. When inputting to V, the first component x_1 of x is normalized to a value within $[-\pi, \pi]$ by adding $\pm 2\pi k$ to it for some $k \in \mathbb{Z}$.

In what follows, we simulate four different settings, whose learning objective is to swing up and eventually settle down the pendulum at the upright position $\theta_t = 2\pi k$ for some $k \in \mathbb{Z}$, under the torque limit $|U_t| \leq u_{\mathsf{max}}$. For each case, we basically consider the reward function r given by (31) and (35) with

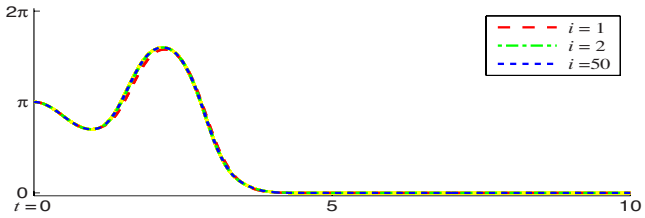
$$s(\mathfrak{u}) = u_{\mathsf{max}} \tanh(\mathfrak{u}/u_{\mathsf{max}}). \tag{50}$$

As the inverted pendulum dynamics is input-affine, this setting corresponds to the concave Hamiltonian formulation in §5.1.1 (with a bounded r if $\mathfrak{r}$ is bounded). The implementation details (the features ϕ, policy evaluation, and policy improvement) are provided in §H; the MATLAB/Octave source code for the simulations is also available online. [9]

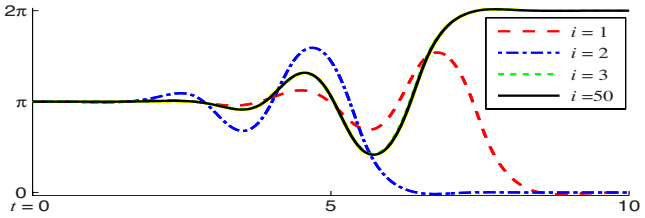
[9] github.com/JaeyoungLee-UoA/PIs-for-RL-Problems-in-CTS/



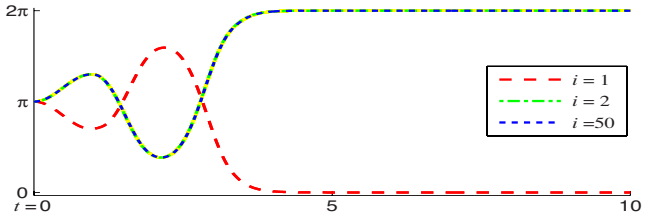
(a) Case 1: concave Hamiltonian with bounded reward — DPI



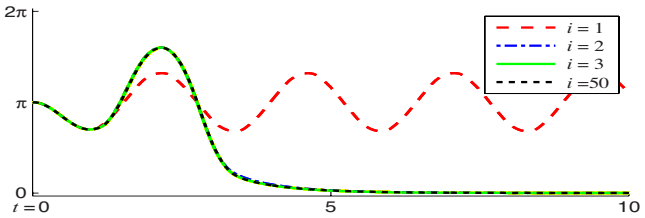
(b) Case 2: optimal control — DPI



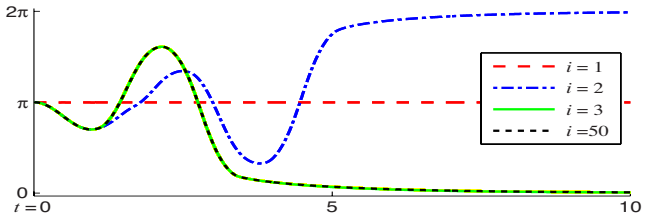
(c) Case 1: concave Hamiltonian with bounded reward —IPI



(d) Case 3: bang-bang control — IPI with $r(x, u) = \cos x_1$



(e) Case 4: bang-bang control with binary reward – DPI



(f) Case 4: bang-bang control with binary reward – IPI

Fig. 1. The pendulum angular-position trajectories ϑ_t during and after PI for each case study. All of the trajectories start from $x = (\pi, 0)$, and the yellow regions correspond to those trajectories ϑ_t generated by the policies obtained at the iterations $i = 3, 4, 5, \cdots, 49$.

6.1 Case 1: Concave Hamiltonian with Bounded Reward

First, we consider the reward function r given by (31) and (35) with $s(\cdot)$ given by (50), $\Gamma = 10^{-2}$, and $\mathfrak{r}(x) = \cos x_1$. As mentioned above, this setting corresponds to the concave Hamiltonian formulation in §5.1, resulting in the following policy improvement update rule (see §H for details):

$$\pi_i(x) \approx \pi(x; \theta_i^*) = -5 \tanh\big(\cos x_1 \cdot \nabla_{x_2} \phi(x) \cdot \theta_i^* / 5\big). \quad (51)$$

As $\mathfrak{r}$ (hence r) is bounded, this setting also corresponds to "discounted RL under Assumption 5.8" in §5.2. Therefore, the initial and subsequent VFs in PIs are all bounded; the properties in §§5.1.1 and 5.2 are all true; the Assumptions in Table 1 w.r.t. §§5.1.1 and 5.2 are also all relaxed.

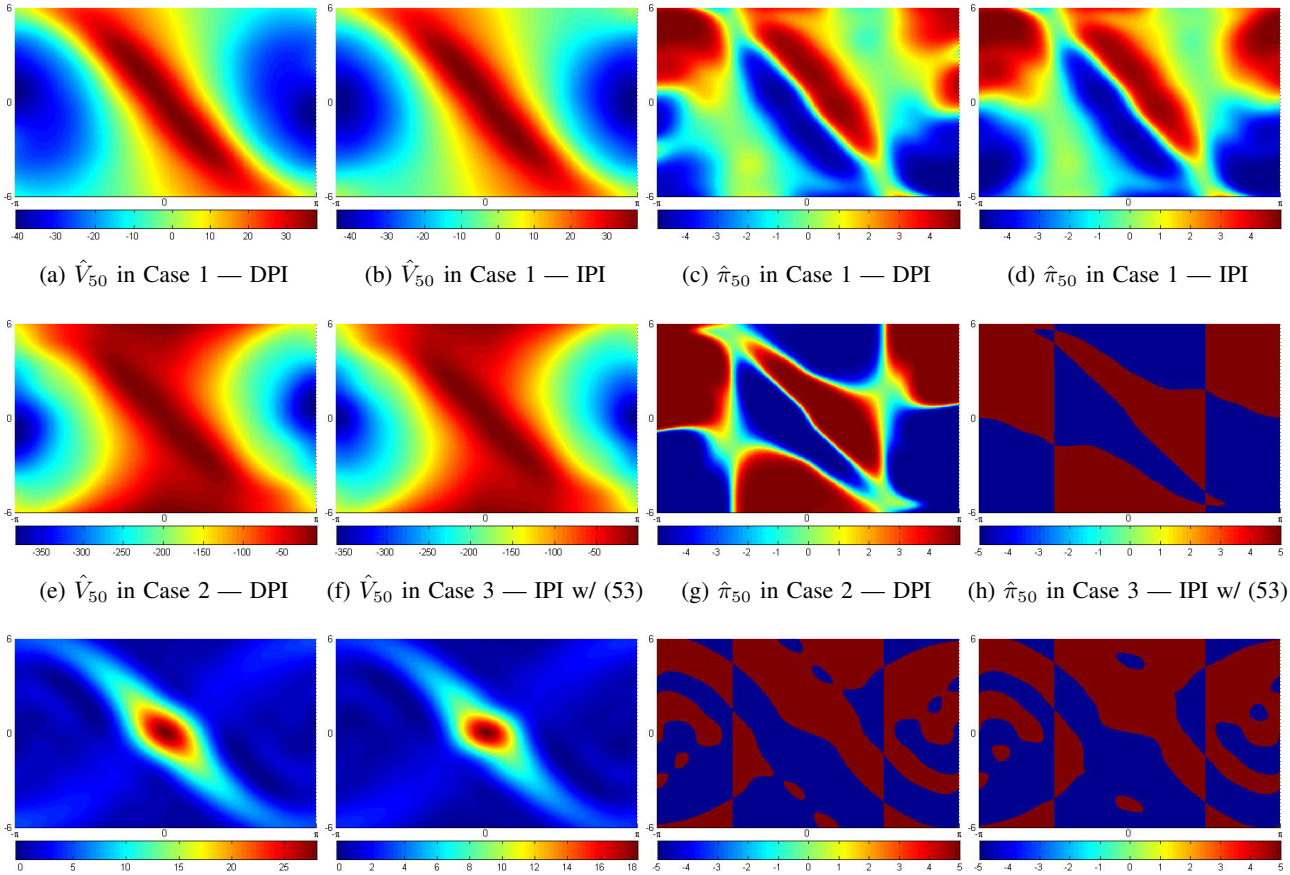
Figs. 1(a), (c) and Figs. 2(a)–(d) show the trajectories of ϑ_t under the policies obtained during PI and the estimates of the optimal solution (v_*, π_*) finally obtained at the iteration $i = 50$, respectively; the yellow regions in Fig. 1 correspond to the trajectories ϑ_t generated by the intermediate policies obtained at the iterations $i = 3, 4, 5, \cdots, 49$. Although both DPI and IPI variants generate rather different trajectories of ϑ_t in Figs. 1(a), (c) due to the difference in the estimates of the VF and policy (e.g., see Figs. 2(a)–(d)), both methods have achieved the learning objective merely after the first iteration. Here, the difference in the ϑ_t-trajectories mainly comes from the different initial behaviors near $\vartheta = \pi$ — see the differences in the policies in Figs. 2(c), (d) (and also the VF estimates in Figs. 2(a), (b)) near the borderlines $\vartheta = \pm\pi$. Also note that both DPI and IPI methods have achieved our learning objective without using an initial stabilizing policy that is usually required in the optimal control setting under the total discounting $\gamma = 1$ (e.g., Abu-Khalaf and Lewis, 2005; Vrabie and Lewis, 2009; Lee et al., 2015).

6.2 Case 2: Optimal Control

A better performance can be obtained if the state reward function $\mathfrak{r}$ in Case 1 is replaced by

$$\mathfrak{r}(x) = -x_1^2 - \epsilon \cdot x_2^2 \quad \text{with } \epsilon = 10^{-2}. \quad (52)$$

This setting corresponds to the nonlinear optimal control introduced and discussed in §5.4. Whenever input to $\mathfrak{r}$ in (52), the first component x_1 is normalized to a value within $[-\pi, \pi]$. In this case, $\mathfrak{r}$ is not bounded due to the existence of the term $-\epsilon \cdot x_2^2$, but Algorithm 3 (without assuming the boundedness of v_{π_0}) can be successfully applied as can be seen from Figs. 1(b), 2(e), and 2(g). Fig. 1(b) illustrates the



(a) $\hat{V}_{50}$ in Case 1 — DPI (b) $\hat{V}_{50}$ in Case 1 — IPI (c) $\hat{\pi}_{50}$ in Case 1 — DPI (d) $\hat{\pi}_{50}$ in Case 1 — IPI

(e) $\hat{V}_{50}$ in Case 2 — DPI (f) $\hat{V}_{50}$ in Case 3 — IPI w/ (53) (g) $\hat{\pi}_{50}$ in Case 2 — DPI (h) $\hat{\pi}_{50}$ in Case 3 — IPI w/ (53)

(i) $\hat{V}_{50}$ in Case 4 — DPI (j) $\hat{V}_{50}$ in Case 4 — IPI (k) $\hat{\pi}_{50}$ in Case 4 — DPI (ℓ) $\hat{\pi}_{50}$ in Case 4 — IPI

Fig. 2. The optimal value function $\hat{V}_{50}(x) = V(x;\theta_i^*)|_{i=50}$ (left sides) and the optimal policy $\hat{\pi}_{50}(x) = \pi(x;\theta_i^*)|_{i=50}$ (right sides), estimated by DPI and IPI variants over Ω. The horizontal and vertical axes correspond to x_1 ($= \vartheta$) and x_2 ($= \dot{\vartheta}$), respectively.

trajectories ϑ_t under the policies obtained by the DPI variant, under (52). Compared with Case 1, this setting gives a better initial and asymptotic performance — every trajectory ϑ_t in Fig. 1(b) is almost the same as the final one (faster convergence of the PI) and converges to the goal state $x = (0, 0)$ more rapidly than any trajectories ϑ_t' s in Case 1. In particular, the initial behavior near $\vartheta = \pm\pi$ has been improved, so that the policies in this case swing up the pendulum much faster than Case 1. One possible explanation about this is that the higher magnitude of the gradient of $\mathfrak{r}$ near $x_1 = \pm\pi$ expedites the initial swing-up process (note, in Case 1, $\nabla\mathfrak{r}(\pm\pi, x_2) = 0$ for any x_2). See also the difference of the final VF and policy in Figs. 2(e), (g) (this case) from those in Figs. 2(a)–(d) (Case 1). The trajectories ϑ_t's for IPI are almost similar to DPI in this case, so omitted.

6.3 *Case 3: Bang-bang Control*

If $\Gamma \to 0$, the reward function r and the policy update rule (51) in Case 1 (§6.1) are simplified to $r(x, u) = \cos x_1$ and

$$\pi_i(x) \approx \pi(x;\theta_i^*) = -u_{\max} \cdot \mathrm{sign}\big(\cos x_1 \cdot \nabla_{x_2}\phi(x) \cdot \theta_i^*\big)$$

(see §H for details), a bang-bang type discrete control. The PI methods can be also applied to optimize this bang-bang type controller. Note that this case is beyond our scope of the theory developed in §§2–5 since the policy is discrete, not continuous. Fig. 1(d) shows the trajectory ϑ_t generated by the IPI variant in Algorithm 3 applied to this bang-bang control framework. Though the fast switching behavior of the control U_t near $x = (0, 0)$ is inevitable, the initial and asymptotic control performance, compared with Case 1, has been increased in the limit $\Gamma \to 0$ up to the performance of optimal control (Case 2).

By limiting $\Gamma \to 0$, the control policy in Case 2 can be also made a bang-bang type control, but in this case, with

$$r(x, u) = -x_1^2 - \epsilon \cdot x_2^2 \quad \text{with } \epsilon = 10^{-2}. \tag{53}$$

We have observed that the performance of the PI methods in this case is almost same as that shown in Fig. 1(d) for the previous case "$r(x, u) = \cos x_1$," derived from Case 1. Figs. 2(f) and (h) show the envelopes of the VF and the bang-bang policy under (53), both of which are consistent with the envelopes for $\Gamma = 10^{-2}$ shown in Figs. 2(e) and (g).

6.4 Case 4: Bang-bang Control with Binary Reward

In RL problems, the reward is often binary and sparely given only at or near the goal state. To investigate this case, we also consider the bang-bang policy given in the previous subsection, but with the binary reward function: $r(x,u) = 1$ if $|x_1| \leq 6/\pi$ and $|x_2| \leq 1/2$ and $r(x,u) = 0$ otherwise, This gives the reward signal $R_t = 1$ near the goal state $x = (0,0)$ only. Figs. 1(e) and (f) illustrate the θ_t-trajectories under the policies generated by the DPI and IPI variants (i.e., Algorithm 3), respectively. Though the initial performance is neither stable ($i = 1$) nor consistent to each other ($i = 1, 2$), both PI methods eventually converge to the same seemingly near-optimal point ($i = 3, 4, \cdots, 50$). Note that the performance after learning ($i = 50$) for both cases is the same as that of Cases 2 and 3 until around $t = 3[s]$ as can be seen from Figs. 1(b) and (d)–(f). Figs. 2(i)–(ℓ) also show the estimates of the optimal VF and policy at $i = 50$. Although the details are a bit different, we can see that both methods finally result in similar consistent estimates of the VF and policy. In this binary reward case, the shapes of the VF shown in Figs. 2(i) and (j) are distinguished from the others illustrated in Figs. 2(a),(b),(e), and (f) due to the reward information condensed near the goal state $x = (0,0)$ only. Even in this situation, our PI methods were able to achieve the goal at the end, as shown in Figs. 1(e) and (f). For the DPI variant, we have simulated this case with $M = 20$, instead of $M = 21$.

6.5 Discussions

We have simulated the variants of DPI and IPI (Algorithm 3) under the four scenarios above. Some of them have achieved the learning objective immediately at the first iteration, and in all of the simulations above, the proposed methods were able to achieve the goal, eventually. On the other hand, the implementations of the PIs have the following issues.

(1) The least-squares solution θ_i^* at each ith policy evaluation minimizes the Bellman error over a finite number of initial states in Ω (as detailed in §H), meaning that it is not the optimal choice to minimize the Bellman error over the entire region Ω. As mentioned in §3, the ideal policy evaluation cannot be implemented precisely — even when Ω is compact, it is a continuous space and thus contains an uncountably infinite number of points that we cannot fully cover in practice.

(2) As the dimension of the data matrix in the least squares is $L \times (NM) = 121 \times 420$ (see §H), calculating the least-squares solution θ_i^* is computationally expensive, and the numerical error (and thus the convergence) is sensitive to the choice of the parameters such as (the number of) the features ϕ, the time step Δt, discounting factor γ, and of course, N and M. In our experiments, we have observed Case 2 (optimal control) was least sensitive to those parameters.

(3) The VF parameterization. Since the pendulum is symmetric at $x_1 = 0$, the VFs and policies obtained in Fig 2 are all symmetric, and thus it might be sufficient to approximate the VF over $[0, \pi] \times [-6, 6] \subset \Omega$, with a less number of weights, and use the symmetry of the problem. Due to the over-parameterization, we have also observed that the weight vector θ_i^* in certain situations never converges but oscillates between two values, even after the VF V_i almost converges over Ω.

All of these algorithmic and practical issues are beyond the scope of this paper and remain as a future work.

7 Conclusions

In this paper, we proposed fundamental PI schemes called DPI (model-based) and IPI (partially model-free) to solve the general RL problem formulated in CTS. We proved their fundamental mathematical properties: admissibility, uniqueness of the solution to the BE, monotone improvement, convergence, and the optimality of the solution to the HJBE. Strong connections to the RL methods in CTS — TD learning and VGB greedy policy update — were made by providing the proposed ones as their ideal PIs. Case studies simplified and improved the proposed PI methods and the theory for them, with strong connections to RL and optimal control in CTS. Numerical simulations were conducted with model-based and partially model-free implementations to support the theory and further investigate the proposed PI methods, under an initial policy that is admissible but not stable. Unlike the existing PI methods in the stability-based framework, an initial *stabilizing* policy is not necessarily required to run the proposed ones. We believe that this work provides the theoretical background, intuition, and improvement to both (i) PI methods in optimal control and (ii) RL methods, to be developed in the future and developed so far.

References

Abu-Khalaf, M. and Lewis, F. L. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, 41(5):779–791, 2005.

Baird III, L. C. Advantage updating. Technical report, DTIC Document, 1993.

Beard, R. W., Saridis, G. N., and Wen, J. T. Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation. *Automatica*, 33(12):2159–2177, 1997.

Bian, T., Jiang, Y., and Jiang, Z.-P. Adaptive dynamic programming and optimal control of nonlinear nonaffine systems. *Automatica*, 50(10):2624–2632, 2014.

Doya, K. Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245, 2000.

Folland, G. B. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

Frémaux, N., Sprekeler, H., and Gerstner, W. Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLoS Comput. Biol.*, 9(4):e1003024, 2013.

Gaitsgory, V., Grüne, L., and Thatcher, N. Stabilization with discounted optimal control. *Syst. Control Lett.*, 82:91–98, 2015.

Haddad, W. M. and Chellaboina, V. *Nonlinear dynamical systems and control: a Lyapunov-based approach*. Princeton University Press, 2008.

Howard, R. A. *Dynamic drogramming and Markov processes*. Tech. Press of MIT and John Wiley & Sons Inc., 1960.

Khalil, H. K. *Nonlinear systems*. Prentice Hall, 2002.

Kiumarsi, B., Kang, W., and Lewis, F. L. H_∞ control of non-affine aerial systems using off-policy reinforcement learning. *Unmanned Systems*, 4(01):51–60, 2016.

Kleinman, D. On an iterative technique for Riccati equation computations. *IEEE Trans. Autom. Cont.*, 13(1):114–115, 1968.

Leake, R. J. and Liu, R.-W. Construction of suboptimal control sequences. *SIAM Journal on Control*, 5(1):54–63, 1967.

Lee, J. and Sutton, R. S. Policy iterations for reinforcement learning problems in continuous time and space — fundamental theory and methods: Appendices. *To appear as an ArXiv preprint.*, 2020a.

Lee, J. Y., Park, J. B., and Choi, Y. H. Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations. *IEEE Trans. Neural Networks and Learning Systems*, 26(5):916–932, 2015.

Lee, J. Y. and Sutton, R. Policy iteration for discounted reinforcement learning problems in continuous time and space. In *Proc. the Multi-disciplinary Conf. Reinforcement Learning and Decision Making (RLDM)*, 2017.

Lewis, F. L. and Vrabie, D. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 9(3):32–50, 2009.

Mehta, P. and Meyn, S. Q-learning and pontryagin's minimum principle. In *Proc. IEEE Int. Conf. Decision and Control, held jointly with the Chinese Control Conference (CDC/CCC)*, pages 3598–3605, 2009.

Modares, H., Lewis, F. L., and Jiang, Z.-P. Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning. *IEEE Trans. Cybern.*, 46(11):2401–2410, 2016.

Modares, H. and Lewis, F. L. Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. *IEEE Transactions on Automatic Control*, 59(11):3051–3056, 2014.

Murray, J. J., Cox, C. J., Lendaris, G. G., and Saeks, R. Adaptive dynamic programming. *IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev.*, 32(2):140–153, 2002.

Murray, J. J., Cox, C. J., and Saeks, R. E. The adaptive dynamic programming theorem. In *Stability and Control of Dynamical Systems with Applications*, pages 379–394. Springer, 2003.

Powell, W. B. *Approximate dynamic programming: solving the curses of dimensionality*. Wiley-Interscience, 2007.

Rekasius, Z. Suboptimal design of intentionally nonlinear controllers. *IEEE Transactions on Automatic Control*, 9(4):380–386, 1964.

Rudin, W. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Saridis, G. N. and Lee, C. S. G. An approximation theory of optimal control for trainable manipulators. *IEEE Trans. Syst. Man Cybern.*, 9(3):152–159, 1979.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: an introduction*. Second Edition, MIT Press, Cambridge, MA (available at http://incompleteideas.net/book/the-book.html), 2018.

Tallec, C., Blier, L., and Ollivier, Y. Making deep Q-learning methods robust to time discretization. In *International Conference on Machine Learning (ICML)*, pages 6096–6104, 2019.

Thomson, B. S., Bruckner, J. B., and Bruckner, A. M. *Elementary real analysis*. Prentice Hall, 2001.

Vrabie, D. and Lewis, F. L. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Netw.*, 22(3):237–246, 2009.

Policy Iterations for Reinforcement Learning Problems in Continuous Time and Space — Fundamental Theory and Methods: Appendices

Jaeyoung Lee,[a] Richard S. Sutton[b]

[a]*Department of Electrical and Computer Eng., University of Waterloo, Waterloo, ON, Canada, N2L 3G1 (jaeyoung.lee@uwaterloo.ca)*

[b]*Department of Computing Science, University of Alberta, Edmonton, AB, Canada, T6G 2E8 (rsutton@ualberta.ca)*

Abstract

This supplementary document provides additional studies and all the details regarding the materials presented by Lee and Sutton (2020b), as listed in the contents below. Roughly speaking, we present related works, details of the theory, algorithms, and implementations, additional case studies, and all the proofs. All the numbers of equations, sections, theorems, lemmas, etc. that do not contain any alphabet will refer to those in the main paper (Lee and Sutton, 2020b), whereas any numbers starting with an alphabet correspond to those in the appendices herein. Additional references cited only in the appendices are also given at the end, separately from the main paper.

A Highlights and Related Works

First, we briefly review the related works from optimal control and RL fields and highlights the main aspects of the proposed PI methods and the underlying theory developed in the main paper (Lee and Sutton, 2020b) and the appendices.

DPI & IPI. The two main PI methods in the work are DPI, whose policy evaluation is associated with the differential BE, and IPI associated with the integral BE. The former is inspired by the model-based PI methods in optimal control (e.g., Rekasius, 1964; Leake and Liu, 1967; Saridis and Lee, 1979; Beard et al., 1997; Abu-Khalaf and Lewis, 2005; Bian et al., 2014) and has a direct connection to TD(0) in CTS (Doya, 2000; Frémaux et al., 2013). As regards to the latter, the integral BE was first introduced by Baird III (1993) in the field of RL and then spotlighted in the optimal control community, resulting in a series of IPI methods applied to a class of input-affine dynamics for optimal regulations (Vrabie and Lewis, 2009; Lee et al., 2015), robust control (Wang, Li, Liu, and Mu, 2016), and (discounted) LQ tracking control (Modares and Lewis, 2014; Zhu, Modares, Peen, Lewis, and Yue, 2015; Modares et al., 2016), with a number of extensions to off-policy IPI methods (e.g., Bian et al., 2014; Lee et al., 2015; Wang et al., 2016; Modares et al., 2016). In our work (Lee and Sutton, 2020b),

(1) the proposed IPI was motivated by the first IPI given by Vrabie and Lewis (2009) for nonlinear optimal regulations;

(2) both DPI and IPI generalize for a broad class of dynamics and reward functions in CTS in §2, which includes the existing RL tasks (Doya, 2000; Mehta and Meyn, 2009; Frémaux et al., 2013) and the tasks of RL and optimal control shown in the case studies in §§5 and G.

Case Studies.

(1) A highlight is in §5.1, which draws the connection to the VGB greedy policy update (Doya, 2000), a general idea of simplifying policy improvement in input-constrained RL problems. There exist similar ideas in the optimal control field for input-constrained (Lyashevskiy, 1996; Abu-Khalaf and Lewis, 2005) and unconstrained optimal regulations (Rekasius, 1964; Saridis and Lee, 1979; Beard et al., 1997; Abu-Khalaf and Lewis, 2005; Vrabie and Lewis, 2009; Lee et al., 2015) under input-affine dynamics, and even for the non-affine dynamics (Bian et al., 2014; Kiumarsi et al., 2016).

(2) The existing PI methods for optimal regulations, presented in the literature above and by Leake and Liu (1967), are strongly linked to §5.4, where we case-study asymptotic stability and fundamental properties of DPI and IPI applied to a general optimal regulation problem with non-affine dynamics and $\gamma \in (0, 1]$. The asymptotic stability conditions given in Theorem 5.13 in §5.4 are similar to and inspired by Gaitsgory et al. (2015, Assumptions 2.3 and 3.8).

(3) Another highlight is the discounted RL problem with *bounded* reward function (§5.2), in which the VF is bounded for any policy; hence the underlying PI theory is dramatically simplified and clear (see Corollary 5.9). This framework is akin to the RL tasks in a finite MDP, where the reward defined for each state transition is bounded (Sutton and Barto, 2018).

§6 provides practical examples and highlights those case studies in §5.1, 5.2, and 5.4 with strong connections to both RL and optimal control.

Admissibility & Asymptotic Stability. Theoretically, since we consider a stability-free RL framework (under the minimal assumptions in §2), asymptotic stability is excluded from the definition of an admissible policy. Here, the notion of admissibility in optimal control was defined with asymptotic stability (e.g., Beard et al., 1997; Abu-Khalaf and Lewis, 2005; Vrabie and Lewis, 2009; Modares and Lewis, 2014; Bian et al., 2014; Lee et al., 2015 to name a few), and this work is the first to define admissibility in CTS *without* asymptotic stability. Conversely, in a general optimal control problem, we also showed that when $\gamma = 1$, admissibility, according to our definition, implies asymptotic stability (if the associated VF is C^1) — see Theorem 5.13 and Remarks 5.14 and 5.19 in §5.4. This means that asymptotic stability can be removed from the definition of admissibility also in *optimal control*. The global existence of the unique state trajectories in discounted optimal control was also investigated in §5.4 under the condition weaker than a Lyapunov's asymptotic stability criterion.

(Mode of) Convergence. We characterized the convergence properties of the PI methods towards the optimal solution in the following three ways. Those three modes provide different convergence conditions and compensate for one another.

(1) In the first characterization, we use Bessaga (1959)'s converge fixed point principle to show that the VFs generated by the PI methods converge to the optimal one in a metric (Theorem 4.5). This first-type convergence, called convergence in a metric, is weaker than locally uniform convergence below but does not impose any other assumptions than the existence and uniqueness of a fixed point that turns out to be the optimal VF.

(2) The second way is to extend Leake and Liu (1967)'s approach that suggests continuity of the PI operator (see Theorem 4.6) as one of the additional conditions for locally uniform convergence.

(3) Lastly, we also generalize the convergence proof from the optimal control literature (Saridis and Lee, 1979; Beard et al., 1997; Murray et al., 2002; Abu-Khalaf and Lewis, 2005; Bian et al., 2014) to our RL framework, which gives the strongest convergence among the three, under a certain condition other than the two above (see Theorem 4.9). In this

direction, we highlight that for the proof of this third type convergence, the gradients of the VFs obtained by the PIs need to be assumed to converge locally uniformly, *even for the existing results in optimal control*, as the convergence of the generated VFs does not imply any convergence of their derivatives (see Remark 5.2).

LQR. In §G.3, we discuss DPI and IPI applied to a class of the LQR tasks (Lancaster and Rodman, 1995, Chapter 16) where bilinear cost terms of states and controls exist. Here, DPI falls into a special case of the general matrix-form PIs (Arnold III, 1984; Mehrmann, 1991), but this study slightly generalizes the existing PI methods for the LQRs (Kleinman, 1968; Vrabie et al., 2009; Lee, Park, and Choi, 2014) by taking such bilinear cost terms into considerations, with the relaxation of the positive definite matrix assumption imposed on the general matrix-form PI (Mehrmann, 1991, Theorem 11.3).

B Notations and Terminologies

In this appendix, we provide a complete list of notations and terminologies used in the main paper and/or the appendices. In any statement, iff and s.t. stand for *if and only if* and *such that*, respectively. "$\doteq$" denotes the equality relationship that is true *by definition*.

B.1 Sets, Vectors, and Matrices

$\mathbb{N}$	set of all natural numbers
$\mathbb{R}$	set of all real numbers
$\mathbb{C}$	set of all complex numbers
$\mathbb{Z}$	set of all integers
$\mathbb{R}^{n\times m}$	set of all n-by-m real matrices
$\mathbb{R}^n$	n-dimensional Euclidean space $\doteq \mathbb{R}^{n\times 1}$

For a matrix $A \in \mathbb{R}^{n\times m}$ and a vector $x \in \mathbb{R}^m$,

A^T	transpose of A
$\operatorname{rank}(A)$	rank of A
$\|x\|$	Euclidean norm of x, i.e., $\|x\| \doteq (x^\mathsf{T}x)^{1/2}$
$\|x\|_\Omega$	distance of x from a subset $\Omega \subset \mathbb{R}^m$, i.e., $\|x\|_\Omega \doteq \inf\{\|x-y\| : y \in \Omega\}$
$\|\|\|A\|\|\|$	induced norm of A, i.e., $\|\|\|A\|\|\| \doteq \sup_{\|x\|=1} \|Ax\|$
I	identity matrix with a compatible dimension

B.2 Euclidean Topology

Let $\Omega \subseteq \mathbb{R}^n$.

Ω° denotes the *interior of* Ω.
$\partial\Omega$ denotes the *boundary of* Ω.
Ω is said to be *compact* iff it is closed and bounded.

If Ω is open, then $\Omega\cup\partial\Omega$ (resp. Ω) is called an *n-dimensional manifold with* (resp. *without*) *boundary*. By this definition, a manifold contains no isolated point.

B.3 Functions, Sequences, and Convergence

Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{R}^m$ be a function.

$f \in \mathrm{C}^k$ (i.e., f *is* C^k) iff the kth order partial derivatives of f all exist and are continuous, over the interior Ω°.
$\nabla f : \Omega^\circ \to \mathbb{R}^{m\times n}$ denotes *the gradient of* f.
f is *locally Lipschitz* iff for each $x \in \Omega$, there exists $L > 0$ and a neighborhood $\mathcal{N}$ of x s.t. for all $y, z \in \mathcal{N}$,

$$\|f(y) - f(z)\| \le L\|y - z\|. \tag{B.1}$$

f is *globally Lipschitz* iff $\exists L > 0$ s.t. (B.1) holds $\forall y, z \in \Omega$.
$f \in \mathrm{C}^1_{\mathsf{Lip}}$ (i.e., f *is* $\mathrm{C}^1_{\mathsf{Lip}}$) iff f is locally Lipschitz and C^1.
f is *odd* iff $f(-x) = f(x)$ for all $x \in \Omega$.
f with $m = n$ is *strictly monotone* iff for each x, $x' \in \Omega$,

$$(f(x) - f(x'))^\mathsf{T}(x - x') > 0 \text{ whenever } x \neq x'.$$

$f(E) \doteq \{f(x) : x \in E\}$, the image of $E \subseteq \Omega$ under f.

$f : [0,\infty) \to [0,\infty)$ is said to be $\mathcal{K}_\infty$ iff $f(0) = 0$ and f is strictly increasing, radially unbounded, and continuous;

A sequence $\langle a_i\rangle_{i=1}^\infty$ is abbreviated as $\langle a_i\rangle$ or a_i for notational simplicity. A sequence of function $\langle f_i\rangle$ *converges (to f)*

pointwise iff $f_i(x) \to f(x)$ for each $x \in \Omega$;
uniformly on $E \subseteq \Omega$ iff $\sup_{x\in E} \|f_i(x) - f(x)\| \to 0$;
locally uniformly iff for each $x \in \Omega$, there is a neighborhood of x on which $f_i \to f$ uniformly.

For any two functions $f_1, f_2 : \mathbb{R}^n \to [-\infty, \infty)$, we write

$$f_1 \leqslant f_2 \iff f_1(x) \le f_2(x) \quad x \in \mathbb{R}^n.$$

A function $f : \Omega \to \mathbb{R}$ is said to be

positive semidefinite iff $f(0) = 0$ and $f \geqslant 0$;
negative semidefinite iff $-v$ positive semidefinite;
positive definite iff $f(0) = 0$ and $f(x) > 0$ for all $x \neq 0$;
negative definite iff $-f$ positive definite;
radially unbounded iff $\inf_{\|x\|\ge r} |f(x)| \to \infty$ as $r \to \infty$;
convex iff for each $x, x' \in \Omega$ and $\beta \in (0,1)$,
(1) $x_\beta \doteq \beta x + (1-\beta)x' \in \Omega$ (i.e., Ω is *convex*),
(2) $f(x_\beta) \le \beta \cdot f(x) + (1-\beta)\cdot f(x')$;
concave iff $-f$ is convex;

strictly convex iff f is convex and for any $\beta \in (0,1)$, $f(x_\beta) < \beta \cdot f(x) + (1-\beta) \cdot f(x')$ whenever $x \neq x'$;

strictly concave iff $-f$ is strictly convex.

A square matrix $P \in \mathbb{R}^{n\times n}$ is

positive (semi)definite iff so is $z \mapsto z^\mathsf{T} P z$ and $P^\mathsf{T} = P$;
negative definite iff $-P$ is positive definite.

For $P, P' \in \mathbb{R}^{n\times n}$, we denote $P < P'$ (resp. $P \leq P'$) iff $P' - P$ is positive definite (resp. positive semidefinite).

B.4 Reinforcement Learning

l	dimension $\in \mathbb{N}$ of the state space $\mathcal{X}$
m	dimension $\in \mathbb{N}$ of action spaces (e.g., $\mathcal{U}$ and $\mathcal{A}$)

An *action space* is an m-dimensional manifold in $\mathbb{R}^m$ with or without boundary hence has no isolated point by definition.

$\mathcal{X}, \mathcal{X}^\mathsf{T}$	state space $\mathcal{X} \doteq \mathbb{R}^l$ and $\mathcal{X}^\mathsf{T} \doteq \mathbb{R}^{1\times l}$
$\mathcal{U}$	action space $\subseteq \mathbb{R}^m$
$\mathcal{A}$	a transformed action space $\subseteq \mathbb{R}^m$ (§5.1)
$\mathbb{T}$	time space $\mathbb{T} \doteq [0,\infty)$
f, f^x	dynamics $f : \mathcal{X}\times\mathcal{U} \to \mathcal{X}$ and $f^x(u) \doteq f(x,u)$
f_d	drift dynamics $f_\mathsf{d} : \mathcal{X} \to \mathcal{X}$
f_c	input-coupling dynamics $f_\mathsf{c} : \mathcal{X}\times\mathcal{U} \to \mathcal{X}$
F_c	input-coupling matrix $F_\mathsf{c} : \mathcal{X} \to \mathbb{R}^{n\times m}$ (§5.1)
r, r^x	reward function $r : \mathcal{X}\times\mathcal{U} \to \mathbb{R}$; $r^x(u) \doteq r(x,u)$
r_max	the reward maximum $\max_{(x,u)} r(x,u)$
γ	discount factor $\in (0,1]$
α	attenuation rate $\alpha \doteq -\ln\gamma \geq 0$
h	Hamiltonian function $h : \mathcal{X}\times\mathcal{U}\times\mathcal{X}^\mathsf{T} \to \mathbb{R}$
u_*	maximal function $u_*(x,p) \in \arg\max_u h(x,u,p)$
a_*	transformed maximal function (§5.1.2)
t	time variable $\in \mathbb{T}$
η	time horizon $\in (0,\infty]$
X_t	state vector $\in \mathcal{X}$ at time t
$\dot{X}_t$	the time derivative $\in \mathcal{X}$ of X_t at time t
U_t	action (also called control) vector $\in \mathcal{U}$ at time t
A_t	a transformed action vector $\in \mathcal{A}$ at time t
R_t	reward at time t, i.e., $r(X_t, U_t) \in \mathbb{R}$
$\mathfrak{R}_\eta$	discounted cumulative reward up to horizon η
$\dot{v}$	time derivative $dv(X_t)/dt = \nabla v(X_t) f(X_t, U_t)$

A policy is a continuous function from $\mathcal{X}$ to $\mathcal{U}$; for a policy π,

$\mathbb{G}^x_\pi[Y]$	value Y when $X_0 = x$ and $U_t = \pi(X_t)\ \forall t \in \mathbb{T}$
v_π	value function (VF) with respect to π
$\bar{v}$	a uniform upper-bound of VFs ($\bar{v} = 0$ for $\gamma = 1$ and $\bar{v} = r_\mathsf{max}/\alpha$ otherwise — see Lemma 2.1)
f_π	closed-loop dynamics $f_\pi(x) \doteq f(x, \pi(x))$
r_π	closed-loop reward function $r_\pi(x) \doteq r(x, \pi(x))$
π'	improved/maximal policy $\pi' \succcurlyeq \pi$, i.e., $v_{\pi'} \geqslant v_\pi$

When f_π is locally Lipschitz, $t_\mathsf{max}(x;\pi)$ is the minimal time s.t. $\forall t \geq t_\mathsf{max}(x;\pi)$, *no* state $\mathbb{G}^x_\pi[X_t]$ exists (see §5.3).

Π_a	set of all admissible policies
Π_Lip	set of all locally Lipschitz policies
$\mathcal{V}_\mathsf{a}$	set of all admissible VFs
d, d_Ω	a metric and uniform pseudo-metric on $\mathcal{V}_\mathsf{a}$
$\mathcal{T}$	PI operator
v_*	a solution to the HJBE or the optimal VF
v^*	a unique fixed point of $\mathcal{T}$
π_*	a HJB or the optimal policy

B.5 Policy Iteration

i	iteration index $\in \mathbb{N}$ of the PIs
v_i, V_i	solution to the BE at iteration i; $V_i \doteq v_i/\Delta t$
$\hat{v}_*$	limit function $\hat{v}_*(x) \doteq \sup_i v_i(x) = \lim_{i\to\infty} v_i(x)$
π_0, π_i	initial and improved policies at the ith iteration
Δt	small time step $(0 < \Delta t \ll 1)$
γ_d	discrete-time discount factor $\doteq \gamma^{\Delta t}$
$\hat{\gamma}_\mathsf{d}$	an approximation of $\gamma_\mathsf{d} \approx \hat{\gamma}_\mathsf{d} \doteq 1 - \alpha_\mathsf{d}$
α_d	discrete-time attenuation rate $\doteq \alpha\Delta t = -\ln\gamma_\mathsf{d}$

B.6 Optimal Control and LQRs

c	cost function $c \doteq -r$
c_π	closed-loop cost function $c_\pi(x) \doteq c(x, \pi(x))$
C_t	cost at time t, i.e., $c(X_t, U_t) \in \mathbb{R}$
J_π	cost value function $J_\pi \doteq -v_\pi$
J_i, J_*	$J_i \doteq -v_i$ and $J_* \doteq -v_*$
Π_0	set of all policies $\pi \in \Pi_\mathsf{Lip}$ s.t. $\pi(0) = 0$

Let $A \in \mathbb{R}^{l\times l}$, $B \in \mathbb{R}^{m\times l}$, and $C \in \mathbb{R}^{p\times l}$. Then,

A is *Hurwitz* iff every eigenvalue has a negative real part;
(A, B) *stabilizable* iff $\exists K \in \mathbb{R}^{l\times m}$ s.t. $A - BK$ is Hurwitz;
(C, A) *observable* iff for any $\eta > 0$, the initial state X_0 can be determined from the history $\{(CX_t, U_t)\}_{t\in[0,\eta]}$, where $\{X_t\}_{t\in[0,\eta]}$ satisfies $\dot{X}_t = AX_t + BU_t$.

C Replacement of the Boundary Condition with an Inequality

In §2.2, Lee and Sutton (2020b) showed that the boundary condition (12), i.e., for a given horizon $\eta > 0$,

$$\lim_{k\to\infty} \mathbb{G}_\pi^x\big[\,\gamma^{k\cdot\eta}\cdot v(X_{k\cdot\eta})\,\big] = 0 \qquad \forall x \in \mathcal{X} \tag{12}$$

is necessary and sufficient for a solution v to the BE (10) or (11) to be equal to the VF v_π. In other words, the boundary condition (12) ensures the uniqueness of the solution v to the BE. However, except for a few cases, (12) is hard or even impossible to check as (12) is a condition in the limit $\eta \to \infty$. In this appendix, we introduce an alternative condition on v and π that does not depend on the horizon or time and is sufficient for (12) to hold when v is an upper-bounded solution of the BE.

Theorem C.1 (Boundary Condition) *Suppose there exists a continuous function $v : \mathcal{X} \to \mathbb{R}$ satisfying the integral BE* (10) *or, with $v \in \mathrm{C}^1$, the differential BE* (11) *for a policy π. If v is upper-bounded (by zero if $\gamma = 1$) and there exists $\kappa > 0$ s.t. $r_\pi \leqslant \kappa \cdot v$, then v satisfies the boundary condition* (12)*.*

Proof. Suppose v satisfies the integral BE (10) without loss of generality (or, convert the differential BE (11) to (10) via Lemma 2.3 and fix $\eta > 0$). Then, the repetitive applications of (10) to itself k-times result in

$$v(x) = \mathbb{G}_\pi^x\big[\mathfrak{R}_\eta + \gamma^\eta\cdot v(X_\eta)\big] = \mathbb{G}_\pi^x\big[\mathfrak{R}_{2\eta} + \gamma^{2\eta}\cdot v(X_{2\eta})\big] = \;\cdots\; = \mathbb{G}_\pi^x\big[\mathfrak{R}_{k\cdot\eta} + \gamma^{k\cdot\eta}\cdot v(X_{k\cdot\eta})\big] \qquad \forall x \in \mathcal{X}.$$

Rearranging it and substituting $r_\pi \leqslant \kappa \cdot v$ into $\mathfrak{R}_{k\cdot\eta}$ $(= \int_0^{k\cdot\eta} \gamma^t \cdot r(X_t,\, U_t)\; dt)$, we obtain for $J_t(x) \doteq -\mathbb{G}_\pi^x[\gamma^t\, v(X_t)]$:

$$J_{k\cdot\eta}(x) \leq J_0(x) - \kappa \int_0^{k\cdot\eta} J_t(x)\; dt \qquad \forall x \in \mathcal{X} \quad \forall k \in \mathbb{N}$$

where the dependencies on π and γ are implicit. Then, the application of Grönwall-Bellman inequality (Khalil, 2002) yields

$$-e^{-\alpha\cdot(k\cdot\eta)}\cdot M \leq J_{k\cdot\eta}(x) \leq J_0(x)\cdot e^{-\kappa\cdot(k\cdot\eta)} \qquad \forall x \in \mathcal{X} \quad \forall k \in \mathbb{N}$$

for an upper-bound $M \in \mathbb{R}$ of v (note: $\gamma = e^{-\alpha}$). Take $M = 0$ whenever $\gamma = 1$. Then, since $\kappa > 0$, $\alpha \geq 0$, and $M = 0$ whenever $\alpha = 0$, both left and right sides converge to zero as $k \to \infty$, resulting in $\lim_{k\to\infty} J_{k\cdot\eta}(x) = 0$ for all $x \in \mathcal{X}$. □

Combining Theorems 2.5 and C.1, we obtain the following Corollary.

Corollary C.2 (Policy Evaluation) *Under the given conditions in Theorem C.1, π is admissible and $v = v_\pi$.*

In short, if $v \in \mathrm{C}^1$ is a solution to the BE (10) or (11) and upper-bounded (by zero if $\gamma = 1$), then the inequality $r_\pi \leqslant \kappa \cdot v$ for some $\kappa > 0$ implies the boundary condition (12) (but not vice versa), hence $v = v_\pi \in \mathcal{V}_{\mathsf{a}}$. The conditions in Theorem C.1 are particularly related to the optimal control framework in §5.4 but can be also applied to any case in this paper to replace the boundary condition (12). For example, the boundary condition (29) can be replaced by the following two conditions:

(1) there exists $\kappa_i > 0$ s.t. $r_{\pi_{i-1}} \leqslant \kappa_i \cdot v_i$,

(2) v_i is upper-bounded (by zero if $\gamma = 1$).

Under these conditions, Corollary C.2 can replace Theorem 2.5 and the boundary condition (29), in the proofs and statements of the Theorems (e.g., see Theorem 4.1 in §4; also Theorem 5.18 in §5.4, with Theorem 5.16; for their proofs see §§I.2 and I.3).

D Existence and Uniqueness of the Maximal Function u_*

This appendix provides the details about the existence and uniqueness of the maximal function u_* in §2.3 satisfying

$$u_*(x,p) \in \arg\max_{u\in\mathcal{U}} h(x,u,p) \;\; \forall (x,p) \in \mathcal{X}\times\mathcal{X}^{\mathsf{T}} \tag{14}$$

by which a maximal policy π' over $\pi \in \Pi_{\mathsf{a}}$ defined as a continuous function $\pi' : \mathcal{X} \to \mathcal{U}$ s.t.

$$\pi'(x) \in \arg\max_{u\in\mathcal{U}} h(x,u,\nabla v_\pi(x)) \;\; \forall x \in \mathcal{X} \tag{15}$$

can be represented as a closed form:

$$\pi'(x) = u_*(x, \nabla v_\pi(x)). \tag{16}$$

(1) **(Existence)** If $\mathcal{U}$ is *compact*, then for each $(x,p) \in \mathcal{X} \times \mathcal{X}^{\mathsf{T}}$, the maximum of the function $u \mapsto h(x,u,p)$ exists by continuity of the Hamiltonian function h (Rudin, 1964, Theorem 4.16). That is, *a function* $u_* : \mathcal{X} \times \mathcal{X}^{\mathsf{T}} \to \mathcal{U}$ *satisfying* (14) *always exists whenever* $\mathcal{U}$ *is compact.*

(2) **(Uniqueness)** If $\mathcal{U}$ is *convex* and the function $u \mapsto h(x,u,p)$ is *concave* and C^1 for each $(x,p) \in \mathcal{X} \times \mathcal{X}^{\mathsf{T}}$, then the maximization (14) falls into a convex optimization in which any regular point $\bar{u} \in \mathcal{U}^{\mathrm{o}}$ such that

$$\partial h(x,\bar{u},p)/\partial \bar{u} = 0,$$

if exists, belongs to the argmax-set in (14) (Sundaram, 1996, Theorem 7.15) and thus can be the maximal argument $u_*(x,p)$ satisfying (14). In this case, $\pi'(x)$ in (15) corresponds to a regular point $\bar{u}$ for $p = \nabla v_\pi(x)$. Besides, as exemplified in §5.1, if $u \mapsto h(x,u,p)$ is *strictly concave*, then such a regular point $\bar{u}$, if exists, is unique, meaning that $u_*(x,p)$ in (14) is determined *uniquely* (Sundaram, 1996, Theorems 7.14 and 7.15), hence so is each $\pi'(x)$ by (16).

E Theory of Optimality

This appendix provides a theory of optimality regarding (i) an HJB solution (v_*, π_*):

$$\forall x \in \mathcal{X} : \begin{cases} \alpha \cdot v_*(x) = \max\limits_{u \in \mathcal{U}} h(x,u,\nabla v_*(x)) & (17) \\ \pi_*(x) \in \arg\max\limits_{u \in \mathcal{U}} h(x,u,\nabla v_*(x)) & (18) \end{cases}$$

and (ii) a fixed point v^* of $\mathcal{T}$ (i.e., $v^* \in \mathcal{V}_{\mathsf{a}}$ s.t. $\mathcal{T}v^* = v^*$). Here, note that a fixed point v^* of $\mathcal{T}$ is always a solution to the HJBE (17) by Proposition 4.3 (but not vice versa). Hence, if every solution v_* to the HJBE (17) is proven to be optimal, then so is every fixed point v^* of $\mathcal{T}$. We first state the following theorem regarding the optimality of the HJB solution (v_*, π_*).

Theorem E.1 (Optimality) *If a solution* $v_* \in \mathrm{C}^1$ *to the HJBE* (17) *exists and is upper-bounded (by zero if* $\gamma = 1$*), then for any policy* π_* *satisfying* (18)*,*

a. π_* *is admissible and* $v_* \leqslant v_{\pi_*}$*;*

b. $v_\pi \leqslant v_*$ *if* π *satisfies the boundary condition:*

$$\lim_{t\to\infty} \mathbb{G}_\pi^x\big[\,\gamma^t \cdot v_*(X_t)\,\big] = 0 \qquad \forall x \in \mathcal{X} \tag{E.1}$$

(conversely, (E.1) *is true if* π *is admissible and* $v_\pi \leqslant v_*$*);*

c. $v_* = v_{\pi_*}$ *if either (i) the boundary condition* (E.1) *is true for* $\pi = \pi_*$ *or (ii)* $r_{\pi_*} \leqslant \kappa \cdot v_*$ *holds for a constant* $\kappa > 0$*;*

d. (v_*, π_*) *is optimal if* $v \leqslant v_*$ *for any* $v \in \mathcal{V}_{\mathsf{a}}$.

Proof. **a**. Substituting (18) into the HJBE (17), we have

$$\alpha \cdot v_*(x) = h(x, \pi_*(x), \nabla v_*(x)) \qquad \forall x \in \mathcal{X}. \tag{E.2}$$

Then, π_* is admissible and $v_* \leqslant v_{\pi_*}$ by Lemma 2.6.

b. By the HJBE (17), v_* and any policy π satisfy

$$\alpha \cdot v_*(x) \geq h(x, \pi(x), \nabla v_*(x)) \qquad \forall x \in \mathcal{X},$$

hence if π satisfies (E.1), then applying Lemma 2.3 and taking the limit $\eta \to \infty$ results in

$$v_*(x) \geq \underbrace{\lim_{\eta\to\infty} \mathbb{G}_\pi^x\left[\int_0^\eta \gamma^t \cdot R_t\, dt\right]}_{=v_\pi(x)} + \underbrace{\lim_{\eta\to\infty} \mathbb{G}_\pi^x[\gamma^\eta \cdot v_*(X_\eta)]}_{=0} = v_\pi(x) \qquad \forall x \in \mathcal{X}$$

which proves the second statement. Conversely, if π is admissible and $v_\pi \leqslant v_*$, then Proposition 2.4 and the upper-boundedness of v_* (by zero if $\gamma = 1$) results in

$$0 = \lim_{t\to\infty} \mathbb{G}_\pi^x\big[\,\gamma^t \cdot v_\pi(X_t)\,\big] \leq \lim_{t\to\infty} \mathbb{G}_\pi^x\big[\,\gamma^t \cdot v_*(X_t)\,\big] \leq \sup_{x\in\mathcal{X}} v_*(x) \cdot \lim_{t\to\infty} \gamma^t \leq 0$$

implying the boundary condition (E.1).

c. The application of Theorems 2.5 and Corollary C.2 to (E.2) directly proves $v_* = v_{\pi_*}$ under the respective conditions.

d. By the first part and the condition, π_* is admissible and $v \leqslant v_* \leqslant v_{\pi_*}$ for any $v \in \mathcal{V}_\mathsf{a}$; substituting $v = v_{\pi_*}$ results in $v_* = v_{\pi_*}$, which and the condition completes the proof. □

Under the upper-boundedness of $v_* \in \mathrm{C}^1$ in Theorem E.1, any policy π_* given by (18) dominates all policies π's s.t. the boundary condition (E.1) holds for v_*. On the other hand, certain additional conditions (e.g., (E.1) holds for all admissible policies π's) are required in order to have the optimality condition "$v \leqslant v_*$ for all $v \in \mathcal{V}_\mathsf{a}$" in Theorem E.1d (e.g., see case studies in §E.2 and G.2)

E.1 Sufficient Conditions for Optimality

Based on the properties of PIs — convergence (Theorems 4.2, 4.5, 4.6, and 4.9) and monotonicity (Theorem 4.1) — we provide sufficient conditions for optimality, where the notion of "optimality" can be interpreted in a weaker sense than or in a similar manner to that shown in Theorem 2.8 (e.g., see (E.4)). In the latter case, once v_* is the optimal VF, any policy π_* satisfying (18) comes to be optimal ($\because v_* \leqslant v_{\pi_*}$ by Theorem 2.7 and $v_{\pi_*} \leqslant v_*$ by optimality, hence $v_* = v_{\pi_*}$). Specifically, we establish the notions of weak and strong optimality with the following convergence properties introduced in §4.1:

- (C1) **(weak convergence)** $\mathcal{T}^{i-1} v \to v_*$ in a metric;
- (C2) **(strong convergence)** $\mathcal{T}^{i-1} v \to v_*$ locally uniformly;
- (C3) **(additional convergence)** $\nabla(\mathcal{T}^{i-1} v) \to \nabla v_*$ locally uniformly and $\pi_i \to \pi_*$ pointwise,

where we replaced v_i with $\mathcal{T}^{i-1} v$ and $v_1 = v$.

First, we show that Assumption 4.4 alone is sufficient for v^* therein to be weak optimal, i.e., optimal in a metric.

Corollary E.2 *Under Assumption 4.4, there exists a metric d on $\mathcal{V}_\mathsf{a}$ s.t. $\mathcal{T}$ is a contraction under d and for every $v \in \mathcal{V}_\mathsf{a}$,*

$$v \leqslant \mathcal{T} v \leqslant \mathcal{T}^2 v \leqslant \cdots \leqslant \mathcal{T}^N v \xrightarrow{N \to \infty} v^*, \tag{E.3}$$

where the convergence is in the metric d.

Proof. Apply Theorems 4.1 and 4.5. □

Corollary E.2 characterizes v^* as the optimal VF in the weak sense (C1) — as the unique limit point in a metric d, of every monotonically increasing sequence of VFs generated by applying $\mathcal{T}$ recursively (or one of the PI methods). Under the metric d, $\mathcal{T}$ is continuous since it is a contraction. Also, note that v^* is a solution v_* to the HJBE (17) by Proposition 4.3.

Although the weak optimality of v^* in Corollary E.2 looks reasonable, the downside is that convergence (E.3) and continuity of $\mathcal{T}$ are w.r.t. an *unknown* metric d. With continuity of $\mathcal{T}$ under the uniform pseudometric d_Ω, a stronger characterization of v^* is possible, as shown in the next corollary.

Corollary E.3 *If $\lim_{N\to\infty} \mathcal{T}^N v \in \mathcal{V}_\mathsf{a}$ for every $v \in \mathcal{V}_\mathsf{a}$ and for each compact subset Ω of $\mathcal{X}$, $\mathcal{T}$ is continuous under d_Ω, then under Assumption 4.4, $v \leqslant v^*$ for every $v \in \mathcal{V}_\mathsf{a}$.*

Proof. Note that $\mathcal{T}^N v_1 = v_{N+1} \to \hat{v}_*$ pointwise by Theorem 4.2a and then apply Theorems 4.1 and 4.6. □

Under the given conditions on $\mathcal{T}$, Corollary E.3 states that v^* in Assumption 4.4 is truly the optimal VF over $\mathcal{V}_\mathsf{a}$, i.e., over all admissible VFs. This characterization of optimality:

$$v^* \in \mathcal{V}_\mathsf{a} \text{ and } v \leqslant v^* \text{ for every } v \in \mathcal{V}_\mathsf{a} \tag{E.4}$$

is exactly the same as that in Theorem 2.8 and obviously stronger than that in Corollary E.2. Moreover, under (E.4), the uniqueness of the fixed point v^* of $\mathcal{T}$ can be replaced by that of the solution v_* to the HJBE over $\mathcal{V}_\mathsf{a}$ as shown below.

Proposition E.4 *Suppose $v_* \in \mathcal{V}_\mathsf{a}$ is the optimal VF. Then it is a fixed point of $\mathcal{T}$. Moreover, such a fixed point is unique if so is the solution of the HJBE* (17) *over $\mathcal{V}_\mathsf{a}$.*

Proof. Let $v_* \in \mathcal{V}_\mathsf{a}$ be the optimal VF. Then, it satisfies the HJBE (17) by Theorem 2.8, hence we have $v_* \leqslant \mathcal{T} v_*$ by Theorem 2.7. By optimality, $\mathcal{T} v_* \leqslant v_*$ is obvious. Therefore, $v_* = \mathcal{T} v_*$, i.e., v_* is a fixed point of $\mathcal{T}$. Next, suppose v_* is the unique solution to the HJBE (17), but there exists another VF $v'_* \neq v_*$ s.t. $v'_* = \mathcal{T} v'_*$. Then, v'_* is a solution to the HJBE by Proposition 4.3 and thus by the uniqueness, $v'_* = v_*$, a contradiction. Therefore, if v_* is a unique solution to the HJBE (17) over $\mathcal{V}_\mathsf{a}$, then it is a unique fixed point of $\mathcal{T}$. □

Corollary E.5 *Suppose that Assumption 4.8 holds for any initial admissible policy π_0. Then, under Assumptions 4.7 and 4.11, the HJBE* (17) *has a unique solution v_* over C^1 s.t.*

(1) optimality (E.4) *and Assumption 4.4 are true for $v^* = v_*$;*

(2) for each initial admissible policy π_0, there exists a function π_* s.t. (18) *holds and the generated VFs and policies satisfy the stronger convergence, i.e.,* (C2) *and* (C3)*.*

Proof. Theorems 4.1 and 4.9 imply that for a given admissible initial policy π_0, there exists a solution $v_* \in \mathrm{C}^1$ to the HJBE (17) s.t. (i) $v_{\pi_0} \leqslant v_*$ and (ii) the convergence (C2) and (C3) hold for a function π_* satisfying (18). Since the solution v_* is now unique over C^1 by Assumption 4.11 and π_0 is arbitrary, the former implies that $v \leqslant v_*$ for any $v \in \mathcal{V}_\mathsf{a}$. Moreover, by Theorem E.1d, v_* is the optimal VF and thus satisfies the strong optimality (E.4); since $\mathcal{V}_\mathsf{a} \subset \mathrm{C}^1$ by (3), v_* is the unique solution of the HJBE over $\mathcal{V}_\mathsf{a}$ ($\subset \mathrm{C}^1$). Therefore, v_* is the unique fixed point of $\mathcal{T}$ (i.e., Assumption 4.4 holds for $v^* = v_*$) by Proposition E.4, which completes the proof. □

Note that under the given conditions in Corollary E.3 or E.5, (E.3) holds with *locally uniform convergence* (C2) (apply Theorem 4.1 for monotonicity) — stronger than convergence (C1) in a metric shown in Corollary E.2. In addition, Corollary E.5 provides the additional convergence (C3) without employing the PI operator $\mathcal{T}$ and any assumptions imposed on it. We note that even the stronger (i.e., locally uniform) convergence of $\langle \pi_i \rangle$ than that in (C3) can be obtained in the concave Hamiltonian formulation in §5.1, with both Assumption 4.7 and 4.8b for any $\pi_0 \in \Pi_\mathsf{a}$ in Corollary E.5 *relaxed*.

In summary, we characterize v_* in the Corollaries as a *unique* VF to which $\langle \mathcal{T}^N v \rangle$ for any $v \in \mathcal{V}_\mathsf{a}$ monotonically converges (i.e., satisfies (E.3) for $v^* = v_*$) in their respective manners. Here, the uniqueness was assumed and is truly necessary — otherwise, some sequence of VFs generated by PIs may converge to another VF $v'_* \neq v_*$. In this case, the optimality of v_* becomes vague and not decidable unless $v'_* \leqslant v_*$ for any of such VFs v'_*. Since an optimal VF v_* is unique over $\mathcal{V}_\mathsf{a}$ as discussed in §2.4, any two different VFs $v_*, v'_* \in \mathcal{V}_\mathsf{a}$ cannot be the optimal at the same time.

A similar characterization of v_* is possible without assuming the uniqueness of the solution to the HJBE and without the assumptions on PI, but by proving or imposing (i) the boundary condition (E.1) for a class of policies and (ii) one of the two conditions on (v_*, π_*) in Theorem E.1c. This approach will be used in the next subsection (§E.2) to characterize the optimality of v_* (and π_*) under the given respective frameworks and conditions therein, without assuming the uniqueness of v_*.

E.2 Case Studies of Optimality

We now provide and discuss the condition(s) for optimality of the HJB solution (v_*, π_*) under certain classes of RL problems shown in §5 Case Studies — specifically, the cases presented in §§5.1, 5.2, and 5.4.

Concave Hamiltonian Formulation (§5.1). Under (30) and (31), Corollary E.5 can be simplified and strengthened with the assumptions on the policies and policy improvement relaxed.

Corollary E.6 *If Assumption 4.8a holds for any initial admissible policy π_0, then under* (30)*,* (31)*, and Assumption 4.11, there exists a unique HJB solution (v_*, π_*) over $\mathcal{V}_\mathsf{a} \times \Pi_\mathsf{a}$ s.t. Assumption 4.4 holds for $v^* = v_*$, $\pi \preccurlyeq \pi_*$ for all $\pi \in \Pi_\mathsf{a}$, $v_* = v_{\pi_*}$, and for any initial admissible policy π_0, $v_i \to v_*$, $\nabla v_i \to \nabla v_*$, and $\pi_i \to \pi_*$, all locally uniformly.*

Proof. Combine Lemma I.7 with Corollary E.5. Also note that the HJB policy π_* satisfying (18) is uniquely determined under (30) and (31) — see §5.1.1 for details. □

Here, we have directly extended Corollary E.5 to E.6 above in the same way as extending Theorem 4.9 to 5.1, by applying Lemma I.7 in the concave Hamiltonian formulation (30) and (31). Therefore, as discussed in Remark 5.3 and §5.1.2, Corollary E.6 (specifically, Lemma I.7) can be further extended to

(1) the input-affine case where the reward function r satisfies the conditions in Remark 5.3 and $(x, \mathfrak{u}) \mapsto \sigma^x(\mathfrak{u})$ is continuous;

(2) the non-affine case (39) and (40) in a similar manner to Theorem 5.4, even when φ and $\mathfrak{c}$ depend on the state $x \in \mathcal{X}$.

Discounted RL Problems with Bounded VFs (§§5.2). In this case, we can dramatically improve the optimality theory with respect to the solution v_* to the HJBE (17) and the HJB policy π_* in (18) (of course, under the Assumptions made in §2).

Theorem E.7 *Let $\gamma \in (0,1)$. If the HJBE* (17) *has a bounded* C^1 *solution v_*, then for any HJB policy π_* satisfying* (18)*,*

(1) v_{π_*} is bounded (hence, admissible) and $v_* = v_{\pi_*}$;

(2) $\pi \preccurlyeq \pi_*$ for any admissible policy π.

Moreover, v_* is the unique solution to the HJBE (17) *over all bounded* C^1 *functions $v : \mathcal{X} \to \mathbb{R}$.*

Proof. The first two parts can be proven by applying Proposition 5.6 with $v = v_*$ and Theorem E.1b and c. For the uniqueness of v_*, note that if v'_* is another bounded C^1 solution to the HJBE, then we have $v_* \leqslant v'_*$ and $v'_* \leqslant v_*$, hence $v_* = v'_*$. □

Nonlinear Optimal Control (§5.4). Under the assumptions and notations in §5.4, the optimality of an HJB solution (J_*, π_*), with $J_* \doteq -v_*$, can be characterized as follows, without assuming the existence and uniqueness of the state trajectories.

Theorem E.8 *Under the assumptions and notations in §5.4, if there exists a HJB solution (v_*, π_*) of* (17) *and* (19) *s.t.*

(1) J_* is $\mathrm{C}^1_{\mathsf{Lip}}$, positive definite, and radially unbounded;

(2) $\alpha\kappa \cdot J_* \leqslant c_{\pi_*}$ for some $\kappa > 0$,

then, $\pi_* \in \Pi_{\mathsf{a}}$, $J_* = J_{\pi_*}$, and $J_* \leqslant J_\pi$ for any $\pi \in \Pi$ such that

$$\lim_{t\to\infty} \mathbb{G}^x_\pi[\gamma^t \cdot J_*(X_t)] = 0 \qquad \forall x \in \mathcal{X}. \tag{E.5}$$

Proof. The HJB policy π_* satisfy (19); v_* $(= -J_*)$ is $\mathrm{C}^1_{\mathsf{Lip}}$ and negative definite. Hence, $\pi_* \in \Pi_0$ by Lemma I.12. Moreover, the HJBE (17), (18), and the positive definiteness of c, with $J_* = -v_*$ and $c = -r$, imply that

$$\dot{J}_*(x, \pi_*(x)) = \alpha \cdot J_*(x) - c_{\pi_*}(x) \leq \alpha \cdot J_*(x) \quad \forall x \in \mathcal{X}.$$

J_* is continuous, positive definite, and radially unbounded. Hence, by Lemma I.11, there exist $\mathcal{K}_\infty$ functions ρ_1 and ρ_2 s.t.

$$\rho_1(\|x\|) \leq J_*(x) \leq \rho_2(\|x\|) \qquad \forall x \in \mathcal{X}.$$

Therefore, the application of Lemma I.10 proves that $\pi_* \in \Pi$. The remaining proof is divided into two folds.

(1) When "$\gamma = 1$," the HJBE (17) and (18) is reduced to $\dot{J}_*(x, \pi_*(x)) = -c_{\pi_*}(x)$ $\forall x \in \mathcal{X}$, where c_{π_*} is positive definite by Lemma 5.11. Therefore, $x_{\mathsf{e}} = 0$ under π_* is globally asymptotically stable (Khalil, 2002, Theorem 4.2), with J_* as the radially-unbounded Lyapunov function, and Theorem 5.15 results in $\pi_* \in \Pi_{\mathsf{a}}$ and $J_* = J_{\pi_*}$.

(2) For $\gamma \in (0,1)$ (i.e., $\alpha > 0$), "$\alpha\kappa J_* \leqslant c_{\pi_*}$" can be rewritten as "$\kappa_* J_* \leqslant c_{\pi_*}$ for $\kappa_* \doteq \alpha\kappa > 0$." Since the HJBE (17) and (18) imply the differential BE (11) for $v = v_*$ and $\pi = \pi_*$, we have $\pi_* \in \Pi_{\mathsf{a}}$ and $J_* = J_{\pi_*}$ by Theorem 5.16.

For both cases, we have $\pi_* \in \Pi_{\mathsf{a}}$ and $J_* = J_{\pi_*}$; the last part "$J_* \leqslant J_\pi$ $\forall \pi \in \Pi$ s.t. (E.5) holds" is obvious by Theorem E.1b. □

The conditions on (J_*, π_*) in Theorem E.8 can be considered a limit version of the three conditions presented in §5.4:

(1) $\pi_0 \in \Pi_{\mathsf{a}}$,
(2) $J_i \in \mathrm{C}^1_{\mathsf{Lip}}$ is positive definite and radially unbounded,
(3) there exists $\kappa_i > 0$ such that $\alpha\kappa_i \cdot J_i \leqslant c_{\pi_{i-1}}$.

So, similarly to the third one, (i) the condition $\alpha\kappa \cdot J_* \leqslant c_{\pi_*}$ is always true if $\gamma = 1$ by $\alpha = 0$ and (43); (ii) in the discounted case $\gamma \in (0,1)$, it can reduce to a simpler one: $\kappa_* J_* \leqslant c_{\pi_*}$ for some $\kappa_* > 0$; (iii) when $\kappa \in (0,1)$, the condition $\alpha\kappa \cdot J_* \leqslant c_{\pi_*}$ is weaker than both of the stability conditions $\alpha J_* \leqslant c_{\pi_*}$ and (E.6) below corresponding to (44) and (45), respectively.

Remark. *Suppose $\pi_* \in \Pi_{\mathsf{a}}$ and $J_* = J_{\pi_*}$ in Theorem E.8 are true. Then, $x_{\mathsf{e}} = 0$ under π_* is asymptotically stable if*

$$\alpha J_*(x) < c_{\pi_*}(x) \qquad \forall x \in \mathcal{X} \setminus \{0\} \tag{E.6}$$

by Theorem 5.13. Note that (E.6) *is weaker than the stability condition given by Gaitsgory et al. (2015, Assumption 2.3):*

$$\kappa_* J_*(x) \leq c(x,u) \qquad \forall (x,u) \in \mathcal{X} \times \mathcal{U}, \qquad \textit{for some } \kappa_* > \alpha.$$

This inequality and the positive definiteness of J_{π_*} (by Lemma 5.12a) imply $\alpha J_*(x) < \kappa_* J_*(x) \le c_{\pi_*}(x)$ for all $x \in \mathcal{X} \setminus \{0\}$ and thus (E.6) *(but not vice versa). The other condition given by Gaitsgory et al. (2015, Assumption 3.8) for global asymptotic stability can be replaced with the radial unboundedness of J_* (see Lemma I.11 in §I).*

Remark. *The boundary condition* (E.5) *is true for all policies $\pi \in \Pi$ s.t. $\lim_{t\to\infty} \mathbb{G}_\pi^x[X_t] = 0\ \forall x \in \mathcal{X}$, including those that globally asymptotically stabilizes the equilibrium point $x_{\mathsf{e}} = 0$, as in the proof of Theorem 5.15 (see §I.3). On the other hand, when discounted,* (E.5) *contains the cases where the state trajectories are globally bounded as in §G.2 or even diverge exponentially such as in the discounted LQR case in §G.3.*

F A Pathological Example (Kiumarsi et al., 2016)

Presented in this appendix is a counter-example where the dynamics is simple but non-affine, and the design of the reward function r is critical. In this example, (i) a naive choice of r fails to give a closed-form solution of policy improvement and the HJBE; (ii) in the unconstrained case, such a choice results in a pathological Hamiltonian h such that the solutions (i.e., π' in (15) for $\pi \in \Pi_{\mathsf{a}}$, v_* in the HJBE (17), and π_* in (18)) do not exist. We encourage the readers to read §D beforehand and for a technique to resolve such an issue, see also §5.1.2.

Consider the scalar dynamics ($l = m = 1$) with the action space $\mathcal{U} = [-u_{\mathsf{max}}, u_{\mathsf{max}}]$ for $u_{\mathsf{max}} \in (0, \infty]$:

$$\dot{X}_t = X_t^3 + U_t^3.$$

Suppose that the reward function r given by (31) and (35) with $\Gamma = 1$, that is, $r(x,u) = \mathfrak{r}(x) - \mathfrak{c}(u)$ for a continuous function $\mathfrak{r} : \mathcal{X} \to \mathbb{R}$ and $\mathfrak{c} : \mathcal{U} \to \mathbb{R}$ given by

$$\mathfrak{c}(u) = \lim_{v \to u} \int_0^v (s^{\mathsf{T}})^{-1}(\mathfrak{u})\, d\mathfrak{u}.$$

Then, the Hamiltonian $h : \mathbb{R} \times \mathcal{U} \times \mathbb{R} \to \mathbb{R}$ in this case is given by

$$h(x,u,p) = \mathfrak{r}(x) - \mathfrak{c}(u) + p \cdot (x^3 + u^3). \tag{F.1}$$

(Input-constrained Case) First, we consider $\mathcal{U} = [-1, 1]$ with $s = \tanh$. In this case, since $\mathcal{U}$ is compact, the maximal function $u_*(x,p)$ satisfying (14) for each $(x,p) \in \mathbb{R}^2$ exists (see §D). However, a regular point $u \in (-1,1)$ s.t.

$$\partial h(x,u,p)/\partial u = -\tanh^{-1} u + 3pu^2 = 0 \tag{F.2}$$

cannot be expressed in a closed form since (F.2) is nonlinear in u.

(Unconstrained Case) Next, consider (36), that is, $u_{\mathsf{max}} = \infty$ and $s(u) = u/2$. In this case, the maximal function u_* does not exist since $\mathfrak{c}(u) = u^2$ and thus for any $p > 0$ and $x \in \mathbb{R}$, the Hamiltonian (F.1) satisfies

$$\lim_{u\to\infty} h(x,u,p) = \lim_{u\to-\infty} h(x,u,-p) = \infty.$$

Therefore, except the trivial cases $\nabla v_\pi = 0$ and $\nabla v_* = 0$, the maximal policy π' in (15) and the solution v_* to the HJBE (17) (and accordingly, π_* in (18)) fail to exist since so do the maxima in those respective equations. Note that the regular points u s.t. $\partial h(x,u,p)/\partial u = 0$ explicitly given by $u = 0$ and $u = 2/(3p)$ are the local maximum and the local minimum, respectively, but the global maximum does not exist in this case.

The issue in both cases above is that even though $\mathfrak{c}$ is strictly convex, h *is not (strictly) concave* due to the cubic term u^3 in the dynamics $f(x,u) = x^3 + u^3$. This means that the uniqueness of u_* is not guaranteed, and the existing regular points u's satisfying $\partial h(x,u,p)/\partial u = 0$ are not necessarily the maximum of the Hamiltonian $h(x,u,p)$ (see §D).

G Additional Case Studies

This appendix provides the additional case studies with (strong) connections to case studies in §5 and our theory on PIs presented in the main work (Lee and Sutton, 2020b) and §E.

G.1 General Concave Hamiltonian Formulation

Here, we generalize the methods and results in §5.1.1 to the original RL problem (1). The core idea is to introduce a continuous bijection $\psi : \mathcal{U}^{\circ} \to \mathbb{R}^m$ (which has a continuous inverse ψ^{-1} by Lemma I.4) and an m-dimensional action-dynamics:

$$\dot{\mathfrak{U}}_t = A_t, \quad A_t \in \mathcal{A} \tag{G.1}$$

where $\mathcal{A} \subseteq \mathbb{R}^m$ is an action space, and the differential action trajectory $t \mapsto A_t$ is a continuous function from $\mathbb{T}$ to $\mathcal{A}$, determining the rate of change of $\mathfrak{U}_t$ for all $t \in \mathbb{T}$, by (G.1); the effective action $\mathfrak{U}_t \in \mathbb{R}^m$ generates the real action U_t by

$$U_t = \psi^{-1}(\mathfrak{U}_t). \tag{G.2}$$

Under (G.1) and (G.2), the results for the concave Hamiltonian formulation in §5.1 can be applied to the RL problem with the affine dynamics

$$\begin{bmatrix} \dot{X}_t \\ \dot{\mathfrak{U}}_t \end{bmatrix} = \begin{bmatrix} f(X_t, \psi^{-1}(\mathfrak{U}_t)) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ I \end{bmatrix} A_t,$$

with $(X_t, \mathfrak{U}_t) \in \mathbb{R}^{l \times m}$ considered as its state and $A_t \in \mathcal{A}$ as the action, and the extended reward function r_{e}:

$$r_{\mathsf{e}}(x, \mathfrak{u}, a) \doteq r(x, \psi^{-1}(\mathfrak{u})) - \mathfrak{c}(a),$$

where the real action $U_t \in \mathcal{U}$ is determined by (G.2). Here, $\mathfrak{c} : \mathcal{A} \to \mathbb{R}$ satisfies the same properties as $\mathfrak{c}$ in (31) and can be $(x, \mathfrak{u})$-dependent in the same way to the x-dependent $\mathfrak{c}$ in (37) (Remark 5.3) — see §5.1. Note that the resulting IPI will be model-free — it does not explicitly depend on the input-coupling dynamics f_{c} in (25) and, of course, f_{d}. When $\mathcal{U} = \mathcal{A} = \mathbb{R}^m$, similar ideas were presented by Murray et al. (2002) for input-affine optimal control and Lee, Park, and Choi (2012) for LQRs.

G.2 Discounted RL with Bounded State Trajectories

When $\gamma \in (0, 1)$ and the state trajectories are bounded, similar properties and results to "§5.2 Discounted RL with Bounded VF" can be obtained as shown below.

Definition. *The state trajectories under π are said to be globally bounded iff for each $x \in \mathcal{X}$, $t \mapsto \mathbb{G}^x_\pi[X_t]$ is bounded over $\mathbb{T}$.*

Proposition G.1 *If the state trajectories under π are globally bounded, and v is continuous, then under $\gamma \in (0, 1)$, they satisfy the boundary condition* (12)*.*

Proof. Since $t \mapsto \mathbb{G}^x_\pi[X_t]$ is bounded and v is continuous, $t \mapsto \mathbb{G}^x_\pi[v(X^x_t)]$ is also bounded, for each $x \in \mathcal{X}$. Hence, the proof can be done by applying Lemma I.9 in §I. □

Corollary G.2 (Policy Evaluation) *Let $\gamma \in (0, 1)$ and the state trajectories under π be globally bounded. Then, π is admissible, and $v = v_\pi$ is the unique solution to the BEs* (10) *and* (11) *over all continuous and* C^1 *functions, respectively.*

Proof. Apply Theorem 2.5 and Proposition G.1. □

By Corollary G.2, as long as the state trajectories under π_{i-1} are globally bounded and $\gamma \in (0, 1)$, π_{i-1} is admissible, and the ith step of the PI methods can run without assuming the boundary condition (29) that is shown to be true by Proposition G.1 in this case. On the other hand, the VF is not necessarily bounded in this case (see the next example, LQR, in §G.3 in which the admissible VF is always quadratic), and it is a bit unclear when and how the state trajectories are bounded. Some stability-related conditions sufficient for global boundedness of the state trajectories are:

(1) input-to-state stability (Khalil, 2002, Definition 4.7), ensuring that the state trajectories are globally bounded under *any given policy* whenever $\mathcal{U}$ is bounded;

(2) global asymptotic stability (e.g., see nonlinear optimal control in §5.4 and the LQR in §G.3);

(3) global ultimate boundedness of the state trajectories (Khalil, 2002, Definition 4.6), which is stronger than the global boundedness of the state trajectories but weaker than global asymptotic stability.

In general, stability of the system (1) implies boundedness of the state trajectories within some region, but *not vice versa*.

Note that global boundedness of the state trajectories under π, including the above three special cases, guarantees their global existence and uniqueness over the entire time interval, under locally Lipschitz f_π, as can be shown by applying the following proposition for all $x \in \mathcal{X}$ (see also §5.3 for a general case with locally Lipschitz f_π, and §§5.4 and G.3 for its applications).

Proposition G.3 *Let f_π be locally Lipschitz and $x \in \mathcal{X}$. If there exists a compact subset $\Omega_x \subset \mathcal{X}$ s.t. $t \mapsto \mathbb{G}_\pi^x[X_t]$ lies entirely in Ω_x, then $t \mapsto \mathbb{G}_\pi^x[X_t]$ is uniquely defined over $\mathbb{T}$.*

Proof. See (Khalil, 2002, Theorem 3.3) or (Haddad and Chellaboina, 2008, Corollary 2.5). □

The HJB solution (v_*, π_*) can be also characterized in the discounted case as the optimal solution among all the policies that make the state trajectories globally bounded.

Corollary G.4 *Suppose $\gamma \in (0,1)$ and the HJBE* (17) *has an upper-bounded solution $v_* \in \mathrm{C}^1$. Then,*

(1) π_* is admissible and $v_* \geqslant v_{\pi_*}$ for any policy π_* s.t. (18) *holds;*

(2) moreover, if the state trajectories under π (resp. π_*) are globally bounded, then $v_\pi \leqslant v_*$ (resp. $v_* = v_{\pi_*}$).

Proof. Obvious by Theorem E.1a–c and Proposition G.1 with $v = v_*$. □

G.3 Linear Quadratic Regulations (LQRs)

A linear quadratic regulation (LQR) consists of

$$\begin{cases} \text{a linear dynamics: } f(x,u) = A^0 x + Bu, \\ \text{the unconstrained action space: } \mathcal{U} = \mathbb{R}^m, \\ \text{a quadratic positive cost function: } c(x,u) = \begin{bmatrix} x^\mathsf{T} & u^\mathsf{T} \end{bmatrix} \mathcal{W} \begin{bmatrix} x \\ u \end{bmatrix} \geq 0, \text{ with } \mathcal{W} \doteq \begin{bmatrix} S & E \\ E^\mathsf{T} & \Gamma \end{bmatrix}, \end{cases} \tag{G.3}$$

where (A^0, B, S) for $A^0 \in \mathbb{R}^{l \times l}$, $B \in \mathbb{R}^{l \times m}$, $S \in \mathbb{R}^{l \times l}$ is stabilizable and observable, $\mathcal{W} \in \mathbb{R}^{(l+m)\times(l+m)}$ is positive semidefinite and nondegenerate,[10] and $\Gamma \in \mathbb{R}^{m \times m}$ is positive definite. Note that the LQR (G.3) falls into a special case of the nonlinear optimal control in §5.4 whenever the matrix $\mathcal{W}$ is positive definite. On other other hand, f^x is affine and r^x is strictly concave for each $x \in \mathcal{X}$, with its dynamics satisfying (30) for

$$f_\mathsf{d}(x) = A^0 x \text{ and } F_\mathsf{c}(x) = B$$

and its reward function r $(= -c)$ satisfying (37) for

$$\mathfrak{r}(x) = -x^\mathsf{T} S x \text{ and } \mathfrak{c}(x,u) = u^\mathsf{T}\Gamma u + 2x^\mathsf{T} E u.$$

Moreover, whenever $E = 0$, it becomes (31) with $\mathfrak{c}$ given by $\mathfrak{c}(u) = u^\mathsf{T}\Gamma u$, the unconstrained case "(35) and (36)." Therefore, the LQR (G.3) is an example of the concave Hamiltonian formulation in §5.1.1. Also note that in LQR, f is obviously globally Lipschitz, ensuring the global existence of the unique state trajectories under any globally Lipschitz policy, e.g., the linear one below (Khalil, 2002, Theorem 3.2; see also Chen, 1998).

In an LQR (G.3), if a policy π is linear, i.e., $\pi(x) = -Kx$ $(K \in \mathbb{R}^{m \times l})$, then J_π $(\doteq -v_\pi)$ is quadratic, if finite, and can be expressed as $J_\pi(x) = x^\mathsf{T} P_\pi x$ for a positive definite matrix $P_\pi \in \mathbb{R}^{l \times l}$ (e.g., see Lancaster and Rodman, 1995, Lemma 16.3.2 with Theorem 16.3.3.(d); Lee et al., 2014, Section 2). Moreover, the maximal policy π' in (38) is linear again and can be represented as

$$\pi'(x) = -K'x \text{ with } K' = \Gamma^{-1}(B^\mathsf{T} P_\pi + E^\mathsf{T}).$$

[10] $\mathcal{W}$ is nondegenerate iff $\operatorname{rank}(\mathcal{W}) = \operatorname{rank}(S) + m$, which is true when $\mathcal{W}$ is positive definite or $E = 0$.

Algorithm 4: IPI and DPI for the LQR (G.3)

Initialize: $\pi_0(x) = -K_0 x$, the initial admissible policy; $i \leftarrow 1$;

repeat (under the LQR formulation (G.3))

  Policy Evaluation: given policy $\pi_{i-1}(x) = -K_{i-1}x$, find a quadratic function $v_i(x) = -x^\mathsf{T} P_i x$ such that

    (IPI) v_i satisfies the BE (10) for some $\eta > 0$; or **(DPI)** $P_i \in \mathbb{R}^{l\times l}$ satisfies the matrix formula (G.4);

  Policy Improvement: $K_i \leftarrow \Gamma^{-1}(B^\mathsf{T} P_i + E^\mathsf{T})$;

  $i \leftarrow i + 1$;

until *convergence is met.*

This observation gives IPI and DPI for the LQR (G.3) shown in Algorithm 4, where DPI solves the matrix equation:

$$(A^\alpha_{i-1})^\mathsf{T} P_i + P_i A^\alpha_{i-1} = K^\mathsf{T}_{i-1} E^\mathsf{T} + E K_{i-1} - S - K^\mathsf{T}_{i-1} \Gamma K_{i-1}, \tag{G.4}$$

at each ith step of policy evaluation. Here, we denote

$$A^\alpha_{i-1} \doteq A^\alpha - BK_{i-1} \text{ for } A^\alpha \doteq A^0 - \alpha I/2$$

where $I \in \mathbb{R}^{l\times l}$ denotes the identity matrix. Note that DPI (and IPI — see Theorem G.5a) in Algorithm 4 is equivalent to the existing matrix-form PIs (Arnold III, 1984; Mehrmann, 1991; see also Kleinman, 1968; Lee et al., 2014 for the case $E = 0$). In addition, if $\mathcal{W}$ is positive definite, then rearranging (48) using (G.4) yields the very stability condition:

$$(A^0_{i-1})^\mathsf{T} P_i + P_i A^0_{i-1} \text{ is negative definite,}$$

for J_i $(= -v_i)$ to be the Lyapunov function for the linear dynamics $f(x, u) = A^0 x + Bu$ under the policy $\pi_{i-1}(x) = K_{i-1}x$ (Khalil, 2002, Theorem 4.6).

In fact, if the policy π is linear, the process X^α_t generated by

$$\dot{X}^\alpha_t = A^\alpha X^\alpha_t + BU^\alpha_t \tag{G.5}$$

and $U^\alpha_t = \pi(X^\alpha_t)$ for all $t \in \mathbb{T}$ yields the following expression (G.6) of J_π, *without the discount factor γ (or rate α) in its cumulative cost* (Anderson and Moore, 1989):

$$J_\pi(x) \doteq \mathbb{G}^x_\pi\left[\int_0^\infty e^{-\alpha t}\cdot C_t\, dt\right] = \mathbb{G}^{x,\alpha}_\pi\left[\int_0^\infty C_t\, dt\right],^{11} \tag{G.6}$$

where $\mathbb{G}^{x,\alpha}_\pi[Y]$ means $\mathbb{G}^x_\pi[Y]$ if $\alpha = 0$ but otherwise the value Y with respect to the state $X_t = X^\alpha_t$ and the action $U_t = U^\alpha_t$ for all $t \in \mathbb{T}$; $C_t = c(X_t, U_t)$ is the quadratic cost at time t. Here, (A^α, B, S) is stabilizable and observable since so is (A^0, B, S) (see §I.4). Therefore, any discounted LQR can be transformed into an equivalent undiscounted total one, simply by replacing A^0 with A^α.

After transformation into (G.5) and (G.6), we can see that a linear policy π is admissible *iif* X^α_t under π converges to 0 (see Lancaster and Rodman, 1995, Proposition 16.2.9); the convergence $X^\alpha_t \to 0$ implies that any quadratic function J $(= -v)$, say $J(x) = x^\mathsf{T} P x$ for some $P \in \mathbb{R}^{l\times l}$, satisfies the boundary condition (12) since

$$\mathbb{G}^x_\pi\big[\gamma^t J(X_t)\big] = \mathbb{G}^x_\pi\big[e^{-\alpha t}\cdot X^\mathsf{T}_t P X_t\big] = \mathbb{G}^{x,\alpha}_\pi\big[J(X_t)\big] \longrightarrow 0 \text{ as } t \to \infty.^{11}$$

Therefore, by Theorem 4.1, π_i in Algorithm 4 is admissible and $P_i = P_{\pi_{i-1}}$ for all $i \in \mathbb{N}$, but *without assuming the boundary condition* (29) that is true in LQR, as shown above.

As regards to the HJB solution (v_*, π_*) and the Assumptions in §4, the applications of the LQR theory (Lancaster and Rodman, 1995, Theorem 16.3.3), Proposition E.4, and Lemma I.7a with Remark 5.3 to (G.5) and (G.6) show that

(1) (v_*, π_*) satisfying the HJBE (17) and (18) exists;

[11] The equality comes from the fact: $\mathbb{G}^x_\pi[e^{-\alpha t/2} X_t] = \mathbb{G}^x_\pi[e^{-\alpha t/2} e^{(A^0 - BK)t} X_0] = \mathbb{G}^x_\pi[e^{(A^\alpha - BK)t} X_0] = \mathbb{G}^{x,\alpha}_\pi[X_t]$ for $\pi(x) = -Kx$.

(2) J_* ($\doteq -v_*$) and π_* are optimal and given by

$$\begin{cases} J_*(x) = x^\mathsf{T} P_* x \text{ for a positive definite } P_* \in \mathbb{R}^{l\times l}, \\ \pi_*(x) = -K_* x \text{ with } K_* \doteq \Gamma^{-1}(B^\mathsf{T} P_* + E^\mathsf{T}); \end{cases}$$

(3) Assumptions 4.4, 4.7, and 4.11 are all true.

Applying the theory developed in this work, we also obtain the following result regarding the PIs applied to the LQR.

Theorem G.5 *The sequences $\langle K_i \rangle$ and $\langle P_i \rangle$ generated by Algorithm 4 satisfy the followings:*

a. $\forall i \in \mathbb{N}$*:* $\pi_i(x) = -K_i x$ *is admissible and* $P_i = P_{\pi_{i-1}}$,

b. $0 < P_* \leq \cdots \leq P_{i+1} \leq P_i \leq \cdots \leq P_1$,

c. $\lim_{i\to\infty} P_i = P_*$ *and* $\lim_{i\to\infty} K_i = K_*$.

Proof. First, Theorem 4.1 and the optimality of P_* prove the first and second parts. Next, Theorem 4.2 implies that there exists $P \in \mathbb{R}^{l\times l}$ s.t. $P_i \to P$ (see §I.4). Let $M_\Omega \doteq \sup_{x\in\Omega} \|x\| < \infty$ for a compact subset $\Omega \subset \mathcal{X}$. Then, we have

$$0 \leq \sup_{x\in\Omega} \big\|(P_i - P)x\big\| \leq \sup_{x\in\Omega} \big(|||P_i - P||| \cdot \|x\|\big) = M_\Omega \cdot |||P_i - P|||,$$

where $M_\Omega \cdot |||P_i - P||| \to 0$ by $P_i \to P$. Hence, ∇v_i given by $\nabla v_i(x) = -2x^\mathsf{T} P_i$ converges uniformly on any compact subset of $\mathcal{X}$ and by Lemma I.1, locally uniformly. Finally, by Theorem 5.1 with Remark 5.3, $P = P_*$ and $K_i \to K_*$. □

By extending the existing analytical results to the LQR (G.3), we can see more: *the convergence $P_i \to P_*$ is quadratic* (see §I.4). Therefore, PI methods have faster convergence rates than linear in both discrete and continuous domains: it is finite in a finite MDP (Powell, 2007; Sutton and Barto, 2018) and quadratic in the LQR (G.3). Moreover, it could also imply the local quadratic convergence $v_i \to v_*$ for a class of nonlinear optimal control problems in §5.4 as the nonlinear problem can be approximated near the equilibrium point $(x_\mathsf{e}, u_\mathsf{e}) = (0, 0)$ by an LQR (G.3) with

$$A^0 = \nabla_x f(0,0), \quad B = \nabla_u f(0,0), \quad \mathcal{W} = \nabla^2 c(0,0)$$

whenever the gradient $\nabla f(0,0) \in \mathbb{R}^{l\times(l+m)}$ and the Hessian $\nabla^2 c(0,0) \in \mathbb{R}^{(l+m)\times(l+m)}$ exist; $\nabla_x f$ and $\nabla_u f$ denote the gradients of $f(x,u)$ w.r.t. x and u, respectively. Therefore, the rate of convergence is possibly, locally quadratic for the nonlinear optimal control problem in §5.4 when its linearization $(A^0, B, \mathcal{W})$ above exists and satisfies the assumptions on the LQR shown in this appendix.

H Implementation Details

This appendix provides details of the implementations of the PI methods (i.e., Algorithm 3) experimented in §6.

H.1 Structure of the VF Approximator V_i

Recall that in §6, the solution to the policy evaluation, V_i, is represented by a linear function approximator V as

$$V_i(x) \approx V(x; \theta_i) \doteq \theta_i^\mathsf{T} \phi(x), \tag{49}$$

for its weights $\theta_i \in \mathbb{R}^L$ and features $\phi : \mathcal{X} \to \mathbb{R}^L$, with the number of features $L = 121$. Since the policy improvement needs a differentiable structure, we choose radial basis functions (RBFs) as the features ϕ, rather than using (tile-coded) binary ones (Sutton and Barto, 2018). Hence, the j-th component of the feature vector ϕ is given by

$$\phi_j(x) = \exp\big(-(x - c_j)^\mathsf{T} \Sigma^{-1} (x - c_j)\big)$$

where $\Sigma \doteq \mathrm{diag}\{1, 2\}$ is a weighting matrix, and $\{c_j \in \Omega : 1 \leq j \leq L\}$ is the set of RBF center points c_j that are uniformly distributed within the compact region $\Omega = [-\pi, \pi] \times [-6, 6] \subset \mathcal{X}$. In the simulations in §6, we choose $L = 11 \times 11 = 121$; the set of center points $\{c_j\}$ includes the origin $(0, 0)$ and some boundaries of Ω. Also note that whenever inputting to the features ϕ, the first component x_1 of x is normalized to a value within $[-\pi, \pi]$ by adding $\pm 2\pi k$ to it for some $k \in \mathbb{Z}$.

H.2 Least-Squares Solution of Policy Evaluation

In the experiments in §6, the policy evaluation (or the BE) in Algorithm 3 is solved by batch least squares, over the set of initial states $\{x_k : 1 \le k \le N \times M\}$, uniformly distributed as the $(N \times M)$-grid points over Ω, where N, $M \in \mathbb{N}$ are the total numbers of the grids in the x_1- and x_2-directions, respectively. We chose $N = 20$ and $M = 21$, so the total 420 number of grid points x_k's in Ω are considered to determine the least-squares solution θ_i^* in each policy evaluation, except the DPI variant in Case 4 where we used $M = 20$ instead of 21.

To describe the batch least square solution θ_i^*, note that under the approximation (49), the BEs of the variants of DPI and IPI in Algorithm 3 can be expressed at each point $x = x_k$ as

$$y_i^{\mathsf{T}}(x_k) \cdot \theta_i + \varepsilon_i(x_k) = r(x_k, \pi_{i-1}(x_k)), \tag{H.1}$$

where $\varepsilon_i : \mathcal{X} \to \mathbb{R}$ is the approximation error for each case, and $y_i : \mathcal{X} \to \mathbb{R}^L$ is given by

$$y_i(x) = \begin{cases} \mathbb{G}^x_{\pi_{i-1}}\big[\phi(X_0) - \gamma_{\mathsf{d}} \cdot \phi(X_{\Delta t})\big] & \text{for the variant of IPI,} \\ \alpha_{\mathsf{d}} \cdot \phi(x) - \Delta t \cdot \nabla\phi(x) \cdot f_{\pi_{i-1}}(x) & \text{for the variant of DPI.} \end{cases}$$

Concatenating the vectors as and denoting them by

$$\begin{aligned} \mathcal{Y}_i &\doteq \big[y_i(x_1) \;\; y_i(x_2) \;\; \cdots \;\; y_i(x_{NM})\big] \\ \mathcal{E}_i &\doteq \big[\varepsilon_i(x_1) \;\; \varepsilon_i(x_2) \;\; \cdots \;\; \varepsilon_i(x_{NM})\big]^{\mathsf{T}} \\ \mathcal{Z}_i &\doteq \big[r(x_1, \pi_{i-1}(x_1)) \;\cdots\; r(x_{NM}, \pi_{i-1}(x_{NM}))\big]^{\mathsf{T}} \end{aligned}$$

the expression (H.1) can be compactly rewritten as

$$\mathcal{Y}_i^{\mathsf{T}} \cdot \theta_i + \mathcal{E}_i = \mathcal{Z}_i,$$

and the batch least-squares solution θ_i^* minimizing the approximation error $\mathcal{J}(\theta_i) \doteq \frac{1}{2}\|\mathcal{E}_i\|^2$ over $\{x_k\}$ is given by

$$\theta_i^* = \big(\mathcal{Y}_i \mathcal{Y}_i^{\mathsf{T}}\big)^{-1} \mathcal{Y}_i \, \mathcal{Z}_i$$

so long as $\operatorname{rank}(\mathcal{Y}_i) = L$. We implement each ith policy evaluation by collecting data $\mathcal{Y}_i$ and $\mathcal{Z}_i$ at the distinct points $\{x_k\} \subset \Omega$ and then perform the batch least squares to find the minimizing solution θ_i^*.

H.3 Reward Function and Policy Improvement Update Rule

Recall that each experimental case in §6 basically considers the reward function r given by (31) and (35) with (50), that is,

$$r(x,u) = \mathfrak{r}(x) - \mathfrak{c}(u), \;\; \text{with } \mathfrak{c}(u) = \lim_{v \to u} \int_0^v (s^{\mathsf{T}})^{-1}(\mathfrak{u}) \cdot \Gamma \, d\mathfrak{u} \text{ and } s(\mathfrak{u}) = u_{\mathsf{max}} \tanh(\mathfrak{u}/u_{\mathsf{max}}) \tag{H.2}$$

where $\Gamma > 0$ and the sigmoid function s gives the following expressions of the functions σ in (33) and $\mathfrak{c}$:

$$\begin{aligned} \sigma(\mathfrak{u}) &= u_{\mathsf{max}} \tanh(\Gamma^{-1} \cdot \mathfrak{u}/u_{\mathsf{max}}) = 5 \tanh\big((5\Gamma)^{-1} \cdot \mathfrak{u}\big), \\ \mathfrak{c}(u) &= \Gamma \cdot (u_{\mathsf{max}}^2/2) \cdot \ln\big(u_+^{u_+} \cdot u_-^{u_-}\big) = 12.5 \cdot \Gamma \cdot \ln\big(u_+^{u_+} \cdot u_-^{u_-}\big) \end{aligned}$$

for $u_\pm \doteq 1 \pm u/u_{\mathsf{max}}$. Here, note that $\mathfrak{c}(u)$ is finite for all $u \in \mathcal{U}$ and has its maximum at the end points $u = \pm u_{\mathsf{max}}$ as $\mathfrak{c}(\pm u_{\mathsf{max}}) = \Gamma \cdot (u_{\mathsf{max}}^2 \ln 4)/2 \approx 17.3287 \cdot \Gamma$.

As the inverted pendulum dynamics is input-affine, the above reward setting (H.2) corresponds to the concave Hamiltonian formulation in §5.1.1. Hence, the policy improvement becomes the following simple update rule:

$$\pi_i(x) \approx \pi(x; \theta_i^*) = \sigma\big(\Delta t \cdot F_{\mathsf{c}}^{\mathsf{T}}(x) \cdot \nabla V^{\mathsf{T}}(x; \theta_i^*)\big) = -5 \tanh\left(\frac{1}{5\Gamma} \cdot \cos x_1 \cdot \nabla_{x_2}\phi(x) \cdot \theta_i^*\right) \tag{H.3}$$

where $\nabla_{x_2}\phi(x) \in \mathbb{R}^{1 \times L}$ denotes the gradient of $\phi(x_1, x_2)$ with respect to the second component x_2. Cases 1 and 2 in §6 are associated with the above update rule (H.3).

In the limit $\Gamma \to 0^+$, it is obvious that $\sigma(\mathfrak{u}) \to u_{\mathsf{max}} \cdot \operatorname{sign}(\mathfrak{u})$ and $\mathfrak{c}(u) \to 0$. In this case, thereby, the reward function (H.2) and the policy improvement update rule (H.3) become $r(x,u) = \mathfrak{r}(x)$ and

$$\pi_i(x) \approx \pi(x;\theta_i^*) = -u_{\mathsf{max}} \cdot \operatorname{sign}\big(\cos x_1 \cdot \nabla_{x_2}\phi(x)\cdot\theta_i^*\big).$$

Cases 3 and 4 in §6 consider this type of bang-bang policies, with continuous (Case 3) and binary state-reward $\mathfrak{r}$ (Case 4).

I Proofs

In this appendix, we provide all the proofs of the Theorems, Lemmas, Propositions, and some Corollaries stated in the main work (Lee and Sutton, 2020b). For the proof of properties of locally uniform convergence, the following lemma is necessary.

Lemma I.1 (Remmert, 1991) *A sequence of functions $g_i : \mathcal{X} \to \mathbb{R}^n$ converges to g locally uniformly iff $g_i \to g$ uniformly on any compact subsets of $\mathcal{X}$.*

I.1 Proofs in §2 Preliminaries

Proof of Lemma 2.1. For any policy π and any $x \in \mathcal{X}$,

$$v_\pi(x) \le \lim_{\eta\to\infty}\left(r_{\mathsf{max}} \cdot \int_0^\eta \gamma^t\, dt\right) = \begin{cases} r_{\mathsf{max}}/\alpha & \text{for } \gamma \in (0,1), \\ 0 & \text{for } \gamma = 1 \end{cases}$$

(note that $r_{\mathsf{max}} = 0$ when $\gamma = 1$). This proves the statement with $\overline{v} = r_{\mathsf{max}}/\alpha$ for $0 < \gamma < 1$ and $\overline{v} = 0$ for $\gamma = 1$. □

Proof of Proposition 2.2. If the reward R_t under a policy π satisfies (4) with $\kappa < \alpha$, then we obtain from the definitions

$$v_\pi(x) = \int_0^\infty e^{-\alpha t}\cdot \mathbb{G}_\pi^x[R_t]\, dt \ge \rho(x)\cdot\int_0^\infty e^{-(\alpha-\kappa)t}\, dt = (\alpha-\kappa)^{-1}\cdot\rho(x) > -\infty \qquad \forall x \in \mathcal{X},$$

which also shows that v_π is lower-bounded if so is ρ. Then, the proof is completed by Lemma 2.1. □

Proof of Lemma 2.3. By the standard calculus and $\alpha \doteq -\ln\gamma$,

$$\frac{d}{dt}\big(\gamma^t\cdot v(X_t)\big) = \gamma^t\cdot\big(\dot v(X_t,U_t) - \alpha\cdot v(X_t)\big).$$

Hence, applying (7) and noting that $h(x,u,\nabla v(x)) = r(x,u) + \dot v(x,u)$, we obtain that for any $t \ge 0$ and $x \in \mathcal{X}$,

$$0 \sim \mathbb{G}_\pi^x\Big[\gamma^t\cdot\big(h(X_t,U_t,\nabla v(X_t)) - \alpha\cdot v(X_t)\big)\Big] = \mathbb{G}_\pi^x\Big[\gamma^t\cdot\big(R_t + \dot v(X_t,U_t) - \alpha\cdot v(X_t)\big)\Big] = \mathbb{G}_\pi^x\left[\gamma^t\cdot R_t + \frac{d}{dt}\big(\gamma^t\cdot v(X_t)\big)\right]$$

where $\sim$ is equal to $=$, $\le$, or $\ge$. Then, integrating it from $t = 0$ to $t = \eta$ yields (6).

For the proof of the opposite direction, assume that v satisfies (6). Then, rearranging (6) as

$$\big(1-\gamma^\eta\big)\cdot v(x) \sim \mathbb{G}_\pi^x\Big[\mathfrak{R}_\eta + \gamma^\eta\cdot\big(v(X_\eta) - v(X_0)\big)\Big] \qquad \forall x \in \mathcal{X} \quad \forall \eta > 0,$$

dividing it by η, and limiting $\eta \to 0$ yields

$$-\ln\gamma\cdot v(x) \sim r_\pi(x) + \dot v(x,\pi(x)) \qquad \forall x \in \mathcal{X},$$

which implies (7) since $\alpha = -\ln\gamma$ and $h(x,\pi(x),\nabla v(x)) = r_\pi(x) + \dot v(x,\pi(x))$. □

Proof of Proposition 2.4. Fix $x \in \mathcal{X}$ and take the limit $\eta \to \infty$ of (8) to have

$$v_\pi(x) = \lim_{\eta\to\infty}\mathbb{G}_\pi^x\big[\,\mathfrak{R}_\eta + \gamma^\eta\cdot v_\pi(X_\eta)\,\big] = v_\pi(x) + \lim_{\eta\to\infty}\mathbb{G}_\pi^x\big[\gamma^\eta\cdot v_\pi(X_\eta)\big].$$

Hence, noting that $v_\pi(x)$ is finite by $\pi \in \Pi_{\mathsf{a}}$, we obtain the boundary condition $\lim_{\eta\to\infty}\mathbb{G}_\pi^x\big[\gamma^\eta\cdot v_\pi(X_\eta)\big] = 0$. Since $x \in \mathcal{X}$ is arbitrary, the proof is completed. □

Proof of Theorem 2.5. Suppose v satisfies the integral BE (10) without loss of generality (or, convert the differential BE (11) to (10) via Lemma 2.3 and fix $\eta > 0$). Then, the repetitive applications of (10) to itself k-times result in

$$v(x) = \mathbb{G}_\pi^x\big[\mathfrak{R}_\eta + \gamma^\eta \cdot v(X_\eta)\big] = \mathbb{G}_\pi^x\big[\mathfrak{R}_{2\eta} + \gamma^{2\eta} \cdot v(X_{2\eta})\big] = \ \cdots \ = \mathbb{G}_\pi^x\big[\mathfrak{R}_{k\cdot\eta} + \gamma^{k\cdot\eta} \cdot v(X_{k\cdot\eta})\big] \qquad \forall x \in \mathcal{X}.$$

Taking the limit $k \to \infty$ and substituting (12), we obtain

$$v(x) = \underbrace{\lim_{k\to\infty} \mathbb{G}_\pi^x\big[\mathfrak{R}_{k\cdot\eta}\big]}_{=v_\pi(x)} + \underbrace{\lim_{k\to\infty} \mathbb{G}_\pi^x\big[\gamma^{k\cdot\eta} \cdot v(X_{k\cdot\eta})\big]}_{=0} = v_\pi(x) \qquad \forall x \in \mathcal{X}.$$

Therefore, $v = v_\pi$ and since $v(x)$ is finite for each $x \in \mathcal{X}$, $\pi \in \Pi_{\mathsf{a}}$. The converse is obvious by Proposition 2.4. □

Proof of Lemma 2.6. By Lemma 2.3, the inequality (13) is equivalent to

$$v(x) \le \mathbb{G}_{\pi'}^x\big[\mathfrak{R}_\eta + \gamma^\eta \cdot v(X_\eta)\big] \qquad \forall x \in \mathcal{X} \quad \forall \eta > 0.$$

Then, taking the limit supremum at $\eta \to \infty$, we obtain for each $x \in \mathcal{X}$:

$$v(x) \le v_{\pi'}(x) + \limsup_{\eta\to\infty} \ \mathbb{G}_{\pi'}^x\big[\gamma^\eta \cdot v(X_\eta)\big] \le v_{\pi'}(x) \qquad \forall x \in \mathcal{X}$$

where we have substituted

$$\limsup_{\eta\to\infty} \ \mathbb{G}_{\pi'}^x[\gamma^\eta \cdot v(X_\eta)] \le \sup_{x\in\mathcal{X}} v(x) \cdot \lim_{\eta\to\infty} \gamma^\eta \le 0$$

which is true since v is upper-bounded (by zero if $\gamma = 1$) and $\gamma \in (0, 1]$. Since $v(x)$ is finite for all $x \in \mathcal{X}$, we have

$$-\infty < v(x) \le v_{\pi'}(x) \le \overline{v} < \infty \qquad \forall x \in \mathcal{X}$$

by Lemma 2.1. Therefore, π' is admissible and $v \leqslant v_{\pi'}$. □

Proof of Theorem 2.7. The policy π' given by (15) satisfies:

$$h(x, \pi'(x), \nabla v_\pi(x)) \ge h(x, \pi(x), \nabla v_\pi(x)) = \alpha \cdot v_\pi(x) \qquad \forall x \in \mathcal{X}$$

where we substituted the differential BE (9). Therefore, the applications of Lemmas 2.1 and 2.6 directly prove the theorem. □

Proof of Theorem 2.8. By optimality and Lemma 2.1, $v \leqslant v_* \leqslant \overline{v}$ for any $v \in \mathcal{V}_{\mathsf{a}}$, implying $v_* \in \mathcal{V}_{\mathsf{a}}$. Moreover, the maximal policy π'_* over π_* is also optimal since $v_{\pi_*} (= v_*) \leqslant v_{\pi'_*}$ by Theorem 2.7 and $v_{\pi'_*} \leqslant v_*$ by optimality. Therefore, the differential BE (9) w.r.t. the policy $\pi = \pi'_*$ and the policy improvement (15) for $\pi' = \pi'_*$ and $\pi = \pi_*$, all with $v_{\pi_*} = v_* = v_{\pi'_*}$, result in the HJBE (17). Finally, comparing the HJBE (17) with the differential BE (9) for $\pi = \pi_*$ and $v_\pi = v_*$, we have (18). □

I.2 Proofs in §4 Fundamental Properties of PIs

Proof of Theorem 4.1. π_0 is admissible by initialization. Suppose for some $i \in \mathbb{N}$ that π_{i-1} is admissible. Then, $v_i = v_{\pi_{i-1}}$ holds by Theorem 2.5 and the boundary condition (29); π_i is also admissible and $\pi_{i-1} \preccurlyeq \pi_i$ by Theorem 2.7. Therefore, the mathematical induction completes the proof. □

Proof of Theorem 4.2. By Theorem 4.1 and Lemma 2.1, we have

$$v_1(x) \le \cdots \le v_i(x) \le v_{i+1}(x) \le \cdots \le \overline{v} < \infty \text{ for each fixed } x \in \mathcal{X}.$$

That is, the sequence $\langle v_i(x) \rangle$ in $\mathbb{R}$ is monotonically increasing and upper bounded by a constant $\overline{v} \in \mathbb{R}$. Hence, $v_i(x)$ converges to $\hat{v}_*(x) \doteq \sup_{i\in\mathbb{N}} v_i(x)$ by monotone convergence theorem (Thomson et al., 2001, Theorem 2.28), implying the pointwise convergence $v_i \to \hat{v}_*$.

Next, since every admissible VF is assumed C^1 (see (3)) and $v_i = v_{\pi_{i-1}}$ is admissible by Theorem 4.1, v_i is continuous for each $i \in \mathbb{N}$. Hence, $\hat{v}_*$ is lower semicontinuous (Folland, 1999, Proposition 7.11c) and the monotone sequence $\langle v_i \rangle$ converges to $\hat{v}_*$ *uniformly on* Ω if Ω is compact and $\hat{v}_*$ is *continuous over* Ω by Dini's theorem (Rudin, 1964, Theorem 7.13). Finally, $v_i \to \hat{v}_*$ uniformly on any compact $\Omega \subset \mathcal{X}$ if $\hat{v}_*$ is continuous, hence the last statement is obvious by Lemma I.1. □

Proof of Proposition 4.3. Since v^* is a fixed point of $\mathcal{T}$, we have $\mathcal{T}v^* = v^* \in \mathcal{V}_{\mathsf{a}}$. Let π^* be a policy s.t. $v_{\pi^*} = \mathcal{T}v^*$ and

$$\pi^*(x) \in \arg\max_{u \in \mathcal{U}} h(x, u, \nabla v^*(x)) \qquad \forall x \in \mathcal{X}. \tag{I.1}$$

Then, we have $v_{\pi^*} = v^* \in \mathcal{V}_{\mathsf{a}}$, meaning that π^* is admissible. Since any admissible policy π satisfies the differential BE (9), it is true for $\pi = \pi^*$, that is, $\alpha \cdot v_{\pi^*}(x) = h(x, \pi^*(x), \nabla v_{\pi^*}(x))$ for all $x \in \mathcal{X}$, from which and $v_{\pi^*} = v^*$ we finally obtain

$$\alpha \cdot v^*(x) = h(x, \pi^*(x), \nabla v^*(x)) \qquad \forall x \in \mathcal{X}.$$

Therefore, the substitution of (I.1) concludes that a fixed point v^* of $\mathcal{T}$ is a solution v_* to the HJBE (17). □

Proof of Theorem 4.5. By Lemma I.2 below and Assumption 4.4, v^* is a unique fixed point of $\mathcal{T}^N$ for all $N \in \mathbb{N}$. Hence, Bessaga (1959)'s converse of the Banach fixed point theorem ensures that there exists a metric d on $\mathcal{V}_{\mathsf{a}}$ such that $(\mathcal{V}_{\mathsf{a}}, d)$ is a complete metric space and $\mathcal{T}$ is a contraction under d. Then, as v^* is the unique fixed point of $\mathcal{T}$, the Banach fixed point theorem (e.g., Kirk and Sims, 2013, Theorem 2.2) shows

$$\forall v_1 \in \mathcal{V}_{\mathsf{a}} : \lim_{N\to\infty} v_{N+1} = \lim_{N\to\infty} \mathcal{T}^N v_1 = v^* \text{ in the metric } d,$$

implying the convergence $v_i \to v^*$ in the metric d. □

Lemma I.2 *If v^* is a unique fixed point of $\mathcal{T}$, then it is a unique fixed point of $\mathcal{T}^N$ for any $N \in \mathbb{N}$.*

Proof. Suppose v^* is the unique fixed point of $\mathcal{T}$. Then, it is also a fixed point of $\mathcal{T}^N$ for any $N \in \mathbb{N}$ since

$$\mathcal{T}^N v^* = \mathcal{T}^{N-1}[\mathcal{T}v^*] = \mathcal{T}^{N-1}v^* = \cdots = \mathcal{T}v^* = v^*.$$

To show that v^* is the *unique* fixed point of $\mathcal{T}^N$ *for all* $N \in \mathbb{N}$ by contradiction, suppose that there exist $M \in \mathbb{N}$ and $v \in \mathcal{V}_{\mathsf{a}}$ s.t. $\mathcal{T}^M v = v \neq v^*$. Then, the repetitive applications of Theorem 2.7 result in

$$v \leqslant \mathcal{T}v \leqslant \mathcal{T}^2 v \leqslant \cdots \leqslant \mathcal{T}^M v = v$$

and thus $\mathcal{T}v = v$. Since v^* is the *unique* fixed point of $\mathcal{T}$, we have a contradiction, $v = v^*$. Therefore, v^* is the unique fixed point of $\mathcal{T}^N$ for all $N \in \mathbb{N}$, and the proof is completed. □

Proof of Theorem 4.6. $\hat{v}_* \in \mathcal{V}_{\mathsf{a}}$ and (3) imply $\hat{v}_* \in \mathrm{C}^1$ and thus continuity of $\hat{v}_*$. Hence, v_i locally uniformly converges to $\hat{v}_*$ by Theorem 4.2c. This and Lemma I.1 imply that for each compact subset Ω of $\mathcal{X}$, $v_i \to \hat{v}_*$ in the uniform pseudometric d_Ω. By this and continuity of $\mathcal{T}$ under d_Ω, we have

$$\hat{v}_* = \lim_{i\to\infty} v_{i+1} = \lim_{i\to\infty} \mathcal{T}v_i = \mathcal{T}\Big(\lim_{i\to\infty} v_i\Big) = \mathcal{T}\hat{v}_* \text{ in the metric } d_\Omega,$$

implying $d_\Omega(\hat{v}_*, \mathcal{T}\hat{v}_*) = 0$ for every compact subset $\Omega \subset \mathcal{X}$, hence $\hat{v}_* = \mathcal{T}\hat{v}_*$. Therefore, we finally have $\hat{v}_* = v^*$ by Assumption 4.4, which completes the proof. □

Proof of Theorem 4.9. ∇v_i converges locally uniformly by Assumption 4.8a. Hence, for each $x \in \mathcal{X}$, there is a neighborhood $\mathcal{N}_x$ of x on which ∇v_i converges uniformly. Since a neighborhood $\mathcal{N}_x$ of x contains an open ball $\mathcal{B}_x \doteq \{y \in \mathcal{X} : \|x - y\| < r\}$ centered at x, for some $r > 0$, and every open ball in $\mathcal{X}$ is convex, Lemma I.3 below ensures that for every $x \in \mathcal{X}$, $\hat{v}_*$ is C^1 over $\mathcal{B}_x$ and $\nabla v_i \to \nabla \hat{v}_*$ uniformly on $\mathcal{B}_x$. This and $\mathcal{X} = \bigcup_{x\in\mathcal{X}} \mathcal{B}_x$ establish that $\hat{v}_*$ is C^1 and

$$\nabla v_i \to \nabla \hat{v}_* \text{ locally uniformly.} \tag{I.2}$$

Since $\hat{v}_*$ is continuous ($\because$ it is C^1), Theorem 4.2c implies that $v_i \to \hat{v}_*$ locally uniformly. Let $\hat{\pi}_* : \mathcal{X} \to \mathcal{U}$ be the function to which $\langle \pi_i \rangle$ converges pointwise. Such a function $\hat{\pi}_*$ exists by Assumption 4.8b. Then, since each of the ith policy π_i satisfies

$$\pi_i(x) \in \arg\max_{u \in \mathcal{U}} h(x, u, \nabla v_i(x)) \qquad \forall x \in \mathcal{X},$$

Assumption 4.7 and (I.2) imply that the limit function $\hat{\pi}_*$ holds

$$\hat{\pi}_*(x) \in \arg\max_{u \in \mathcal{U}} h(x, u, \nabla \hat{v}_*(x)) \qquad \forall x \in \mathcal{X}. \tag{I.3}$$

Note that for each $i \in \mathbb{N}$, $v_i = v_{\pi_{i-1}} \in \mathcal{V}_{\mathsf{a}}$ by Theorem 4.1, hence π_{i-1} satisfies the differential BE (9) for $\pi = \pi_{i-1}$. That is,

$$\alpha \cdot v_i(x) = h(x, \pi_{i-1}(x), \nabla v_i(x)) \qquad \forall x \in \mathcal{X} \quad \forall i \in \mathbb{N}.$$

Then, taking the pointwise limit $i \to \infty$ on both sides and using continuity of h and (I.3) results in

$$\alpha \cdot \hat{v}_*(x) = h(x, \hat{\pi}_*(x), \nabla \hat{v}_*(x)) = \max_{u \in \mathcal{U}} h(x, u, \nabla \hat{v}_*(x)) \qquad \forall x \in \mathcal{X}. \tag{I.4}$$

Here, (I.4) and (I.3) are exactly the HJBE (17) and (18), respectively, for $v_* = \hat{v}_*$ and $\pi_* = \hat{\pi}_*$, completing the proof. □

Lemma I.3 *If ∇v_i uniformly converges on an open convex subset $\mathcal{S} \subset \mathcal{X}$, then*

$$\hat{v}_* \textit{ is } \mathrm{C}^1 \textit{ over } \mathcal{S} \textit{ and } \nabla v_i \to \nabla \hat{v}_* \textit{ uniformly on } \mathcal{S}.$$

Proof. Let $x \in \mathcal{S}$ and e_j be the unit vector in $\mathcal{X}$ $(= \mathbb{R}^l)$ whose jth element is 1 and 0 otherwise. Since $\mathcal{S}$ is open, there exists $\theta > 0$ s.t. for each j, both

$$x_j^+ \doteq x + \frac{\theta}{2} \cdot e_j \text{ and } x_j^- \doteq x - \frac{\theta}{2} \cdot e_j$$

belong to $\mathcal{S}$. Define a function $g_j : [0,1] \to \mathcal{S}$ as

$$g_j(\beta) \doteq \beta x_j^+ + (1-\beta) x_j^- \text{ for } \beta \in [0,1],$$

where the dependencies on x and θ are implicit; by convexity of $\mathcal{S}$, $g_j(\beta) \in \mathcal{S}$ for all $\beta \in [0,1]$. Then, the composition $v_i \circ g_j$ pointwise converges to $\hat{v}_* \circ g_j$ by Theorem 4.2a. Moreover, the derivative $(v_i \circ g_j)'$ (w.r.t. β) can be expressed by chain rule as

$$(v_i \circ g_j)'(\beta) = \theta \cdot \nabla v_i(g_j(\beta))\, e_j = \theta \cdot \left. \frac{\partial v_i(z)}{\partial z_j} \right|_{z = g_j(\beta)}$$

which reveals that $(v_i \circ g_j)'$ is continuous and converges uniformly on $[0,1]$ since so is ∇v_i on $\mathcal{S}$ (note that $v_i = v_{\pi_{i-1}} \in \mathrm{C}^1$ by Theorem 4.1 and the regularity Assumption (3)). Hence, the application of (Thomson et al., 2001, Theorem 9.34) shows that $\hat{v}_* \circ g_j$ is differentiable (w.r.t. β) and

$$(v_i \circ g_j)' \to (\hat{v}_* \circ g_j)' \text{ uniformly on } [0,1]. \tag{I.5}$$

By definition, the derivative $(\hat{v}_* \circ g_j)'(\beta)$ at $\beta = 1/2$ satisfies

$$(\hat{v}_* \circ g_j)'(1/2) = \lim_{\epsilon \to 0} \frac{\hat{v}_*(g_j(1/2+\epsilon)) - \hat{v}_*(g_j(1/2))}{\epsilon} = \lim_{\epsilon \to 0} \frac{\hat{v}_*(x + \epsilon\theta \cdot e_j) - \hat{v}_*(x)}{\epsilon} = \theta \cdot \frac{\partial \hat{v}_*(x)}{\partial x_j}.$$

Since this is true for any $j = \{1, 2, \cdots, l\}$ and any $x \in \mathcal{S}$, the gradient $\nabla \hat{v}_*$ exists over $\mathcal{S}$. Moreover, (I.5) at $\beta = 1/2$ implies

$$\frac{\partial v_i(x)}{\partial x_j} \to \frac{\partial \hat{v}_*(x)}{\partial x_j} \qquad \forall x \in \mathcal{S} \quad \forall j \in \{1, 2, \cdots, l\},$$

hence ∇v_i uniformly converges to $\nabla \hat{v}_*$ on $\mathcal{S}$. This also implies that the convergent point $\nabla \hat{v}_*$ is continuous over $\mathcal{S}$ (Rudin, 1964, Theorem 7.12). That is, $\hat{v}_*$ is C^1 over $\mathcal{S}$. □

I.3 Proofs in §5 Case Studies

Here, we prove the mathematical statements, including Theorems, Lemmas, Propositions, and some Corollaries, w.r.t. each case study presented in the main paper (Lee and Sutton, 2020b, §5). For some proofs, the following lemmas are required.

Lemma I.4 *For any two action spaces $\mathcal{U}$ and $\mathcal{A}$, the inverse of a continuous bijection $g : \mathcal{U}^{\mathrm{o}} \to \mathcal{A}^{\mathrm{o}}$ is continuous.*

Proof. By definitions, $\mathcal{U}^{\mathrm{o}}$ and $\mathcal{A}^{\mathrm{o}}$ are open sets in $\mathbb{R}^m$. Hence, Brouwer (1911)'s invariance of domain theorem implies that $g(\mathcal{O})$ for every open subset $\mathcal{O} \subseteq \mathcal{U}^{\mathrm{o}}$ is open. Hence, the inverse g^{-1} is continuous (Rudin, 1964, Theorem 4.8). □

Lemma I.5 *Let $\psi : \mathcal{X}^2 \to \mathcal{U}$ be a continuous function. If a sequence $\langle g_i \rangle$ of continuous functions $g_i : \mathcal{X} \to \mathcal{X}$ converges to g locally uniformly, then so does $x \mapsto \psi(x, g_i(x))$ to $x \mapsto \psi(x, g(x))$.*

Proof. Let $\Omega \subset \mathcal{X}$ be compact. Then, by locally uniform convergence $g_i \to g$ and Lemma I.1, we have the followings:

(1) $\langle g_i \rangle$ is uniformly equicontinuous over any compact subset $S \subset \mathcal{X}$ (Rudin, 1964, Theorem 7.24), that is,

$$\text{given } \delta > 0, \text{ there exists } \delta' > 0 \text{ s.t. } \big(\|x - x'\| < \delta' \implies \|g_i(x) - g_i(x')\| < \delta, \quad \forall x, x' \in S \quad \forall i \in \mathbb{N}\big); \tag{I.6}$$

(2) $\langle g_i \rangle$ is uniformly bounded on Ω (e.g., Rudin, 1964, Theorem 7.25);

(3) g is continuous over Ω (Rudin, 1964, Theorem 7.12), hence the image $g(\Omega)$ is compact (Rudin, 1964, Theorem 4.14).

In short, we have uniform equicontinuity (I.6) over any compact subset $S \subset \mathcal{X}$ and a uniform bound $M > 0$ over Ω, that is,

$$\|g_i(x)\| \le M \text{ and } \|g(x)\| \le M \qquad \forall x \in \Omega \quad \forall i \in \mathbb{N}. \tag{I.7}$$

Next, let $\varepsilon > 0$ and $S_0 \subset \mathcal{X}$ be a compact subset defined as $S_0 \doteq \{y \in \mathcal{X} : \|y\| \le M\}$. Then, the function ψ is uniformly continuous over the compact subset $\Omega \times S_0 \subset \mathcal{X}^2$ (Rudin, 1964, Theorem 4.14), hence there exists $\delta > 0$ such that

$$\|x - x'\| < \delta \text{ and } \|y - y'\| < \delta \quad \implies \quad \big\|\psi(x, y) - \psi(x', y')\big\| < \varepsilon, \qquad \forall x, x' \in \Omega \quad \forall y, y' \in S_0$$

Since $g_i(x), g(x) \in S_0$ for each $x \in \Omega$ and $i \in \mathbb{N}$ by (I.7), the uniform equicontinuity (I.6) over the compact set $S = S_0$ finally results in: for any $x, x' \in \Omega$ and any $i \in \mathbb{N}$,

$$\|x - x'\| < \delta^* \quad \implies \quad \big\|\psi(x, g_i(x)) - \psi(x, g_i(x))\big\| < \varepsilon$$

where $\delta^* \doteq \min\{\delta, \delta'\}$. That is, the sequence of functions $x \mapsto \psi(x, g_i(x))$ is uniformly equicontinuous over Ω. Moreover, for each $x \in \mathcal{X}$, $y \mapsto \psi(x, y)$ is continuous and $g_i(x)$ converges to $g(x)$, hence $\psi(x, g_i(x))$ converges to $\psi(x, g(x))$. Therefore, $x \mapsto \psi(x, g_i(x))$ converges to $x \mapsto \psi(x, g(x))$ uniformly on Ω (Royden, 1988, Lemma 39 in Chapter 7; or see Rudin, 1964, Exercise 16 in Chapter 7); since the compact set Ω is arbitrary, the proof is completed by Lemma I.1. □

Proof of Properties of $\mathfrak{c}$ (§5.1.1). The target properties are w.r.t. the gradient and the gradient inverse of $\mathfrak{c}$ shown below:

(1) $\nabla\mathfrak{c}^\mathsf{T}$ is *bijective*, so that its inverse $\sigma \doteq (\nabla\mathfrak{c}^\mathsf{T})^{-1}$ exists;

(2) $\nabla\mathfrak{c}^\mathsf{T}$ and σ are *strictly monotone* and *continuous*.

where $\mathfrak{c} : \mathcal{U} \to \mathbb{R}$ is a function given in (31). Here, we prove those properties of $\mathfrak{c}$.

To begin with, recall that $\mathfrak{c}$ is assumed *strictly convex*, C^1, and its gradient $\nabla\mathfrak{c}$ is *surjective*, i.e., $\nabla\mathfrak{c}^\mathsf{T}(\mathcal{U}^\mathrm{o}) = \mathbb{R}^m$. First, we focus on $\nabla\mathfrak{c}^\mathsf{T}$. As $\mathfrak{c}$ is assumed C^1, continuity of $\nabla\mathfrak{c}^\mathsf{T}$ is obvious. To prove strict monotonicity, note that $\mathfrak{c}$ satisfies Lemma I.6 below; by adding the two strict inequalities in Lemma I.6 and rearranging it, we obtain

$$(\nabla\mathfrak{c}(u) - \nabla\mathfrak{c}(u'))(u - u') > 0 \qquad \forall u \neq u'. \tag{I.8}$$

Hence, $\nabla\mathfrak{c}^\mathsf{T}$ is strictly monotone.[12] Moreover, $\nabla\mathfrak{c}^\mathsf{T}$ is injective — if not, $\exists u, u' \in \mathcal{U}^\mathrm{o}$ s.t. $u \neq u'$ but $\nabla\mathfrak{c}(u) = \nabla\mathfrak{c}(u')$, which and strict monotonicity of $\nabla\mathfrak{c}$ directly lead us a contradiction "$0 > 0$":

$$0 = 0^\mathsf{T}(u - u') = (\nabla\mathfrak{c}(u) - \nabla\mathfrak{c}(u'))(u - u') > 0.$$

Therefore, the surjective mapping $\nabla\mathfrak{c}^\mathsf{T}$ is also injective and thus bijective. This ensures the existence of the inverse σ; since $\nabla\mathfrak{c}^\mathsf{T}$ is continuous, so is its inverse σ by Lemma I.4 above with $\mathcal{A} = \mathbb{R}^m$.

To prove strict monotonicity of σ, let $u \doteq \sigma(\mathfrak{u})$ and $u' \doteq \sigma(\mathfrak{u}')$ for arbitrary $\mathfrak{u}, \mathfrak{u}' \in \mathbb{R}^m$. Then, we obviously have $\mathfrak{u} = \nabla\mathfrak{c}^\mathsf{T}(u)$ and $\mathfrak{u}' = \nabla\mathfrak{c}^\mathsf{T}(u')$ and thus by (I.8),

$$(\mathfrak{u} - \mathfrak{u}')^\mathsf{T}(\sigma(\mathfrak{u}) - \sigma(\mathfrak{u}')) > 0 \qquad \forall \mathfrak{u} \neq \mathfrak{u}',$$

which completes the proof. □

Lemma I.6 *For a strictly convex* C^1 *function* $\mathfrak{c} : \mathcal{U} \to \mathbb{R}$ *and for any* u, $u' \in \mathcal{U}^\mathrm{o}$ *such that* $u \neq u'$,

$$\begin{cases} \mathfrak{c}(u) > \mathfrak{c}(u') + \nabla\mathfrak{c}(u')(u - u') \\ \mathfrak{c}(u') > \mathfrak{c}(u) + \nabla\mathfrak{c}(u)(u' - u), \end{cases}$$

[12] The converse (i.e., $\nabla\mathfrak{c}^\mathsf{T}$ is strictly monotone $\implies$ $\mathfrak{c}$ is strictly convex) is also true. This equivalence between convexity of $\mathfrak{c}$ and monotonicity of $\nabla\mathfrak{c}^\mathsf{T}$ is known as Kachurovskii (1960)'s theorem.

where the second inequality is due to the interchange of u and u' of the first one.

Proof. Let $g(\beta) \doteq \mathfrak{c}\big(\beta \cdot u + (1-\beta)\cdot u'\big)$ for $\beta \in [0,1]$. Then, g is strictly convex and C^1 since so is $\mathfrak{c}$. By the mean value theorem, there exists $\bar{\beta} \in (0,1)$ such that

$$g(1) - g(0) = g'(\bar{\beta}) > \lim_{\beta \to 0^+} g'(\beta) = \nabla \mathfrak{c}(u')(u - u'),$$

where the strict inequality comes from the fact that the derivative g' of a strictly convex C^1 function g is strictly increasing. [13] Then, the proof is completed by substituting the definition of g into the strict inequality. □

Proof of Theorem 5.1 (§5.1.1). Combine Lemma I.7 below with Theorem 4.9. □

Lemma I.7 *Under* (30) *and* (31)*,*

a. *Assumption 4.7 is true;*

b. *if $\langle \nabla v_i \rangle$ locally uniformly converges to a function ξ, then $\langle \pi_i \rangle$ locally uniformly converges to $\sigma(F_{\mathsf{c}}^{\mathsf{T}} \cdot \xi^{\mathsf{T}})$.*

Proof. a. Under (30) and (31), the maximal function u_* in (14) is uniquely determined by (32) and thus continuous. This also implies that the argmax-set in (14) is a singleton, hence Assumption 4.7 is equivalent to the continuity of $p \mapsto u_*(x,p)$ (see Remark 4.10) which is obviously true by continuity of u_*.

b. For each $i \in \mathbb{N}$, $v_i \in \mathrm{C}^1$ (i.e., ∇v_i is continuous) since $v_i \in \mathcal{V}_{\mathsf{a}}$ by Theorem 4.1 and $\mathcal{V}_{\mathsf{a}} \subset \mathrm{C}^1$ by (3). The function $(x,y) \mapsto \sigma(F_{\mathsf{c}}^{\mathsf{T}}(x)y)$ is also continuous since so are F_{c} and σ. Therefore, applying Lemma I.5 with $\psi(x,y) = \sigma(F_{\mathsf{c}}^{\mathsf{T}}(x)y)$, $g_i = \nabla v_i$, and $g = \xi$, we complete the proof. □

Proof of Theorem 5.4 (§5.1.2). Denoting $a \doteq \varphi(u)$ and considering $a \in \mathcal{A}^{\mathrm{o}}$ as the action transformed from $u \in \mathcal{U}^{\mathrm{o}}$, we can formulate the following input-affine dynamics $\bar{f}$ and the reward function $\bar{r}$ from (39) and (40) as

$$\begin{cases} \bar{f}(x,a) \doteq f(x, \varphi^{-1}(a)) = f_{\mathsf{d}}(x) + F_{\mathsf{c}}(x) \cdot a \\ \bar{r}(x,a) \doteq r(x, \varphi^{-1}(a)) = \mathfrak{r}(x) - \mathfrak{c}(a), \end{cases}$$

both of which are defined for all $(x,a) \in \mathcal{X} \times \mathcal{A}^{\mathrm{o}}$. The associated Hamiltonian $\bar{h} : \mathcal{X} \times \mathcal{A}^{\mathrm{o}} \times \mathcal{X}^{\mathsf{T}} \to \mathbb{R}$ is given by

$$\bar{h}(x,a,p) = \underbrace{\mathfrak{r}(x) - \mathfrak{c}(a)}_{\bar{r}(x,a)} + p \cdot \underbrace{\big(f_{\mathsf{d}}(x) + F_{\mathsf{c}}(x) \cdot a\big)}_{\bar{f}(x,a)} = h(x, \varphi^{-1}(a), p). \tag{I.9}$$

Here, both $a \mapsto \bar{r}(x,a)$ and $a \mapsto \bar{h}(x,a,p)$ are strictly concave and C^1 for each $x \in \mathcal{X}$. Thus, similarly to the maximal function u_* in §5.1, a maximal function $a_* : \mathcal{X} \times \mathcal{X}^{\mathsf{T}} \to \mathcal{A}^{\mathrm{o}}$ such that

$$a_*(x,p) \in \operatorname*{arg\,max}_{a \in \mathcal{A}^{\mathrm{o}}} \bar{h}(x,a,p) \qquad \forall (x,p) \in \mathcal{X} \times \mathcal{X}^{\mathsf{T}}$$

exists and is continuous as it can be uniquely represented as (see also §D)

$$a_*(x,p) = \sigma\big(F_{\mathsf{c}}^{\mathsf{T}}(x)\, p^{\mathsf{T}}\big) \qquad \forall (x,p) \in \mathcal{X} \times \mathcal{X}^{\mathsf{T}}. \tag{I.10}$$

Claim I.8 $\varphi^{-1}\big[a_*(x,p)\big] \in \arg\max_{u \in \mathcal{U}} h(x,u,p) \quad \forall (x,p) \in \mathcal{X} \times \mathcal{X}^{\mathsf{T}}$.

By (I.10) and Claim I.8, a maximal function u_* satisfying (14) for the RL problem (39) and (40) is given by

$$u_*(x,p) = \tilde{\sigma}\big(F_{\mathsf{c}}^{\mathsf{T}}(x)\, p^{\mathsf{T}}\big) \qquad \forall (x,p) \in \mathcal{X} \times \mathcal{X}^{\mathsf{T}}, \tag{I.11}$$

[13] Consider the inequalities for $0 \le x_1 < x_1' < x_2 < x_2' \le 1$:

$$\frac{g(x_1') - g(x_1)}{x_1' - x_1} < \frac{g(x_2) - g(x_1')}{x_2 - x_1'} < \frac{g(x_2') - g(x_2)}{x_2' - x_2}$$

(e.g., see Sundaram, 1996, Theorem 7.5) and take the limits $x_1' \to x_1$ and $x_2' \to x_2$, resulting in $g'(x_1) < g'(x_2)$ for $x_1 < x_2$.

where $\tilde{\sigma}(\mathfrak{u}) \doteq \varphi^{-1}[\sigma(\mathfrak{u})]$. Moreover, u_* is continuous since so are both a_* and the inverse φ^{-1} by (I.10) and Lemma I.4, respectively (see Claim I.8 above). Therefore, substituting (I.11) into (16) and (19) result in the following respective closed-form expressions of a maximal policy π' over $\pi \in \Pi_{\mathsf{a}}$ and a HJB policy π_*:

$$\pi'(x) = \tilde{\sigma}(F_{\mathsf{c}}^{\mathsf{T}}(x)\nabla v_\pi^{\mathsf{T}}(x)) \text{ and } \pi_*(x) = \tilde{\sigma}(F_{\mathsf{c}}^{\mathsf{T}}(x)\nabla v_*^{\mathsf{T}}(x)).$$

Next, substituting (I.9) and $\pi_*(x) = \varphi^{-1}[a_*(x, \nabla v_*(x))]$ into the HJBE (17), we obtain

$$\alpha \cdot v_*(x) = h(x, \pi_*(x), \nabla v_*(x)) = \bar{h}\big(x, a_*(x, \nabla v_*(x)), \nabla v_*(x)\big) = \max_{a \in \mathcal{A}^{\mathrm{o}}} \bar{h}(x, a, \nabla v_*(x)) \qquad \forall x \in \mathcal{X}.$$

In addition, under (39), (40), and (41), the PIs running on the original and transformed RL problems result in the same sequences of VFs since both PIs generate the sequences of the policies $\langle \pi_i \rangle$ and $\langle \varphi(\pi_i) \rangle$, respectively, and

$$f(x, u) = \bar{f}(x, a) \text{ and } r(x, u) = \bar{r}(x, a) \text{ under } a = \varphi(u).$$

Therefore, the application of Theorem 5.1 to the transformed RL problem shows that for both cases, the limit function $\hat{v}_*$ is a solution $v_* \in \mathrm{C}^1$ to the HJBE (17) s.t. $v_i \to v_*$ and $\nabla v_i \to \nabla v_*$ both locally uniformly. For locally uniform convergence of $\langle \pi_i \rangle$ towards π_*, apply Lemma I.5 with $\psi(x, y) = \tilde{\sigma}(F_{\mathsf{c}}^{\mathsf{T}}(x)y)$, $g_i = \nabla v_i$, and $g = \nabla v_*$.

(Proof of Claim I.8). Fix $(x, p) \in \mathcal{X} \times \mathcal{X}^{\mathsf{T}}$ and note that the associated Hamiltonian $\bar{h}$ satisfies (see (I.9))

$$\bar{h}(x, a, p) = h(x, \varphi^{-1}(a), p) \quad \forall a \in \mathcal{A}^{\mathrm{o}}. \tag{I.12}$$

Since φ is a bijection between the interior spaces $\mathcal{U}^{\mathrm{o}}$ and $\mathcal{A}^{\mathrm{o}}$, we have $\varphi(\mathcal{U}^{\mathrm{o}}) = \mathcal{A}^{\mathrm{o}}$, which and (I.12) imply

$$\max_{a \in \mathcal{A}^{\mathrm{o}}} \bar{h}(x, a, p) = \max_{u \in \mathcal{U}^{\mathrm{o}}} \bar{h}(x, \varphi(u), p) = \max_{u \in \mathcal{U}^{\mathrm{o}}} h(x, u, p). \tag{I.13}$$

For simplicity, denote $a_*(x, p)$ by a_* and $u_* \doteq \varphi^{-1}(a_*)$. Here, a_* and u_* belong to the interiors $\mathcal{A}^{\mathrm{o}}$ and $\mathcal{U}^{\mathrm{o}}$, respectively. So,

$$\max_{a \in \mathcal{A}^{\mathrm{o}}} \bar{h}(x, a, p) = \bar{h}(x, a_*, p) = h(x, \varphi^{-1}(a_*), p) = h(x, u_*, p) \qquad \text{by (I.12).}$$

This and (I.13) imply that u_* satisfies

$$u_*(x, p) \in \arg\max_{u \in \mathcal{U}^{\mathrm{o}}} h(x, u, p).$$

This proves the statement if $\partial\mathcal{U} = \varnothing$. If not, suppose that there exists a $\tilde{u} \in \partial\mathcal{U}$ on the boundary $\partial\mathcal{U}$ s.t.

$$h(x, \tilde{u}, p) > h(x, u, p) \text{ for all } u \in \mathcal{U}^{\mathrm{o}}. \tag{I.14}$$

Then, by continuity of h and the definition of a boundary, for $\varepsilon = h(x, \tilde{u}, p) - h(x, u_*, p) > 0$, there exists $\hat{u}_*$ in the interior $\mathcal{U}^{\mathrm{o}}$ s.t. $h(x, \tilde{u}, p) - h(x, \hat{u}_*, p) < \varepsilon$, which implies

$$h(x, \hat{u}_*, p) > h(x, u_*, p),$$

meaning that $u_* \in \mathcal{U}^{\mathrm{o}}$ is not a maximum of the map $u \mapsto h(x, u, p)$ over $\mathcal{U}^{\mathrm{o}}$, a contradiction. Therefore, there is no $\tilde{u} \in \partial\mathcal{U}$ s.t. (I.14) holds; we conclude that u_* is a maximum of the map $u \mapsto h(x, u, p)$ over $\mathcal{U}^{\mathrm{o}} \cup \partial\mathcal{U} = \mathcal{U}$. $\square$

Proof of Proposition 5.5 (§5.2). Fix the policy π and for simplicity, denote with slight abuse of notation

$$v \doteq v_\pi, \quad X_t(x) \doteq \mathbb{G}_\pi^x[X_t], \quad R_t(x) \doteq \mathbb{G}_\pi^x[R_t].$$

Here, the dependencies on the policy π are implicit, and $X_t(x)$ is assumed to exist uniquely for all $t \geq 0$ and all $x \in \mathcal{X}$ (see §2). Also note that $R_t(x) = r_\pi(X_t(x))$ since

$$R_t(x) = \mathbb{G}_\pi^x[R_t] = r_\pi(\mathbb{G}_\pi^x[X_t]) = r_\pi(X_t(x)).$$

Suppose v is bounded and fix $x_0 \in \mathcal{X}$. Then, by continuity of $x \mapsto X_t(x)$ (e.g., Khalil, 2002, Theorem 3.5) and r_π, we have: for any $\eta > 0$ and any $\beta > 0$, there exists $\delta \equiv \delta(\beta, \eta) > 0$ such that

$$\|x - x_0\| < \delta \implies \big|R_t(x) - R_t(x_0)\big| < \beta \quad \forall t \in [0, \eta],$$

from which and the integral BE (8), we obtain that whenever $\|x - x_0\| < \delta$,

$$\big|v(x) - v(x_0)\big| \le \int_0^\eta \gamma^t \cdot \big|R_t(x) - R_t(x_0)\big|\, dt + \gamma^\eta \cdot \big|v\big(X_\eta(x)\big)\big| + \gamma^\eta \cdot \big|v\big(X_\eta(x_0)\big)\big| < \beta \cdot \eta + 2 \cdot \gamma^\eta \cdot M,$$

where $M > 0$ is a bound of v, i.e., a constant such that $\sup_{x\in\mathcal{X}} |v(x)| \le M$. Since $\beta, \eta > 0$ are arbitrary, for given $\varepsilon > 0$, choose $\beta = \varepsilon/2\eta$ and any $\eta > 0$ s.t. $\gamma^\eta < \varepsilon/(4M)$. Then, we conclude that for any $\varepsilon > 0$, there exists $\delta \equiv \delta(\varepsilon) > 0$ s.t.

$$\|x - x_0\| < \delta \implies \big|v(x) - v(x_0)\big| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

the ε-δ statement of the continuity of v ($= v_\pi$) at x_0, and the proof is completed as $x_0 \in \mathcal{X}$ is arbitrary. □

Proof of Proposition 5.6 (§5.2). If v is bouned, then since $\mathbb{G}_\pi^x[v(X_t)] = v\big(\mathbb{G}_\pi^x[X_t]\big)$, $t \mapsto \mathbb{G}_\pi^x[v(X_t)]$ for any policy π is bounded over $\mathbb{T}$ (uniformly in $x \in \mathcal{X}$); v satisfies the boundary condition (12) by Lemma I.9 below. □

Lemma I.9 *In the discounted case, the boundary condition* (12) *is true if $t \mapsto \mathbb{G}_\pi^x[v(X_t)]$ is bounded for each $x \in \mathcal{X}$.*

Proof. For $x \in \mathcal{X}$, let $M_x > 0$ be a constant s.t. $\sup_{t\in\mathbb{T}} \mathbb{G}_\pi^x\big[|v(X_t)|\big] \le M_x$. Then, since $\gamma \in (0, 1)$, we have

$$0 \le \lim_{t\to\infty} \mathbb{G}_\pi^x\big[\,\gamma^t \cdot |v(X_t)|\,\big] \le \lim_{t\to\infty} M_x \cdot \gamma^t = 0 \qquad \forall x \in \mathcal{X},$$

implying the boundary condition (12). □

Proof of Corollary 5.7 (§5.2). For the first part, Proposition 5.6 and Theorem 2.5 ensure $v = v_\pi$, hence v_π is bounded. Next, if v_π is bounded (hence admissible), then we have $v_\pi \leqslant v_{\pi'} \leqslant \overline{v}$ by Theorem 2.7 and Lemma 2.1, and thus $v_{\pi'}$ is also bounded. □

Proof of Corollary 5.9 (§5.2). For any given policy π, we can choose κ and $\rho(x)$ in (4) under Assumption 5.8 as $\kappa = 0$ and a constant function $\rho(x) \equiv \inf\{r(y, u) : (y, u) \in \mathcal{X} \times \mathcal{U}\} \in \mathbb{R}$, hence v_π is bounded by Proposition 2.2; the remaining proof is now obvious by Proposition 5.5 and Corollary 5.7. □

Proof of Theorem 5.10 (§5.3). Since (15) (hence, (13) for $v = v_\pi$) is true for $\pi \in \Pi_{\mathsf{a}}$ and the maximal policy π' over it,

$$\dot{v}_\pi(x, \pi'(x)) \ge -r_{\pi'}(x) + \alpha \cdot v_\pi(x) \ge -(r_{\mathsf{max}} - \alpha \cdot v_\pi(x)) = -\alpha \cdot (\overline{v} - v_\pi(x)) \qquad \forall x \in \mathcal{X},$$

where the last equality comes from Lemma 2.1. Let $J_\pi \doteq \overline{v} - v_\pi$. Then, the inequality can be expressed as

$$\dot{J}_\pi(x, \pi'(x)) \le \alpha J_\pi(x) \qquad \forall x \in \mathcal{X},$$

by substituting $\dot{v}_\pi = -\dot{J}_\pi$ and rearranging it. By $\pi \in \Pi_{\mathsf{a}}$ and the Assumptions, we see that u_* and ∇v_π are locally Lipschitz. Hence, π' given by (16) is locally Lipschitz (i.e., $\pi' \in \Pi_{\mathsf{Lip}}$); the application of Lemma I.10 below results in $t_{\mathsf{max}}(x; \pi') = \infty$ for all $x \in \mathcal{X}$. Now that the state trajectory exists globally and uniquely, we conclude $\pi' \in \Pi_{\mathsf{a}}$ and $\pi \preccurlyeq \pi'$ by Theorem 2.7. □

Lemma I.10 *Suppose there exist a C^1 function $J : \mathcal{X} \to \mathbb{R}$ and $\mathcal{K}_\infty$-functions ρ_1 and ρ_2 such that for all $x \in \mathcal{X}$,*

$$\rho_1(\|x\|_\Omega) \le J(x) \le \rho_2(\|x\|_\Omega) \tag{I.15}$$

$$\dot{J}(x, \pi(x)) \le \alpha J(x) \tag{I.16}$$

for a compact subset $\Omega \subset \mathcal{X}$ and a policy $\pi \in \Pi_{\mathsf{Lip}}$, then $t_{\mathsf{max}}(x; \pi) = \infty$ for all $x \in \mathcal{X}$.

Proof. We denote $t_{\mathsf{max}}(x) \doteq t_{\mathsf{max}}(x; \pi)$ for simplicity and show $t_{\mathsf{max}}(x) = \infty$ for all $x \in \mathcal{X}$. Here, $t_{\mathsf{max}}(x) \in (0, \infty]$ is defined for each $x \in \mathcal{X}$ since π and thus f_π are locally Lipschitz.

First, applying Grönwall (1919)'s inequality to (I.16), we obtain $\mathbb{G}_\pi^x[J(X_t)] \le e^{\alpha t} \cdot J(x)$ and by (I.15),

$$\mathbb{G}_\pi^x[\rho_1(\|X_t\|_\Omega)] \le \mathbb{G}_\pi^x[J(X_t)] \le e^{\alpha t} \cdot \rho_2(\|x\|_\Omega) \qquad \forall x \in \mathcal{X}.$$

Since the inverse of a $\mathcal{K}_\infty$ function ρ_1^{-1} exists and is also $\mathcal{K}_\infty$ (Khalil, 2002, Lemma 4.2), we obtain

$$\mathbb{G}_\pi^x[\|X_t\|_\Omega] \le \rho(x,t) \doteq \rho_1^{-1}\big(e^{\alpha t}\cdot \rho_2(\|x\|_\Omega)\big) \qquad \forall x \in \mathcal{X}.$$

Now, suppose $t_{\mathsf{max}}(x)$ *is finite for some* $x \in \mathcal{X}$. Then, $t \mapsto \mathbb{G}_\pi^x[\|X_t\|_\Omega]$ is bounded by $\rho(x, t_{\mathsf{max}}(x))$ and since Ω is compact, for all $t \in [0, t_{\mathsf{max}}(x))$, the state trajectory $t \mapsto \mathbb{G}_\pi^x[X_t]$ remains within the compact set given by

$$\big\{y \in \mathcal{X} : \|y\|_\Omega \le \rho(x, t_{\mathsf{max}}(x))\big\}$$

hence is uniquely defined for all $t \in \mathbb{T}$ (see Proposition G.3 in §G.2), a contradiction: $t_{\mathsf{max}}(x) = \infty$ for such $x \in \mathcal{X}$ that $t_{\mathsf{max}}(x) < \infty$. Therefore, we conclude $t_{\mathsf{max}}(x) = \infty$ *for all* $x \in \mathcal{X}$. □

Proof of Lemma 5.11. The positive definiteness of c_π is obvious by (43) and $c_\pi(0) = c(0, \pi(0)) = 0$ ($\because \pi(0) = 0$). □

Proof of Lemma 5.12. a. Since $\pi \in \Pi_{\mathsf{a}} \subseteq \Pi_0$, (i) $c_\pi(0) = 0$ by Lemma 5.11; (ii) $x_{\mathsf{e}} = 0$ is an equilibrium point under π ($\because f_\pi(0) = f(0, \pi(0)) = 0$), that is, $\mathbb{G}_\pi^0[X_t] \equiv 0$. Hence, $\mathbb{G}_\pi^0[C_t] = c_\pi\big(\mathbb{G}_\pi^0[X_t]\big) = 0$ for all $t \in \mathbb{T}$, implying $J_\pi(0) = 0$. By $\pi \in \Pi_{\mathsf{a}}$, we also have $t_{\mathsf{max}}(x;\pi) = \infty$ and

$$J_\pi(x) = \mathbb{G}_\pi^x\left[\int_0^\infty \gamma^t \cdot C_t\, dt\right] \in [0, \infty) \qquad \forall x \in \mathcal{X}.$$

Since $c_\pi(0) = 0$ and $c_\pi(x) > 0$ for any $x \neq 0$ by Lemma 5.11, and $t \mapsto C_t$ is continuous, we have that for each $x \neq 0$, there exists $\eta > 0$ s.t. $\mathbb{G}_\pi^x[C_t] > 0$ for all $t \in [0, \eta)$. Therefore, by the integral BE (8),

$$J_\pi(x) > \gamma^\eta \cdot \mathbb{G}_\pi^x[J_\pi(X_\eta)] \ge 0 \qquad \forall x \neq 0,$$

that is, $J_\pi(x) > 0$ for each $x \neq 0$. This and $J_\pi(0) = 0$ prove that J_π is positive definite.

b. and **c.** Since $\pi \in \Pi_{\mathsf{a}}$ satisfies the differential BE (9), we have

$$\dot{J}_\pi(x, \pi(x)) = \alpha J_\pi(x) - c_\pi(x) \quad \forall x \in \mathcal{X}. \tag{I.17}$$

Here, $\dot{J}_\pi(0, \pi(0)) = 0$ since J_π and c_π are positive definite; the proof is now obvious by (44), (45), and (I.17). □

Proof of Theorem 5.13 (§5.4). The proof is obvious by Lemma 5.12 and Lyapunov's stability theorems (Khalil, 2002, Theorems 4.1 and 4.2), except that the asymptotic stability is global when $\gamma = 1$ but J_π is *not* radially unbounded. To prove this case, fix $\pi \in \Pi_{\mathsf{a}}$ and let $x_{\mathsf{e}} = 0$ be asymptotically stable under π. $\mathcal{B}_\pi \subseteq \mathcal{X}$ denotes the basin of attraction under π, i.e., the set of all points $x \in \mathcal{X}$ s.t. $X_t(x) \to 0$ as $t \to 0$, where $X_t(x) \doteq \mathbb{G}_\pi^x[X_t]$ denotes the state trajectory under π starting at $X_0 = x \in \mathcal{X}$. Here, the dependency of $X_t(x)$ on π is implicit. Also note that

(1) since $\mathcal{B}_\pi$ is open (Khalil, 2002, Lemma 8.1), there exists $r > 0$ such that

$$\|x\| < r \quad \Longrightarrow \quad x \in \mathcal{B}_\pi; \tag{I.18}$$

(2) since c_π is positive definite by Lemma 5.11 and continuous by definitions, (46) implies that

$$\varphi_\pi(r) \doteq \inf\{c_\pi(x) : \|x\| \ge r\} > 0; \tag{I.19}$$

(3) by time-invariance $X_{\tau+t}(x) = X_t(X_\tau(x))$,

$$x \notin \mathcal{B}_\pi \quad \Longrightarrow \quad X_t(x) \notin \mathcal{B}_\pi \text{ for all } t \ge 0 \tag{I.20}$$

(if $X_\tau(x) \in \mathcal{B}_\pi$ for some $\tau > 0$, then we have a contradiction "$x \in \mathcal{B}_\pi$" ($\because \lim_{t\to\infty} X_{\tau+t}(x) = \lim_{t\to\infty} X_t(X_\tau(x)) = 0$).

The proof will be done by contradiction. Suppose $\mathcal{B}_\pi \neq \mathcal{X}$. Then, it implies that there exists $x \notin \mathcal{B}_\pi$ in $\mathcal{X}$ and $r > 0$ such that

$$\|X_t(x)\| \ge r \text{ for all } t \in \mathbb{T} \tag{I.21}$$

by (I.20) and then the contraposition of (I.18). Finally, applying (I.19) and (I.21) to the cost value function for $\gamma = 1$ yields

$$J_\pi(x) = \lim_{\eta\to\infty} \int_0^\eta c_\pi(X_t(x))\, dt \ge \varphi_\pi(r) \cdot \lim_{\eta\to\infty} \int_0^\eta 1\, dt = \infty,$$

a contradiction to $\pi \in \Pi_{\mathsf{a}}$. Therefore, $\mathcal{B}_\pi = \mathcal{X}$ and thus the asymptotic stability under $\pi \in \Pi_{\mathsf{a}}$ and $\gamma = 1$ is global. □

Proof of Theorem 5.15. By global asymptotic stability, for each $x \in \mathcal{X}$, the state trajectory $t \mapsto \mathbb{G}_\pi^x[X_t]$ exists uniquely and globally over $\mathbb{T}$, and $\mathbb{G}_\pi^x[X_t] \to 0$ as $t \to \infty$. Hence, continuity of v at 0 and $v(0) = 0$ imply that $\mathbb{G}_\pi^x[\gamma^t\, v(X_t)] \to 0$ as $t \to \infty$, for all $x \in \mathcal{X}$; the proof is completed by Theorem 2.5. □

Proof of Theorem 5.16. Since J is positive definite, it is lower-bounded by zero. Therefore, the application of Corollary C.2 in §C completes the proof (note that $v = -J$ and $r_\pi = -c_\pi$). □

Proof of Theorem 5.17. J_π is (i) positive definite (by Lemma 5.12a), (ii) $\mathrm{C}^1_{\mathsf{Lip}}$ (by $\pi \in \Pi_{\mathsf{a}}$ and the regularity $\mathcal{V}_{\mathsf{a}} \subset \mathrm{C}^1_{\mathsf{Lip}}$), and (iii) radially unbounded. So, Lemma I.11 below implies that there exist $\mathcal{K}_\infty$ functions ρ_1 and ρ_2 s.t.

$$\rho_1(\|x\|) \le J_\pi(x) \le \rho_2(\|x\|) \qquad \forall x \in \mathcal{X}.$$

Moreoever, $\bar{v} = 0$ by Lemma 2.1 since c is positive definite (see (43)), hence $r_{\mathsf{max}} = -\min\{c(x,u) : (x,u) \in \mathcal{X} \times \mathcal{U}\} = 0$. Therefore, we conclude $\pi' \in \Pi_{\mathsf{a}}$ and $J_{\pi'} \leqslant J_\pi$ by Theorem 5.10 with $\Omega = \{0\}$ (i.e., $\|x\|_\Omega = \|x\|$) and Lemma I.12 below. □

Lemma I.11 *(Khalil, 2002, Lemma 4.3) If $J : \mathcal{X} \to \mathbb{R}$ is continuous, positive definite, and radially unbounded, then there exists $\mathcal{K}_\infty$ functions ρ_1 and ρ_2 s.t. $\rho_1(\|x\|) \le J(x) \le \rho_2(\|x\|)$ for all $x \in \mathcal{X}$.*

Lemma I.12 *If a policy π is given by $\pi(x) = u_*(x, \nabla v(x))$ for a negative definite function $v \in \mathrm{C}^1_{\mathsf{Lip}}$ on $\mathcal{X}$, then $\pi \in \Pi_0$.*

Proof. Since v is C^1 and negative definite, $\nabla v(0) = 0$ ($\because x = 0$ is the global maximum). Then, by the definition (5) of the Hamiltonian and the argmax-formula (14) of u_*, we have at $x = 0$:

$$\pi(0) = u_*(0, \nabla v(0)) = u_*(0,0) \in \operatorname*{arg\,max}_{u \in \mathcal{U}} r(0,u).$$

Since $r\ (= -c)$ is negative definite by (43), $(x,u) = (0,0)$ is the global maximum of r, hence $\pi(0) = 0$. Moreover, π is locally Lipschitz since so are both u_* and ∇v. Therefore, we conclude $\pi \in \Pi_0$. □

Proof of Theorem 5.18. $\pi_0 \in \Pi_{\mathsf{a}}$ by assumption. Suppose $\pi_{i-1} \in \Pi_{\mathsf{a}}$ for some $i \in \mathbb{N}$. Then, (47) implies that for $\gamma \in (0,1)$, we have $\tilde{\kappa}_i \cdot J_i \leqslant c_{\pi_{i-1}}$ for $\tilde{\kappa}_i \doteq \alpha\kappa_i > 0$. For $\gamma = 1$, $x_{\mathsf{e}} = 0$ is globally asymptotically stable under π_{i-1} by Theorem 5.13. For both cases, we have $J_i = J_{\pi_{i-1}}$ by Theorems 5.15 and 5.16. Therefore, (i) global asymptotic stability under (48), (ii) $\pi_i \in \Pi_{\mathsf{a}}$, and (iii) $J_{\pi_i} \leqslant J_{\pi_{i-1}}$ are all obvious by the radial unboundedness of J_i (assumed!) and Theorems 5.13 and 5.17. Finally, the mathematical induction completes the proof. □

I.4 Proofs of Some Facts in §G.3 LQRs

In this appendix, for completeness, we provide proofs of some of the facts used in §G.3.

Proof of Stabilizability and Observability of (A^α, B, S). Note that (A,B) is stabilizable iff $\mathrm{rank}\,([\,A - \lambda I \;\; B\,]) = l$ $\forall \lambda \in \mathbb{C}$ such that $\mathrm{Re}\lambda \ge 0$ (Zhou and Doyle, 1998, Theorem 3.2). Since (A^0, B) is stabilizable, therefore, we obtain

$$\mathrm{rank}\,([\,A^\alpha - \lambda I \;\; B\,]) = \mathrm{rank}\big([\,A^0 - (\lambda + \alpha/2)I \;\; B\,]\big) = l \tag{I.22}$$

for all $\lambda \in \mathbb{C}$ such that $\mathrm{Re}\lambda \ge -\alpha/2$ with $\alpha \ge 0$. Hence, (I.22) holds whenever $\mathrm{Re}\lambda \ge 0$, i.e., (A^α, B) is stabilizable. Similarly, (S,A) is observable iff $\mathrm{rank}\,\big([\,A^\mathsf{T} - \lambda I \;\; S\,]\big) = l$ for all $\lambda \in \mathbb{C}$ (Zhou and Doyle, 1998, Theorem 3.3). Since

$$\mathrm{rank}\big([\,(A^\alpha)^\mathsf{T} - \lambda I \;\; S\,]\big) = \mathrm{rank}\big([\,(A^0)^\mathsf{T} - \bar{\lambda} I \;\; S\,]\big) = l$$

for all $\bar{\lambda} \doteq \lambda + \alpha/2 \in \mathbb{C}$ and thus for all $\lambda \in \mathbb{C}$, the observability of (S, A^α) is now obvious by that of (S, A^0). □

Proof of Existence of P s.t. $P_i \to P$. For $x, y \in \mathcal{X}$, let $B_i : \mathcal{X}^2 \to \mathbb{R}$ be defined for $J_i(x) = x^\mathsf{T} P_i x\ (= -v_i(x))$ as

$$B_i(x,y) \doteq J_i(x+y) - J_i(x-y) = 4\, x^\mathsf{T} P_i y,$$

and denote $\hat{J}_* \doteq -\hat{v}_*$. Then, since $J_i \to \hat{J}_*$ pointwise by Theorem 4.2a, B_i pointwise converges to B defined as

$$B(x,y) \doteq \hat{J}_*(x+y) - \hat{J}_*(x-y).$$

Since B_i is bilinear and symmetric, we have the following claim.

Claim I.13 *B is* **a.** *bilinear and* **b.** *symmetric.*

By Claim I.13, there exists a symmetric matrix P s.t. $B(x,y) = 4\,x^\mathsf{T}Py$. Moreover, $\hat{J}_*(0) = 0$ ($\because 0 = J_i(0) \to \hat{J}_*(0)$). Therefore, $\hat{J}_*$ is quadratic (and thus continuous) as shown below:

$$\hat{J}_*(x) = \hat{J}_*(x) - \hat{J}_*(0) = B(x/2, x/2) = x^\mathsf{T}Px.$$

Next, let $\Omega \doteq \{x \in \mathcal{X} \colon \|x\| = 1\}$. Then, Ω is obviously compact, hence $J_i \to \hat{J}_*$ uniformly on Ω by Theorem 4.2b. Moreover, since $\hat{J}_* \leqslant J_i$ for every $i \in \mathbb{N}$ by Theorem 4.1, every $P_i - P$ is positive semidefinite and thus represented as $P_i - P = N_i^\mathsf{T}N_i$ for some $N_i \in \mathbb{R}^{l\times l}$ (Chen, 1998, Theorem 3.7.3). Therefore, by the definition of d_Ω,

$$d_\Omega(J_i, \hat{J}_*) = \sup_{x\in\Omega} \left| x^\mathsf{T}(P_i - P)x \right| = \sup_{\|x\|=1} \|N_i x\|^2 = |||N_i|||^2 = |||N_i^\mathsf{T}N_i||| = |||P_i - P||| \geq 0,$$

where $|||N_i|||^2 = |||N_i^\mathsf{T}N_i|||$ holds since $|||\cdot|||$ is induced by the Euclidean norm $\|\cdot\|$. Finally, since $d_\Omega(J_i, \hat{J}_*) \to 0$ by the uniform convergence $J_i \to \hat{J}_*$ on Ω, we conclude that $|||P_i - P||| \to 0$, i.e., $P_i \to P$.

(Proof of Claim I.13). a. Bilinearity. Since B_i is bilinear (i.e., $B_i(x,y) = 4\,x^\mathsf{T}P_iy$),

$$B_i(x_1 + x_2, y) = B_i(x_1, y) + B_i(x_2, y) \qquad \forall x_1, x_2, y \in \mathcal{X},$$

where both sides converge to $B(x_1 + x_2, y)$ and $B(x_1, y) + B(x_2, y)$, respectively. This proves that for each $y \in \mathcal{X}$, $B(\cdot, y)$ preserves the vector addition. Similarly, we can prove that $B(\alpha x, y) = \alpha B(x, y)$ for all $x, y \in \mathcal{X}$ and $\alpha \in \mathbb{R}$. Therefore, $B(\cdot, y)$ is linear and in the same way, so is $B(x, \cdot)$, meaning that B is bilinear.

b. Symmetry. Since each P_i is symmetric, so is each B_i, hence for all $x, y \in \mathcal{X}$, $B_i(x,y) = B_i(y,x)$; by the pointwise convergence $B_i \to B$, we have $B_i(x,y) \to B(x,y)$ and $B_i(y,x) \to B(y,x)$; by the uniqueness of the limit point, $B(x,y) = B(y,x)$, for all $x \in \mathcal{X}$. Therefore, B is symmetric. □

Proof of Quadratic Convergence $P_i \to P_*$. Note that the matrix formula (G.4) can be rewritten for $i \in \mathbb{N} \setminus \{1\}$ as

$$(A_{i-1}^\alpha)^\mathsf{T}P_i + P_iA_{i-1}^\alpha = -\mathcal{S} - \mathcal{K}_{i-1}^\mathsf{T}\Gamma\mathcal{K}_{i-1}, \tag{I.23}$$

where $\mathcal{S} \doteq S - E\Gamma^{-1}E^\mathsf{T}$ is a Schur complement of $\mathcal{W}$ and thus positive semi-definite (Horn and Johnson, 1990); $\mathcal{K}_{i-1} \doteq \Gamma^{-1}B^\mathsf{T}P_{i-1}$. Here, by the policy improvement and definitions in §G.3, A_{i-1}^α in (I.23) can be rewritten as

$$A_{i-1}^\alpha = \mathcal{A}^\alpha - B\mathcal{K}_{i-1} \text{ for } \mathcal{A}^\alpha \doteq A^0 - \alpha I/2 - B\Gamma^{-1}E^\mathsf{T},$$

where $\mathcal{A}^\alpha$ is different from A^α ($\doteq A - \alpha I/2$). Therefore, one can see that (I.23) is exactly same as the well-known matrix-form PI (Kleinman, 1968) for the LQR (G.3) with A^0, S, and E replaced by $\mathcal{A}^\alpha$, $\mathcal{S}$ and 0, respectively.

Moreover, each policy $\tilde{\pi}_i(x) = -\mathcal{K}_i x$ is admissible by Theorem 4.1, meaning that the states under $\tilde{\pi}_i$ converge to zero as $t \to \infty$ (as discussed in §G.3). This convergence happens for the linear system iff A_i^α ($= \mathcal{A}^\alpha - B\mathcal{K}_i$) is Hurwitz (Chen, 1998; Khalil, 2002) and thus also proves that $(\mathcal{A}^\alpha, B)$ is stabilizable. Since (S, A^α) is observable, so is $(\mathcal{S}, \mathcal{A}^\alpha)$ by Lancaster and Rodman (1995, Lemma 16.2.7) and nondegenerate $\mathcal{W}$. Therefore, the quadratic convergence $P_i \to P_*$ is directly proven by following Kleinman (1968)'s proof (when $(\mathcal{A}^\alpha, B)$ is controllable) or generally, by Lee et al. (2014, Theorem 5 and Remark 4 with $\hbar \to \infty$). Additionally, this approach can provide an alternative proof of Theorem G.5. □

Additional References

Anderson, B. and Moore, J. B. *Optimal control: linear quadratic methods*. Prentice-Hall, Inc., 1989.

Arnold III, W. Numerical Solution of Algebraic Matrix Riccati Equations. Technical report, Naval Weapons Center, China Lake, CA, 1984.

Bessaga, C. On the converse of Banach "fixed-point principle". *Colloquium Mathematicae*, 7(1):41–43, 1959.

Brouwer, L. E. Beweis der invarianz desn-dimensionalen gebiets. *Mathematische Annalen*, 71(3):305–313, 1911.

Chen, C.-T. *Linear system theory and design*. Oxford University Press, Inc., 1998.

Grönwall, T. H. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 1990.

Kachurovskii, R. I. Monotone operators and convex functionals. *Uspekhi Mat. Nauk*, 15(4(94)):213–215, 1960.

Kirk, W. A. and Sims, B. *Handbook of metric fixed point theory*. Springer Science & Business Media, 2013.

Lancaster, P. and Rodman, L. *Algebraic Riccati equations*. Oxford University Press, 1995.

Lee, J. and Sutton, R. S. Policy iterations for reinforcement learning problems in continuous time and space — fundamental theory and methods. *To appear in Automatica.*, 2020b.

Lee, J. Y., Park, J. B., and Choi, Y. H. Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems. *Automatica*, 48(11):2850–2859, 2012.

Lee, J. Y., Park, J. B., and Choi, Y. H. On integral generalized policy iteration for continuous-time linear quadratic regulations. *Automatica*, 50(2):475–489, 2014.

Lyashevskiy, S. Constrained optimization and control of nonlinear systems: new results in optimal control. In *Decision and Control, 1996., Proceedings of the 35th IEEE Conference on*, volume 1, pages 541–546, 1996.

Mehrmann, V. L. *The autonomous linear quadratic control problem: theory and numerical solution*, volume 163. Springer, 1991.

Remmert, R. *Theory of complex functions*, volume 122. Springer Science & Business Media, 1991.

Royden, H. L. Real analysis (third edtion). *New Jersey: Printice-Hall Inc*, 1988.

Sundaram, R. K. *A first course in optimization theory*. Cambridge university press, 1996.

Vrabie, D., Pastravanu, O., Abu-Khalaf, M., and Lewis, F. L. Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, 45(2):477–484, 2009.

Wang, D., Li, C., Liu, D., and Mu, C. Data-based robust optimal control of continuous-time affine nonlinear systems with matched uncertainties. *Information Sciences*, 366:121–133, 2016.

Zhou, K. and Doyle, J. C. *Essentials of robust control*. Prentice hall Upper Saddle River, NJ, 1998.

Zhu, L. M., Modares, H., Peen, G. O., Lewis, F. L., and Yue, B. Adaptive suboptimal output-feedback control for linear systems using integral reinforcement learning. *IEEE Transactions on Control Systems Technology*, 23(1):264–273, 2015.