The Clash of Reinforcement Learning and Control Theory:
Policy Iterations for Reinforcement Learning in Continuous Time and Space

Jae Young Lee and Richard S. Sutton

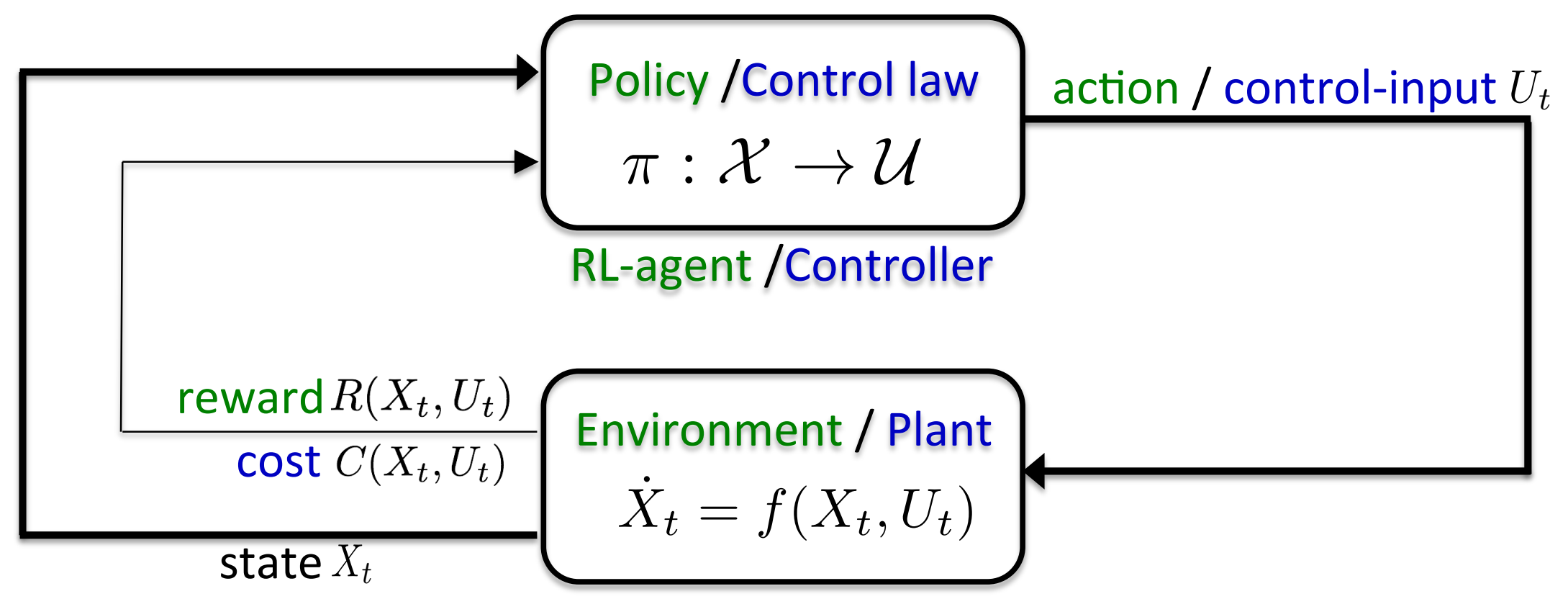
Reinforcement Learning and Artificial Intelligence Lab., Computing Science, University of Alberta, Edmonton, AB, Canada.

Abstract

In decision-making, reinforcement learning (RL) and control theory are two main disciplines in which PI methods have been studied for solving optimal decision/control problems, respectively. Motivated by PI for optimal control in continuous time and space (CTS), this work extends the reinforcement learning (RL) framework to CTS and then proposes the corresponding PI with fundamental theory. Together with the PIs in both disciplines, the theoretical approach and frameworks behind them will be extensively compared and discussed in this interactive talk.

Common Backgrounds

- **Continuous time and space (CTS) settings:**
  1. state space $\mathcal{X} \doteq \mathbb{R}^n$;
  2. action space $\mathcal{U} \subseteq \mathbb{R}^m$ is an m-dim. manifold with boundary;
  3. time space $\mathbb{R}_+ = [0, \infty)$.
- **Continuous feedback interaction components**


Policy /Control law
$\pi : \mathcal{X} \to \mathcal{U}$
RL-agent /Controller
action / control-input U_t
reward $R(X_t, U_t)$
cost $C(X_t, U_t)$
Environment / Plant
$\dot{X}_t = f(X_t, U_t)$
state X_t

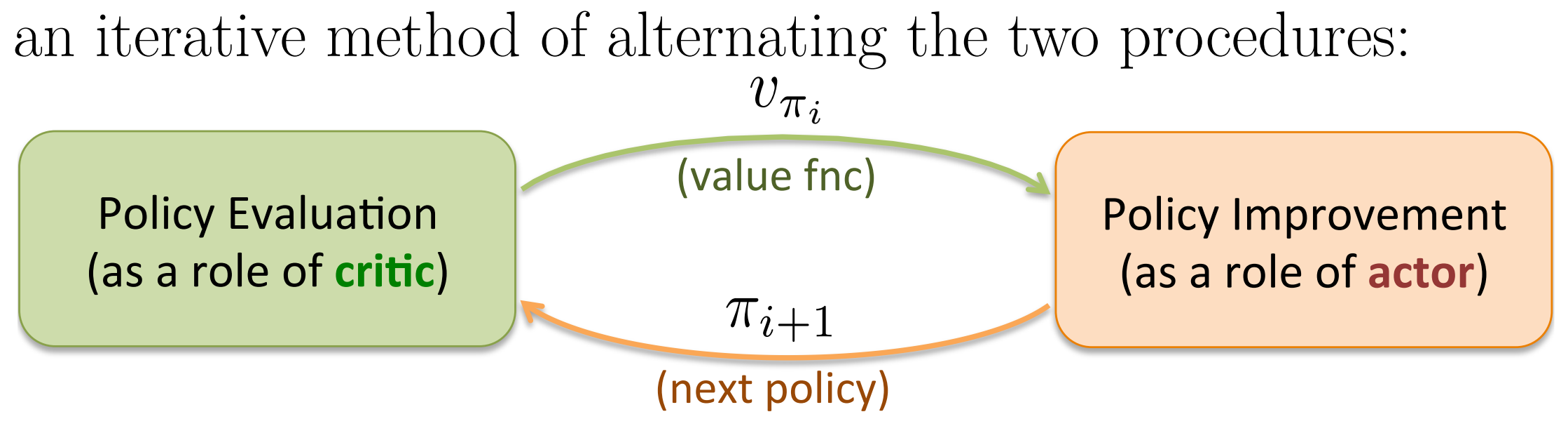
- **Real physical world** is modeled in CTS as

$$\dot{X}_t = f(X_t, U_t) + \text{(noise terms)}$$

The PIs discussed in this talk from both disciplines ignore the noise terms and consider f expressed as

$$f(X_t, U_t) = \underbrace{f_{\mathsf{c}}(X_t, U_t)}_{known} + \underbrace{f_{\mathsf{d}}(X_t)}_{unknown} .$$

- **Policy iteration (PI),** a fundamental principle in RL, is an iterative method of alternating the two procedures:


v_{π_i}
(value fnc)
Policy Evaluation
(as a role of critic)
Policy Improvement
(as a role of actor)
π_{i+1}
(next policy)

Similar concept and algorithms exist in the control theory discipline from 1960s, but their connection to PI is not fully recognized until 2000s. On the other hand, in RL discipline,

State/Action	Discrete		Continuous	
Time	Discrete	Continuous	Discrete	Continuous
Study on PI	Well-developed PI methods and theory			??

- In RL discipline, **value function** (VF) $v_\pi : \mathcal{X} \to \overline{\mathbb{R}}$ is defined as

$$v_\pi(x) \doteq \mathbb{E}_\pi\left[\int_t^\infty \gamma^{\tau - t} R(X_\tau, U_\tau)\, d\tau \middle| X_t = x\right]$$

for $\begin{cases} \text{a discount factor } \gamma \in (0,1); \\ \text{an upper-bounded reward fnc } R : \mathcal{X} \times \mathcal{U} \to \mathbb{R}. \end{cases}$

Here, v_π is always upper-bounded ($\because$ so is R).

- In control discipline, they consider the VF with

"$\gamma = 1$ *and a negative definite reward fnc R,*"

which is strongly connected to **stability theory**.

- **Definition (admissible policy):** a policy π is said to be *admissible* if $v_\pi(x)$ is finite $\forall x \in \mathcal{X}$.

- The **goal** of PI in both RL and optimal control disciplines is to find a best admissible policy π_* (and the corresponding optimal VF v_*) such that

$$v_\pi(x) \leq v_*(x) \;\; \forall x \in \mathcal{X}.$$

- **Problem Settings** in discrete versus continuous domains

	Discrete & Stochastic (e.g., a finite MDP)	*Continuous & Deterministic* $\dot{X}_t = f(X_t, U_t)$
Spaces	$\mathcal{S}$, $\mathcal{A}$, t	$\mathcal{X} = \mathbb{R}^n$, $\mathcal{U} \subseteq \mathbb{R}^m$, $t \in [0, \infty)$
Reward R	$\mathcal{R}^a_{ss'}$ (bounded)	$R(x, u)$ (upper-bounded)
Optimal sol. (v_* and π_*)	well-defined (always)	well-defined (assumed)
State trj. ($\mathbb{E}_\pi[X_\tau]$)	well-defined (always)	well-defined (by stability theory or assumed)
Others	N/A	continuity/regularity of fncs

Brief Review of Stability Theory

- In stability framework, it is assumed that there is a stationary point $X_* = 0 \in \mathcal{X}$ for $U_* = 0 \in \mathcal{U}$ s.t.

$$X_* = f(X_*, U_*).$$

Any non-zero (X_*, U_*) can be transformed to $(0, 0)$.

- **The objective** is to stabilize $X_* = 0$ in the sense that:
  1. a small initial state perturbation from X_* at any time t gives a small perturbation of $X_{t+\tau}$ for all $\tau \geq 0$ from X_*;
  2. $\lim_{\tau \to \infty} X_{t+\tau} = X_*$ for all such initial state $X_t \in \mathcal{X}$.

  Any policy π achieving it is said to be **stabilizing**.

- Closely related to PI in control field, but not in the RL discipline, is **(Lyapunov's) stability theorem**:
  - If there exist two negative definite fncs $v, w : \mathcal{X} \to \mathbb{R}$ s.t. $\nabla v(x) f(x, \pi(x)) \geq -w(x)\ \forall x \in \mathcal{X}$, then π is stabilizing.

VFs in Control and RL Disciplines

- In control discipline,

every admissible policy π is stabilizing

by Lyapunov's stability theorem with $v = v_\pi$ and $w = R(\cdot, \pi(\cdot))$. Here, stability (and thus admissibility) ensures the existence of the bounded unique state trj. X_τ for all future time τ.

- **The discounted VF in RL discipline is not related to stability** since there is no bridge between them. *Our RL problem in CTS is more general but beyond the current stability theory*, which forces us to assume the existence of the unique state trj. X_τ for all future time τ.

- **Bellman equation:** for any given admissible π,

$$\mathbb{E}_\pi\big[\, \delta_t(v_\pi) \,\big|X_t = x\big] = 0 \text{ for all } x \in \mathcal{X},$$

where the Bellman error $\delta_t(v_\pi)$ is given by

$$\underbrace{\int_t^{t'} \gamma^{\tau - t} R(X_\tau, U_\tau) d\tau}_{\text{accumulated reward}} + \gamma^{\Delta t} \cdot \underbrace{v_\pi(X_{t'})}_{\text{next value}} - \underbrace{v_\pi(X_t)}_{\text{current value}}$$

($\Delta t \doteq t' - t$: difference b.t.w the time steps).

- $v = v_\pi$ if v satisfies the Bellman equation and

$$\lim_{\tau \to \infty} \gamma^{\tau - t} \cdot \mathbb{E}_\pi\big[v(X_{t+\tau})\big|X_t = x\big] = 0 \;\; \forall x \in \mathcal{X},$$

which is true in optimal control when $v(X_*) = 0$ by stability; for this in RL problem, $v(X_{t+\tau})$ should not grow in time with the rate higher than $1/\gamma$.

Policy Improvement & Optimality

- **Policy improvement** in CTS is to find an improved policy π' satisfying

$$\pi'(x) \in \arg\max_{U_t \in \mathcal{U}} \lim_{\Delta t \to 0} \mathbb{E}\big[\, \delta_t(v_\pi)/\Delta t \big| X_t = x\big],$$

which is assumed to exist and can be expressed as

$$\pi'(x) \in \arg\max_{u \in \mathcal{U}} \big(R(x, u) + \nabla v_\pi(x) f_{\mathsf{c}}(x, u)\big).$$

- Policy improvement in both disciplines is **partially model-free** and yields the improved policy π'—if π is admissible, then so is π' and

"$v_\pi(x) \leq v_{\pi'}(x)$ for all $x \in \mathcal{X}$."

In the case $\gamma = 1$ w/o stability theory, there needs the condition $\mathbb{E}_{\pi'}[v_\pi(X_{t+\tau})|X_t = x] \leq 0$ in the limit $\tau \to \infty$, which is true in stability framework.

- By principle of optimality, v_* in RL and control disciplines satisfies the optimality equation known as **Hamilton-Jacobi-Bellman equation**:

$$0 = \max_{U_t \in \mathcal{U}} \lim_{\Delta t \to 0} \mathbb{E}\big[\, \delta_t(v_*)/\Delta t \,\big|X_t = x\big] \text{ for all } x \in \mathcal{X}$$

(it is assumed that v_* is the unique HJB solution).

Policy Iteration for RL in CTS

```
Initialize: $i \leftarrow 0$, $\Delta t > 0$, and an admissible policy $\pi_0$;
repeat
    Evaluation: find $v_i$ s.t. $\mathbb{E}_{\pi_i}\big[\delta_t(v_i)\big|X_t = x\big] = 0\ \forall x \in \mathcal{X}$;
    Improvement: find a next policy $\pi_{i+1}$ s.t.
        $\pi_{i+1}(x) \in \arg\max_{u \in \mathcal{U}} \Big(R(x, u) + \nabla v_i(x) f_{\mathsf{c}}(x, u)\Big)\ \forall x \in \mathcal{X}$;
    $i \leftarrow i + 1$;
until convergence is met.
```

Assumption 1. $\forall i \in \mathbb{N} \cup \{0\}$: if π_i is admissible,

$$\lim_{\tau \to \infty} \gamma^{\tau - t} \cdot \mathbb{E}_{\pi_i}\Big[\, v_i(X_{t+\tau}) \,\Big| X_t = x\Big] = 0 \;\;\; \forall x \in \mathcal{X}.$$

Theorem 1. The sequences $\{\pi_i\}$ and $\{v_i\}$ generated by PI under the above assumption satisfy the followings:

1. $\forall i \in \{0, 1, 2, \cdots\}$: $\begin{cases} \pi_{i+1} \text{ is admissible and } v_i = v_{\pi_i}; \\ v_{\pi_i}(x) \leq v_{\pi_{i+1}}(x) \leq v_*(x)\ \forall x \in \mathcal{X}; \end{cases}$
2. v_i converges to v_*
   - w.r.t. a metric on the space of admissible VFs;
   - pointwisely on $\mathcal{X}$ and uniformly on any compact $\Omega \subset \mathcal{X}$ (under certain continuity & admissibility conditions).

PI for Optimal Control in CTS

- PI and its properties in control field is same as that in the RL discipline shown above, except that
  1. (Pros) **Assumption 1** is true (and thus not required);
  2. (Pros) X_τ is uniquely defined for all future time τ;
  3. (Pros) all generated policies π_{i+1} are stabilizing;
  4. (Cons) the initial policy π_0 is required to be stabilizing;
  5. (Cons) by stability-based approach, γ must be equal to 1, and there are restrictions on f and R.

Concluding Remarks

- PI in optimal control is not suitable for RL;
- stability theory cannot cover the *discounted* cases;
- we establish PI for RL problems, theoretical RL backgrounds, in CTS w/o employing stability thm;
- there is a gap b.t.w. stability and RL frameworks.