Policy Iteration for Discounted Reinforcement Learning Problems in Continuous Time and Space

Jae Young Lee*

Reinforcement Learning and Artificial Intelligence
Department of Computing Science
University of Alberta,
4-08 Athabasca Hall, Edmonton, AB, Canada, T6G 2E8
jyounglee@ualberta.ca

Richard S. Sutton

Reinforcement Learning and Artificial Intelligence
Department of Computing Science
University of Alberta,
2-21 Athabasca Hall, Edmonton, AB, Canada T6G 2E8
rsutton@ualberta.ca

Abstract

Recent advances in various fields regarding decision making, especially regarding reinforcement learning (RL), have revealed the interdisciplinary connections among their findings. For example, actor and critic in computational RL are shown to play the same roles of dorsal and ventral striatum; goal-directed and habitual learning is strongly relevant to model-based and model-free computational RL, respectively. Among the different methodologies in those fields, theoretical approach in machine learning community has established the well-defined computational RL framework in discrete domain and a dynamic programming method known as *policy iteration* (PI), both of which served as the fundamentals in computational RL methods. The main focus of this work is to develop such RL framework and a series of PI methods in continuous domain, with its environment modeled by an ordinary differential equation (ODE). Similar to the discrete case, the PI methods are designed to recursively find the best decision-making strategy by iterating *policy evaluation* (as a role of *critic*) and *policy improvement* (as a role of *actor*). Each proposed one is either model-free corresponding to habitual learning, or partially model-free (or partially model-based) corresponding to somewhere between goal-directed (model-based) and habitual (model-free) learning. This work also provides theoretical background and perhaps, the basic principles to RL algorithms with a real physical task which is usually modeled by ODEs. In detail, we propose on-policy PI and then four off-policy PI methods—the two off-policy methods are the ideal PI forms of advantage updating and Q-learning, and the other two are extensions of the existing off-policy PI methods; compared to PI in optimal control, ours do not require an initial stabilizing policy. The mathematical properties of admissibility, monotone improvement, and convergence are all rigorously proven; simulation examples are provided to support the theory.

Keywords: policy iteration, reinforcement learning, off-policy, discounting, continuous time, continuous space, convergence, model-free, partially model-free, ordinary differential equation, actor-critic, Hamilton-Jacobi-Bellman

Acknowledgements

The authors gratefully acknowledge the support of Alberta Innovates – Technology Futures, the Alberta Machine Intelligence Institute, Google Deepmind, and the Natural Sciences and Engineering Research Council of Canada.

*Corresponding author.

1 Introduction

Decision making problems have been studied in various disciplines such as machine learning, neuroscience, psychology, and optimal control, with a number of different methodologies, e.g., simulation-based, theoretical, biological, and empirical approaches. Recent advances in those fields regarding decision making, especially regarding reinforcement learning (RL), have revealed the interdisciplinary connections among their findings. Closely related to this work is that the actor-critic structure in computational RL presumably exists in a brain mechanism of dorsal (actor) and ventral (critic) striatum, and that model-based and model-free computational RL are very relevant to goal-directed and habitual learning in psychology, respectively [5].

Among the different approaches, the theoretical works in machine learning community have established the fundamental mathematical model of the computational RL frameworks, with its environment represented by a finite Markov decision process (MDP), and a dynamic programming method known as *policy iteration* (PI) with its mathematical properties. Here, PI recursively finds the best decision-making strategy, called the optimal policy, with *policy evaluation* as a role of *critic* and *policy improvement* as a role of *actor* [5]. Together with the other dynamic programming methods, PI has served as a fundamental principle to develop computational RL methods in the MDP framework.

On the other hand, different from the MDP environment, the dynamics of real physical world is usually modeled by ordinary differential equations (ODEs) in continuous time and space (CTS). In such continuous domain, a similar PI method has also recently come to be studied in optimal control fields [4]. An interesting point of such a PI method is that it is partially model-free, lying somewhere on the line between goal-directed (model-based) and habitual (model-free) learning in psychology. This is a quite different situation from the MDP case, where PI is completely model-based, but the actor-critic methods developed so far are all model-free and thus thought to be relevant only to habitual behavior [5]. There are also off-policy versions of PI in CTS, each of which is either completely or partially model-free [3]. However, it is not theoretically straightforward to extend such PI methods in CTS from optimal control to the general RL framework.

In this work, we precisely define the RL framework in CTS, with its environment represented by an ODE, and then propose a series of PI methods in that framework—one is the extension of integral PI [3, 4], and the other four are its off-policy versions (two are the ideal PI forms of advantage updating [1, 2] and Q-learning in CTS, and the other two correspond to the off-policy PI methods in [3]). Similar to PI in the MDP environment, their mathematical properties of admissibility, monotone improvement, and convergence to the optimal solution are all rigorously proven. As opposed to the PI methods in optimal control, all of the proposed PI schemes do not require an initial stabilizing policy, by virtue of discounting, while each one still remains to be either completely or partially model-free. Simulation examples are also provided to support the theory. Though our PI methods are not online incremental RL algorithms, we believe that this work provides the theoretical background of and intuition to the RL methods in CTS, e.g., those in [1, 2]. This theoretical work, lying between the fields of machine learning and optimal control, would also provide some motivational links in the future between the RL methods and the findings in neuroscience and psychology in CTS, perhaps similar to the interdisciplinary links between them in discrete time and space. For brevity, all of the Theorem proofs are omitted.

2 RL Problem Formulation in CTS

In the RL problem, $\mathcal{X} \doteq \mathbb{R}^n$ denotes the state space, and the action space $\mathcal{U} \subseteq \mathbb{R}^m$ is a m-dimensional manifold in $\mathbb{R}^m$ with (or without) boundary; $t \geq 0$ denotes a given specific time instant; the environment in CTS is described by an ODE $\dot{X}_\tau = f(X_\tau, U_\tau)$, where $\tau \in [t, \infty)$ is the time variable, $f : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ is a continuous function, $X_\tau \in \mathcal{X}$ is the state vector at time τ with its time-derivative $\dot{X}_\tau \in \mathcal{X}$, and the action trajectory $U_\tau \in \mathcal{U}$ is a right continuous function over $[t, \infty)$. A continuous function $\pi : \mathcal{X} \to \mathcal{U}$ is called a (stationary) policy whenever the state trajectory $\mathbb{E}_\pi[X_\tau | X_t = x]$ is uniquely defined for all $\tau \geq t$ and all $x \in \mathcal{X}$, where $\mathbb{E}_\pi[Z|X_t = x]$ has no stochastic role but means the deterministic value Z when $X_t = x$ and $U_\tau = \pi(X_\tau)$ for all $\tau \geq t$. We will denote $\Delta t > 0$ the time difference, $t' \doteq t + \Delta t$, and $X'_t \doteq X_{t'}$. The RL problem we consider is to find the optimal policy π_* that maximizes the value function (VF) $v_\pi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$

$$v_\pi(x) \doteq \mathbb{E}_\pi[G_t | X_t = x] \text{ with a discounted return } G_t \doteq \int_t^\infty \gamma^{\tau - t} R_\tau \, d\tau \text{ (upper bounded)}, \tag{1}$$

where $R_\tau \doteq R(X_\tau, U_\tau) \in \mathbb{R}$ is the immediate reward at time τ and $\gamma \in (0, 1)$ is the discount factor. The reward function $R : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ here is continuous and upper-bounded. A policy π (or its VF v_π) is said to be admissible, denoted by $\pi \in \Pi_a$ or $v_\pi \in \mathcal{V}_a$, if $v_\pi(x)$ is finite for all $x \in \mathcal{X}$, where Π_a and $\mathcal{V}_a$ denote the set of all admissible policies and VFs, respectively. Note that if R is bounded, then every possible policy π is guaranteed to be admissible and has a bounded value function. In our RL problem, we assume that every $v_\pi \in \mathcal{V}_a$ has its continuous gradient ∇v_π and that there is an optimal *admissible* policy π_* such that $v_\pi(x) \leq v_*(x)$ for any $x \in \mathcal{X}$ and any policy π, where v_* is the optimal VF.

3 Partially Model-Free Policy Iteration

Note that in our CTS case, any $v_\pi \in \mathcal{V}_a$ satisfies

$$\forall x \in \mathcal{X} \text{ and } \forall \Delta t > 0 : \begin{cases} \text{the Bellman equation: } v_\pi(x) = \mathbb{E}_\pi\Big[\mathcal{R}_t + \gamma^{\Delta t}\, v_\pi(X'_t)\Big| X_t = x\Big] \text{ with } \mathcal{R}_t \doteq \int_t^{t'} \gamma^{\tau - t}\, R_\tau \, d\tau; \\ \text{the boundary condition: } \lim_{k\to\infty} \gamma^{k\Delta t}\, \mathbb{E}_\pi[v_\pi(X_{t+k\Delta t})|X_t = x] = 0. \end{cases}$$

The policy improvement operation in CTS is defined in the limit $\Delta t \to 0$ as

$$\begin{aligned} \pi'(x) &\in \arg\max_{U_t \in \mathcal{U}} \lim_{\Delta t \to 0} \frac{1}{\Delta t} \cdot \mathbb{E}\Big[\mathcal{R}_t + \gamma^{\Delta t}\, v_\pi(X'_t) - v_\pi(X_t)\Big| X_t = x\Big] \\ &= \arg\max_{u \in \mathcal{U}} \big(R(x,u) + \dot{v}_\pi(x,u) + \ln\gamma \cdot v_\pi(x)\big) \quad \forall x \in \mathcal{X}, \end{aligned} \tag{2}$$

where π' is the improved policy, and $\dot{v}_\pi(x,u) = \nabla v_\pi(x) f(x,u)$ by chain rule.

Considering any decomposition: $f(x,u) = f_d(x) + f_c(x,u)$, where f_d is an *unknown* drift dynamics and f_c is a known input coupling dynamcs, and noting that any addition or subtraction of u-independent terms does not change the maximization process *with respect to* $u \in \mathcal{U}$, one can express (2) in terms of f_c as

$$\pi'(x) \doteq \arg\max_{u \in \mathcal{U}} \Big(R(x,u) + \nabla v_\pi(x) f_c(x,u)\Big) \tag{3}$$

which we call partially model-free policy improvement. Algorithm 1 is our partially model-free PI which finds the optimal solution without knowing the drift dynamics f_d. It starts with an initial admissible policy π_0; in each i-th step, it finds a function v_i satisfying the Bellman equation (policy evaluation), and then using v_i and f_c, the next policy π_{i+1} is updated by "(3) with $\pi' = \pi_{i+1}$ and $v_\pi = v_i$" (policy improvement). This process is recursively done until convergence. Our main theorem is as follows.

Algorithm 1: Partially Model-Free Policy Iteration

Initialize: $i \leftarrow 0$, $\Delta t > 0$, and $\pi_0 \in \Pi_a$;
repeat
  Policy Evaluation: find $v_i : \mathcal{X} \to \mathbb{R}$ such that $\forall x \in \mathcal{X}$:

$$v_i(x) = \mathbb{E}_{\pi_i}\Big[\mathcal{R}_t + \gamma^{\Delta t}\, v_i(X'_t)\Big| X_t = x\Big];$$

  Policy Improvement: find a policy π_{i+1} such that

$$\pi_{i+1}(x) \in \arg\max_{u \in \mathcal{U}} \Big(R(x,u) + \nabla v_i(x) f_c(x,u)\Big) \ \forall x \in \mathcal{X};$$

  $i \leftarrow i + 1$;
until *convergence is met.*

Assumption 1. *(Policy improvement condition) for each $\pi \in \Pi_a$, there is a policy π' such that* (3) *holds.*

Assumption 2. *(Boundary condition) $\forall i \in \mathbb{Z}_+$: if π_i is admissible, then $\lim_{k\to\infty} \gamma^{k\Delta t}\, \mathbb{E}_{\pi_i}[v_i(X_{t+k\Delta t})|X_t = x] = 0 \quad \forall x \in \mathcal{X}$.*

Assumption 3. *(Uniqueness of optimality) there is one and only one element $w_* \in \mathcal{V}_a$ over $\mathcal{V}_a$ that satisfies*

$$\textit{the Hamilton-Jacobi-Bellman equation: } 0 = \max_{u \in \mathcal{U}} \big(R(x,u) + \dot{w}_*(x,u) + \ln\gamma \cdot w_*(x)\big) \ \ \forall x \in \mathcal{X} \ \ \textit{(in fact, } w_* = v_*\textit{)}.$$

Theorem 1. *The sequences $\{\pi_i\}_{i=0}^\infty$ and $\{v_i\}_{i=0}^\infty$ generated by Algorithm 1 under Assumptions 1–3 satisfy the followings.*

- (P1) *$\pi_{i+1} \in \Pi_a$ and $v_i = v_{\pi_i} \in \mathcal{V}_a$ for all $i \in \mathbb{N} \cup \{0\}$;*
- (P2) *the policy is monotonically improved, i.e., $v_{\pi_0}(x) \le v_{\pi_1}(x) \le \cdots \le v_{\pi_i}(x) \le v_{\pi_{i+1}}(x) \le \cdots \le v_*(x)$ for all $x \in \mathcal{X}$;*
- (P3) *$v_i \to v_*$ with respect to some metric $d : \mathcal{V}_a \times \mathcal{V}_a \to [0,\infty)$, i.e., $\lim_{i\to\infty} d(v_i, v_*) = 0$;*
- (P4) *$v_i \to v_*$ pointwisely on $\mathcal{X}$ and uniformly on any compact subset of $\mathcal{X}$ if:*
  - *(a) the limit function $\hat{v}_* \doteq \lim_{i\to\infty} v_i$ belongs to $\mathcal{V}_a$;*
  - *(b) for every compact subset $\Omega \subset \mathcal{X}$, the policy iteration mapping $v_\pi \mapsto v_{\pi'}$ is continuous with respect to*

$$\textit{the uniform pseudometric } d_\Omega(v,w) \doteq \sup_{x\in\Omega}\big|v(x) - w(x)\big| \ \ (v, w \in \mathcal{V}_a).$$

4 Extensions to Off-Policy Policy Iteration

PI shown in Algorithm 1 can be extended to a series of its off-policy versions that use the behavior policy μ than the target policy π_i to generate the state trajectory X_τ (and the reward R_τ). To describe μ, we extend the concept of a policy established in Section 2. A function $\mu : [t,\infty) \times \mathcal{X} \to \mathcal{U}$ is called a (non-stationary) policy if: 1) $\mu(\tau,\cdot)$ is continuous for each fixed $\tau \ge t$ and $\mu(\cdot,x)$ is right continuous for each fixed $x \in \mathcal{X}$; 2) for each $x \in \mathcal{X}$, the state trajectory $\mathbb{E}^x_\mu[X_\tau]$ is uniquely defined for all $\tau \ge t$. Here, $\mathbb{E}^x_\mu[Z]$ means the deterministic value Z when $X_t = x$ and $U_\tau = \mu(\tau, X_\tau)\ \forall \tau \ge t$. A function $\mu : [t,\infty) \times \mathcal{X} \times \mathcal{U}_0 \to \mathcal{U}$ is said to be an AD policy over $\mathcal{U}_0 \subseteq \mathcal{U}$ if for each fixed $u \in \mathcal{U}_0$, $\mu(\cdot,\cdot,u)$ is a policy and $\mu(t,x,u) = u$ holds for all $x \in \mathcal{X}$ and all $u \in \mathcal{U}_0$. For an AD policy μ, we denote $\mathbb{E}^x_{\mu(\cdot,\cdot,u)}[Z]$ by $\mathbb{E}^{(x,u)}_\mu[Z]$.

By replacing policy evaluation and improvement, we propose four different off-policy PI methods—Advantage PI (API), QPI, Explorized PI (EPI), and Common PI (CPI). In policy evaluation of API, the advantage function a_π defined as $a_\pi(x,u) \doteq R(x,u) + \dot{v}_\pi(x) + \ln\gamma \cdot v_\pi(x)$ [1, 2] is estimated, along with v_π and the constraint $a_\pi(x,\pi(x)) = 0$. In QPI, the Q-function q_π so-defined in CTS as $q_\pi(x,u) = \kappa \cdot v_\pi(x) + a_\pi(x,u)$ for some $\kappa \neq 0$ is estimated in policy evaluation with the discounting $\beta \doteq \gamma \cdot e^\kappa > 0$ that should be different from $\gamma \in (0,1)$. Here, β determines κ in q_π, and the extremely large $|\kappa|$ may result in a significant performance degradation or extremely slow Q-learning [1]. Both a_π in API and q_π in QPI replace the policy improvement (3) with the respective model-free ones. EPI is the direct extension of PI with respect to the behavior policy μ without introducing any other function than v_π. CPI is the model-free modification of EPI when

- (C1) $f_c(x,u) = F_c(x)u$ for a continuous function $F_c : \mathcal{X} \to \mathbb{R}^{n\times m}$ (input-affine dynamics);
- (C2) $\mathcal{U}$ is convex and the reward R is given by $R(x,u) = R_0(x) - S(u)$ for a continuous upper-bounded function R_0 and a strictly convex function S, with its gradient $\nabla S : \mathcal{U} \to \mathbb{R}^{1\times m}$ that is continuous and has its inverse ∇S^{-1} on the interior of the action space domain $\mathcal{U}$.

The key idea of CPI is to estimate the C-function $c_\pi(x) \doteq F_c^\mathsf{T}(x)\nabla v_\pi^\mathsf{T}(x)$ in policy evaluation and then use it in policy improvement under (C1) and (C2) above. Here, note that the maximization (3) (and thus policy improvement of PI, EPI, and CPI) can be dramatically simplified under (C1) and (C2) as $\pi'(x) = \sigma(F_c^\mathsf{T}(x)\nabla v_\pi^\mathsf{T}(x)) = \sigma(c_\pi(x))$ with $\sigma^\mathsf{T} \doteq \nabla S^{-1}$.

The policy evaluation and improvement of the off-policy methods are summarized in Table 1, where we used the compact notations $R^{\pi_i} \doteq R(\cdot,\pi_i(\cdot))$, $a_i^{\pi_i} \doteq a_i(\cdot,\pi_i(\cdot))$, $q_i^{\pi_i} \doteq q_i(\cdot,\pi_i(\cdot))$, $f_c^{\pi_i} \doteq f_c(\cdot,\pi_i(\cdot))$, $\xi_\tau^{\pi_i} \doteq U_\tau - \pi_i(X_\tau)$, and

$$\mathcal{I}_\alpha(Z) \doteq \int_t^{t'} \alpha^{\tau-t} Z(X_\tau, U_\tau)\, d\tau \quad \text{and} \quad D_\alpha(v) \doteq v(X_t) - \alpha^{\Delta t} v(X_t')$$

for brevity. For example, the policy evaluation of Algorithm 1 can be expressed as $\mathbb{E}_{\pi_i}^x[D_\gamma(v_i)] = \mathbb{E}_{\pi_i}^x[\mathcal{I}_\gamma(R)]$. As shown in Table 1, API, QPI, and CPI are model-free while EPI requires the full-knowledge of an input-coupling dynamics f_c to run. On the other hand, while API and QPI explore the whole state-action space $\mathcal{X}\times\mathcal{U}$ to learn their respective functions (v_π, a_π) and q_π for all $(x,u) \in \mathcal{X}\times\mathcal{U}$, EPI and CPI search only the significantly smaller spaces $\mathcal{X}$ and $\mathcal{X}\times\{u_j\}_{j=0}^m$, respectively. This is because EPI and CPI both learn *no AD function* like a_π and q_π (see the last column of Table 1). In CPI, $u_0, u_1, \cdots, u_m \in \mathcal{U}$ in the search space $\mathcal{X}\times\{u_j\}_{j=0}^m$ are any vectors in $\mathcal{U}$ such that $\operatorname{span}\{u_j - u_{j-1}\}_{j=1}^m = \mathbb{R}^m$. Denote $v_i \doteq q_i(\cdot,\pi_i(\cdot))/\kappa$ in the QPI case. Then, all of the four off-policy PI satisfy the following theorem.

Theorem 2. *The sequences $\{\pi_i\}_{i=0}^\infty$ and $\{v_i\}_{i=0}^\infty$ generated by any of the four off-policy PI under Assumptions 1, 2, and 3 satisfy* (P1)–(P4) *in Theorem 1. Moreover, for all $i \in \mathbb{N}\cup\{0\}$: $a_i = a_{\pi_i}$ (API), $q_i = q_{\pi_i}$ (QPI), and $c_i = c_{\pi_i}$ (CPI).*

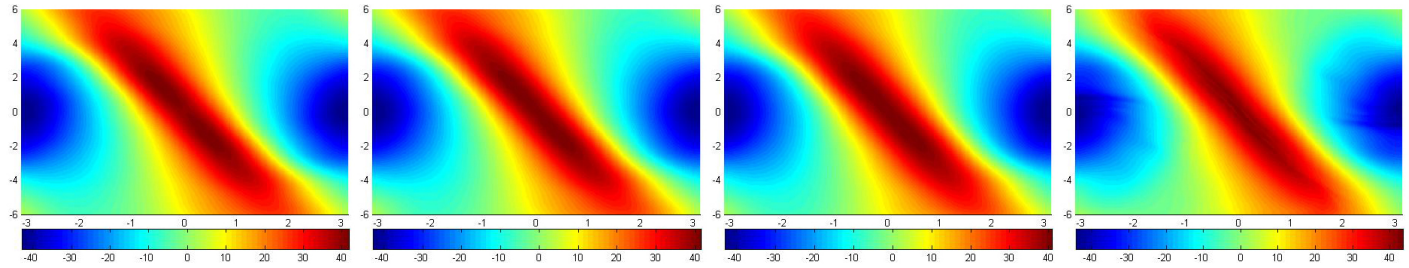
Table 1: Details about the (Partially) Model-Free Off-policy PI methods (API, QPI, EPI, and CPI)

Name	Policy Evaluation	Policy Improvement	Constraint(s)	Search Space	Fnc(s) to be Estimated
API	$\mathbb{E}_\mu^{(x,u)}\big[D_\gamma(v_i)\big] = \mathbb{E}_\mu^{(x,u)}\big[\mathcal{I}_\gamma\big(R - a_i + a_i^{\pi_i}\big)\big]$	$\pi_{i+1}(x) \in \arg\max_{u\in\mathcal{U}} a_i(x,u)$	$a_i^{\pi_i}(x) = 0$	$\mathcal{X}\times\mathcal{U}$	v_π and a_π
QPI	$\mathbb{E}_\mu^{(x,u)}\big[D_\beta\big(q_i^{\pi_i}\big)\big] = \kappa\cdot\mathbb{E}_\mu^{(x,u)}\big[\mathcal{I}_\beta\left(R - q_i\right)\big]$	$\pi_{i+1}(x) \in \arg\max_{u\in\mathcal{U}} q_i(x,u)$	(none)	$\mathcal{X}\times\mathcal{U}$	q_π
EPI	$\mathbb{E}_\mu^x\big[D_\gamma(v_i)\big] = \mathbb{E}_\mu^x\big[\mathcal{I}_\gamma\big(R^{\pi_i} - \nabla v_i\left(f_c - f_c^{\pi_i}\right)\big)\big]$	exactly same as that in Alg. 1	(none)	$\mathcal{X}$	v_π
CPI	$\mathbb{E}_\mu^{(x,u)}\big[D_\gamma\big(v_i\big)\big] = \mathbb{E}_\mu^{(x,u)}\big[\mathcal{I}_\gamma\big(R^{\pi_i} - c_i\,\xi^{\pi_i}\big)\big]$	$\pi_{i+1}(x) = \sigma(c_i(x))$	(C1) & (C2)	$\mathcal{X}\times\{u_j\}$	v_π and c_π

5 Simulation Examples: Applications to a Swing-up Pendulum Task

We simulated the proposed PI methods with the 2nd-order inverted-pendulum model ($n = 2$ and $m = 1$): $\dot{\theta}_\tau = w_\tau$ and $J\dot{w}_\tau = -\varrho\, w_\tau + mgl\sin\theta_\tau + U_\tau$, where θ_τ and w_τ are state variables representing the angular position and velocity of the pendulum at time $\tau \geq 0$, and U_τ is the external torque input to the pendulum at time τ limited as $|U_\tau| \leq U_{\max}$ for $U_{\max} = 5$ [N·m]. The physical parameters were set to $\varrho = 0.01$, $m = l = 1$, $g = 9.81$, and $J = ml^2 = 1$. The state and action spaces in this example are $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{U} = [-U_{\max}, U_{\max}]$; the state vector is $X_\tau \doteq [\,\theta_\tau \;\; w_\tau\,]^\mathsf{T} \in \mathcal{X}$; f in the dynamics is given by $f(x,u) = f_d(x) + F_c(x)u$ with $f_d(x) = [\,x_2 \quad (mgl\,\sin x_1 - \varrho x_2)/J\,]^\mathsf{T}$ and $F_c(x) = [\,0 \;\; 1/J\,]^\mathsf{T}$ and thus satisfies (C1), where $x = [\,x_1 \;\; x_2\,]^\mathsf{T} \in \mathcal{X}$. Note that this inverted pendulum setting is exactly same to that in [2]; since the maximum torque $U_{\max}$ is smaller than mgl, the policy has to swing the pendulum several times to reach the upright position.

Our learning objective in the simulation is to make the pendulum swing up and eventually settle down at the upright position $\theta_{\mathsf{final}} = 2\pi k$ for some integer k. The reward R to achieve such a goal under the torque limit $|U_\tau| \leq U_{\max}$ was set to $R(x,u) = R_0(x) - S(u)$ with $R_0(x) = 10^2\cos x_1$ and $S(u) = \lim_{v\to u}\int_0^v \sigma^{-1}(\xi)\, d\xi$, where $\sigma(\xi) = U_{\max}\tanh(\xi/U_{\max})$ is a sigmoid function that saturates at $\pm U_{\max}$. Notice that our setting satisfies (C1) and (C2) both of which are necessary to simulate CPI with its simple policy improvement $\pi_{i+1}(x) = \sigma(c_i(x))$; both also simplify the policy improvement of PI and EPI



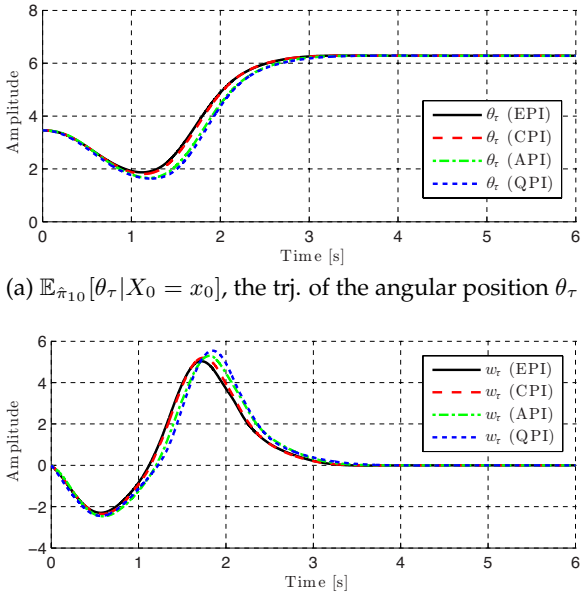
(a) $\hat{v}_{10}(x)$ (EPI) (b) $\hat{v}_{10}(x)$ (CPI) (c) $\hat{v}_{10}(x)$ (API) (d) $\hat{q}_{10}(x, \hat{\pi}_{10}(x))/\kappa$ (QPI)

Figure 1: The estimates of $v_i|_{i=10}$ ($\approx v_*$) done by the off-policy PI methods over the region Ω_{x}; the horizontal and vertical axes correspond to the values of the angular position x_1 and the velocity x_2 of the pendulum; $\hat{v}_i$, $\hat{q}_i$, and $\hat{\pi}_i$ denote the estimates of v_i, q_i, and π_i obtained by running each method. Note that $v_i = q_i(\cdot, \pi_i(\cdot))/\kappa$ in QPI.

as $\pi_{i+1}(x) = \sigma(F_c^{\mathsf{T}}(x)\nabla v_i^{\mathsf{T}}(x))$, but not of API and QPI at all. By integration by parts and $\tanh^{-1}(u/U_{\mathsf{max}}) = \frac{1}{2}\ln(u_+/u_-)$, where $u_\pm \doteq 1 \pm u/U_{\mathsf{max}}$, the action penalty $S(u)$ is explicitly expressed as $S(u) = (U_{\mathsf{max}}^2/2) \cdot \ln\left(u_+^{u_+} \cdot u_-^{u_-}\right)$ which is finite over $\mathcal{U}$ and has its minimum ($= 0$) at $u = 0$ and its maximum (≈ 17.3287) at $u = \pm U_{\mathsf{max}}$. This establishes the boundedness of the reward R over $\mathcal{X} \times \mathcal{U}$ and thereby, admissibility and boundedness of v_π for any policy π (see Section 2).

The initial policy π_0 and the parameters in all of the methods were set to $\pi_0 = 0$, $\gamma = 0.1$, $\Delta t = 10$ [ms], and in QPI, $\beta = 1$. The behavior policy μ used in the off-policy simulations was $\mu = 0$ (EPI) and $\mu(t, x, u) = u$ (API, QPI, and CPI). In API and QPI, the next target policy π_{i+1} is given by $\pi_{i+1}(x) \approx \sigma(y_i(x))$, where $y_i(x)$ is the output of a radial basis function network (RBFN) to be trained by policy improvement using a_i and q_i, respectively. The functions v_i, a_i, q_i, and c_i were all approximated by RBFNs as well. Instead of the whole spaces $\mathcal{X}$ and $\mathcal{X} \times \mathcal{U}$, we considered their compact regions $\Omega_{\mathsf{x}} \doteq [-\pi, \pi] \times [-6, 6]$ and $\Omega_{\mathsf{x}} \times \mathcal{U}$; since our inverted-pendulum system and the VF are 2π-periodic in the angular position x_1, the state value $x \in \mathcal{X}$ was normalized to $\bar{x} \in [-\pi, \pi] \times \mathbb{R}$ whenever input to the RBFNs. Further details about the RBFNs and the implementation methods are all omitted for brevity; the result of PI (Algorithm 1) is also omitted since it is almost exactly same to that of EPI.

Fig. 1 shows the estimated values of v_{π_i} over Ω_{x} at $i = 10$. Here, v_{π_i} can be considered to be approximately equal to the optimal one v_* after convergence. As shown in Fig. 1, the landscapes of the final VF estimates generated by different PI methods are all consistent and approximately equal to each other. The same goes for the state trajectories in Fig. 2 generated under the estimated policy $\hat{\pi}_i$ of π_i finally obtained at $i = 10$ by each off-policy method. They also all achieved the learning objective with $\theta_{\mathsf{final}} = 2\pi$ at around $t = 3$ [s]. Note that in our case, the initial policy $\pi_0 = 0$ was not asymptotically stabilizing while it should be in the PI methods *under the optimal control setting (without discounting)* to achieve such learning objective [3, 4].



(a) $\mathbb{E}_{\hat{\pi}_{10}}[\theta_\tau | X_0 = x_0]$, the trj. of the angular position θ_τ

(b) $\mathbb{E}_{\hat{\pi}_{10}}[w_\tau | X_0 = x_0]$, the trj. of the angular velocity w_τ

Figure 2: State trjs. under $x_0 = [\,1.1\pi \;\; 0\,]^{\mathsf{T}}$ and $\hat{\pi}_{10}$.

References

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

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

[3] J. Y. Lee, J. B. Park, and Y. H. Choi. 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.

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

[5] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. Second Edition in Progress, MIT Press, Cambridge, MA, 2017.