Lambda-Policy Iteration with Randomization for Contractive Models with Infinite Policies: Well-Posedness and Convergence*

Yuchao Li, Karl H. Johansson, and Jonas Mårtensson
Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden

*This was supported by the Swedish Foundation for Strategic Research, the Swedish Research Council, and the Knut and Alice Wallenberg Foundation.

Abstract

Abstract dynamic programming (DP) models are used to analyze λ-policy iteration with randomization (λ-PIR) algorithms. Particularly, contractive models with infinite policies are considered and it is shown that well-posedness of the λ-operator plays a central role in the algorithm. In addition, we identify the conditions required to guarantee convergence with probability one when the policy space is infinite. Guided by the analysis, we exemplify a data-driven approximated implementation of the algorithm for estimation of optimal costs of constrained control problems, where promising numerical results are found.

Motivations

λ-PIR, proposed in [1], belongs to the broad class of policy iteration (PI) methods. In particular, it brings to bears the rich results for implementations due to its close connections to

- **TD(λ):** temporal difference (TD) learning ideas;
- **Proximal algorithm:** prominent methods in convex optimization [2];
- **Value iteration:** a principle method for DP.

However, no analysis is given for problems with infinite states and/or infinite policies.

Problems

- **Well-posedness:**

  Is the λ-PIR well-posed for problems with infinite states and policies?

- **Convergence:**

  Given the λ-PIR is well-posed, will it converge to the optimal?

Preliminaries

Given state space X, control space U, and policy space $\mathcal{M} = \{\mu \,|\, \mu(x) \in U(x), \forall x \in X\}$, we study the mappings of the form $H : X \times U \times \mathcal{R}(X) \to \mathbb{R}$, and the ones

$$(T_\mu J)(x) = H(x, \mu(x), J),$$

$$(TJ)(x) = \inf_{\mu \in \mathcal{M}} (T_\mu J)(x).$$

Principle properties are:

- **Uniform contraction:**

  For some $\alpha \in (0,1)$, $\forall J, J' \in \mathcal{B}(X)$, $\mu \in \mathcal{M}$, it holds that

  $$\|T_\mu J - T_\mu J'\| \leq \alpha \|J - J'\|.$$

- **Monotonicity:**

  $\forall J, J' \in \mathcal{B}(X)$, it holds that $J \leq J'$ implies $\forall x \in X$, $u \in U(x)$,

  $$H(x,u,J) \leq H(x,u,J').$$

Main Results

The operator, named as λ-operator, is

$$(T_\mu^{(\lambda)} J)(x) = (1-\lambda) \sum_{\ell=1}^{\infty} \lambda^{\ell-1} (T_\mu^\ell J)(x). \quad (1)$$

Given $J_k \in \mathcal{B}(X)$ and $p_k \in (0,1)$, λ-PIR computes the policy μ^k and cost approximate J_{k+1} as

$$T_{\mu^k} J_k = T J_k;\; J_{k+1} = \begin{cases} T_{\mu^k} J_k, & p_k, \\ T_{\mu^k}^{(\lambda)} J_k, & \text{o.w.} \end{cases} \quad (2)$$

1 Well-posedness

Theorem 1 *Let the set of mappings $T_\mu : \mathcal{B}(X) \to \mathcal{B}(X)$, $\mu \in \mathcal{M}$, satisfy the contraction property. Consider the mappings $T_\mu^{(w)}$ defined point-wise as*

$$(T_\mu^{(w)} J)(x) = \sum_{\ell=1}^{\infty} w_\ell(x) (T_\mu^\ell J)(x),\; x \in X, \quad (3)$$

with $w_\ell(x) \geq 0$ and $\sum_{\ell=1}^{\infty} w_\ell(x) = 1$. Then the range of $T_\mu^{(w)}$ is a subset of $\mathcal{B}(X)$, viz., $T_\mu^{(w)} : \mathcal{B}(X) \to \mathcal{B}(X)$; and $T_\mu^{(w)}$ is a contraction.

2 Convergence

Theorem 2 *Let relevant assumptions hold. Given $J_0 \in \mathcal{B}(X)$ such that $TJ_0 \leq J_0$, the sequence $\{J_k\}_{k=0}^\infty$ generated by algorithm (2) converges in norm to J^* with probability one.*

Corollary 2.1 *Let $H(\cdot,\cdot,\cdot)$ have the form*

$$H(x,u,J) = \int_X \big(g(x,u,y) + \alpha J(y)\big) d\mathbb{P}(y|x,u) \quad (4)$$

where $g : X \times U \times X \to \mathbb{R}$, $\alpha \in (0,1)$ and $\mathbb{P}(\cdot|x,u)$ is the probability measure conditioned on (x,u) for certain MDP. Let $v(x) = 1\ \forall x \in X$, and relevant assumptions hold. Given arbitrary $J_0 \in \mathcal{B}(X)$, the sequence $\{J_k\}_{k=0}^\infty$ generated by algorithm (2) converges in norm to J^* with probability one.

Numerical Example

Consider a torsional pendulum system:

$$\dot{\phi} = \omega,\; \dot{\omega} = M^{-1}(-mgl \sin\phi - \gamma\omega + \tau),$$

with state and control spaces constrained in compact sets. It is suitably discretized and the dynamics on the state boundaries are tailored to have the assumptions hold.

- The closed loop system behavior greatly improved after 5 λ-PIR iterations, see Fig. 1.

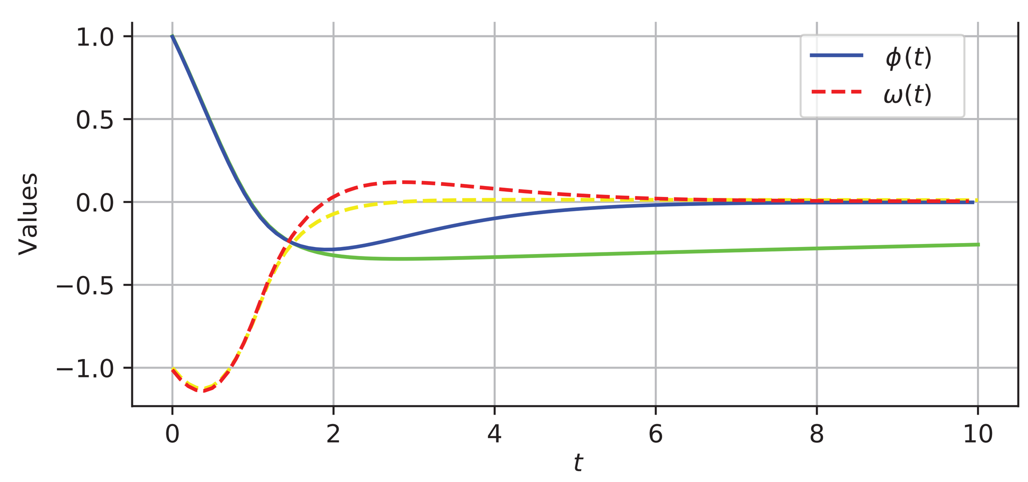

Figure 1: Closed loop system trajectory before (yellow and green) and after training (red and blue).

- The cost function converges after 5 iterations, see Figs. 2 and 3 for plots along the axes where $\omega = 0$ and $\phi = 0$.

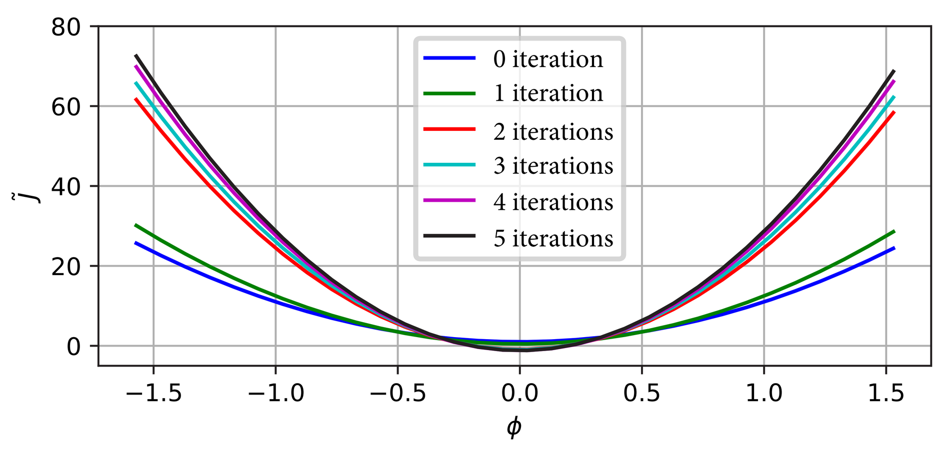

Figure 2: Cost function along the axis $\omega = 0$ after different training iterations.

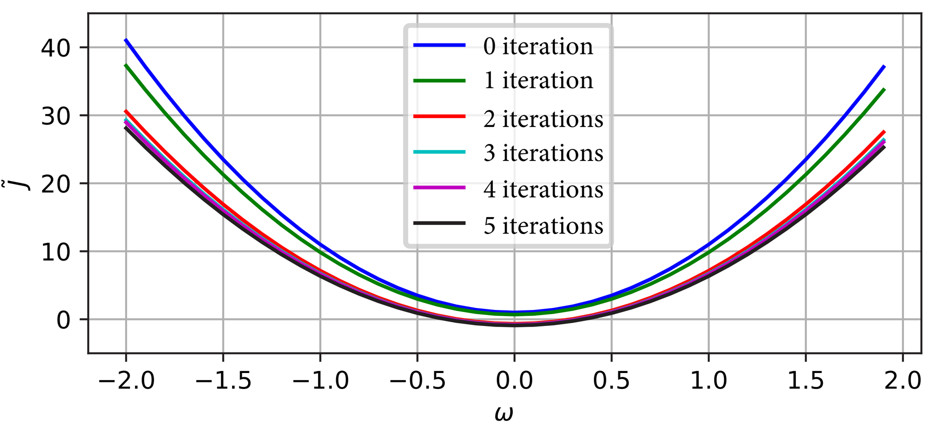

Figure 3: Cost function along the axis $\phi = 0$ after different training iterations.

- λ-PIR shows faster convergence against VI; and requires less computational efforts to obtain training samples for the cost function when compared with OPI [3], see Figs. 4 and 5.

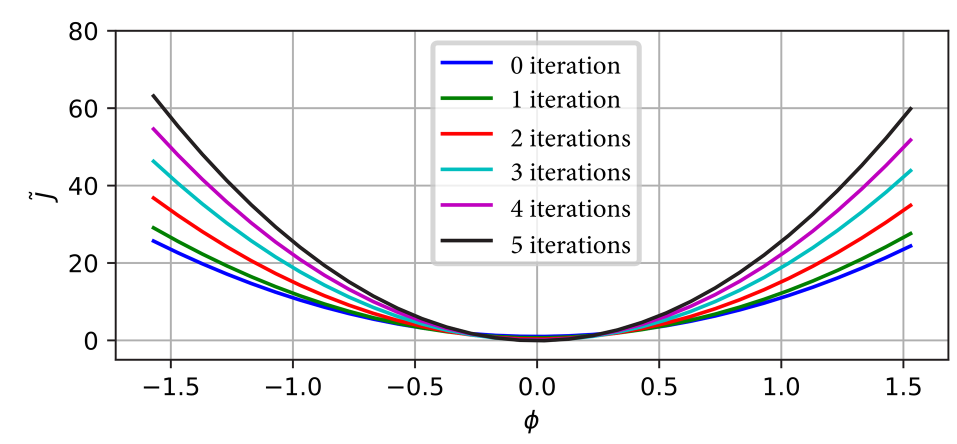

Figure 4: Cost functions of VI along the axis $\omega = 0$.

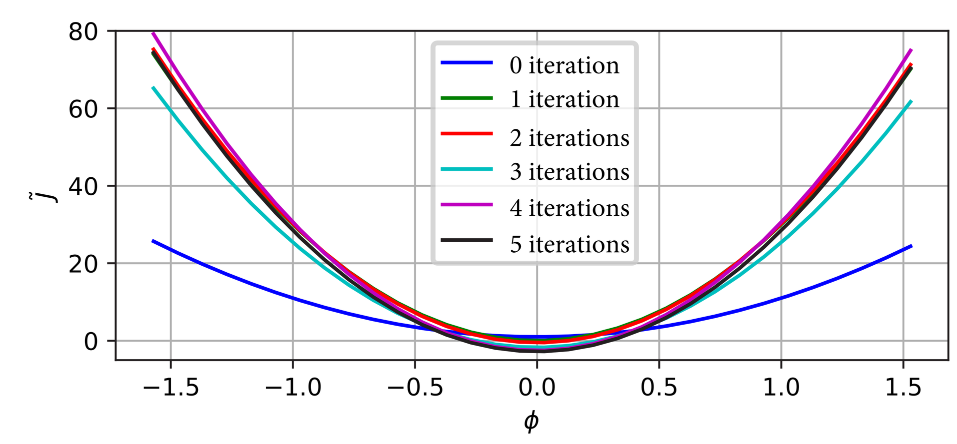

Figure 5: Cost functions of OPI along the axis $\omega = 0$.

References

[1] D. P. Bertsekas. *Abstract dynamic programming*. Athena Scientific, 2nd edition, 2018.

[2] D. P. Bertsekas. Proximal algorithms and temporal difference methods for solving fixed point problems. *Computational Optimization and Applications*, 70(3):709–736, 2018.

[3] B. Scherrer, *et al.* Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, 16:1629–1676, 2015.