Regularized Anderson Acceleration for Off-Policy Deep Reinforcement Learning

Wenjie Shi, Shiji Song, Hui Wu, Ya-Chu Hsu, Cheng Wu, Gao Huang
Tsinghua University, Beijing, China

MOTIVATION

Sample inefficiency

Existing methods

- To learn the model of system dynamics.
- Applying off-policy training scheme to reuse past experience.

Our method

- RL is closely linked to fixed-point iteration: the optimal policy can be found by solving a fixed-point problem of Bellman operator.
- Anderson acceleration is a method capable of speeding up the computation of fixed point iterations.

METHOD

Anderson Acceleration for Policy Iteration

For a policy iteration, suppose that in addition to the current estimate Q^{π_k}, the $m-1$ previous estimates $Q^{\pi_{k-1}}$,..., $Q^{\pi_{k-m+1}}$ are also known. Then, a linear combination of estimates Q^{π_i} with coefficients α_i reads

$$Q_\alpha^k=\sum_{i=1}^m \alpha_i Q^{\pi_{k-m+i}} \text{ with } \sum_{i=1}^m \alpha_i=1.$$

We define combined Bellman operator $\mathcal{T}_c$ as follows

$$\mathcal{T}_c Q_\alpha^k=\sum_{i=1}^m \alpha_i \mathcal{T} Q^{\pi_{k-m+i}}.$$

Then, one searches a vector α^k that minimizes the objective function defined as the combined Bellman residuals among the entire state-action space $\mathcal{S}\times\mathcal{T}$,

$$\alpha^k=\underset{\alpha\in\mathbb{R}^m}{\arg\min}\; J(\alpha)=\underset{\alpha\in\mathbb{R}^m}{\arg\min}\left\|\sum_{i=1}^m \alpha_i(\mathcal{T}Q^{\pi_{k-m+i}}-Q^{\pi_{k-m+i}})\right\|, \quad \text{s.t.} \sum_{i=1}^m \alpha_i=1.$$

Regularized variant with function approximation

Challenges

- Sweeping entire state-action space is intractable.
- Function approximation errors are unavoidable.
- The solution may suffer from ill-conditioning when the squared Bellman residuals matrix is rank-deficient.

Adding a regularization term to the objective function,

$$\tilde\alpha^k=\underset{\alpha\in\mathbb{R}^m}{\arg\min}\left\|\sum_{i=1}^m \alpha_i(\mathcal{T}Q^{\pi_{k-m+i}}-Q^{\pi_{k-m+i}}+e_{k-m+i})\right\|+\lambda\|\alpha\|^2, \quad \text{s.t.} \sum_{i=1}^m \alpha_i=1.$$

where e_{k-m+i} represents the perturbation induced by function approximation errors, The solution can be obtained analytically

$$\tilde\alpha^k=\frac{(\widetilde\Delta_k^T\widetilde\Delta_k+\lambda I)^{-1}\mathbf{1}}{\mathbf{1}^T(\widetilde\Delta_k^T\widetilde\Delta_k+\lambda I)^{-1}\mathbf{1}}, \quad \begin{aligned}\widetilde\Delta_k&=[\widetilde\delta_{k-m+1},...,\widetilde\delta_k]\in\mathbb{R}^{N_A\times m}\\ \widetilde\delta_i&=\mathcal{T}Q^{\pi_i}-Q^{\pi_i}+e_i\in\mathbb{R}^{N_A}\end{aligned}$$

Proposition 1: *Consider two identical PIs $\mathcal{I}_1$ and $\mathcal{I}_2$ with function approximation. $\mathcal{I}_2$ is implemented with RAA and takes into account approximation errors, whereas $\mathcal{I}_1$ is only implemented with vanilla Anderson acceleration. Let α^k and $\tilde\alpha^k$ be the coefficient vectors of $\mathcal{I}_1$ and $\mathcal{I}_2$ respectively. Then, we have the following bounds*

$$\|\tilde\alpha^k\|\le\sqrt{\frac{\lambda+\|\widetilde\Delta_k\|^2}{m\lambda}}, \qquad \|\tilde\alpha^k-\alpha^k\|\le\frac{\|\widetilde\Delta_k^T\widetilde\Delta_k-\Delta_k^T\Delta_k\|+\lambda}{\lambda}\|\alpha^k\|.$$

Regularized Anderson acceleration for actor-critic

Under the paradigm of off-policy deep RL, RAA variant of policy iteration degrades into the Bellman equation

RAA-Dueling-DQN:

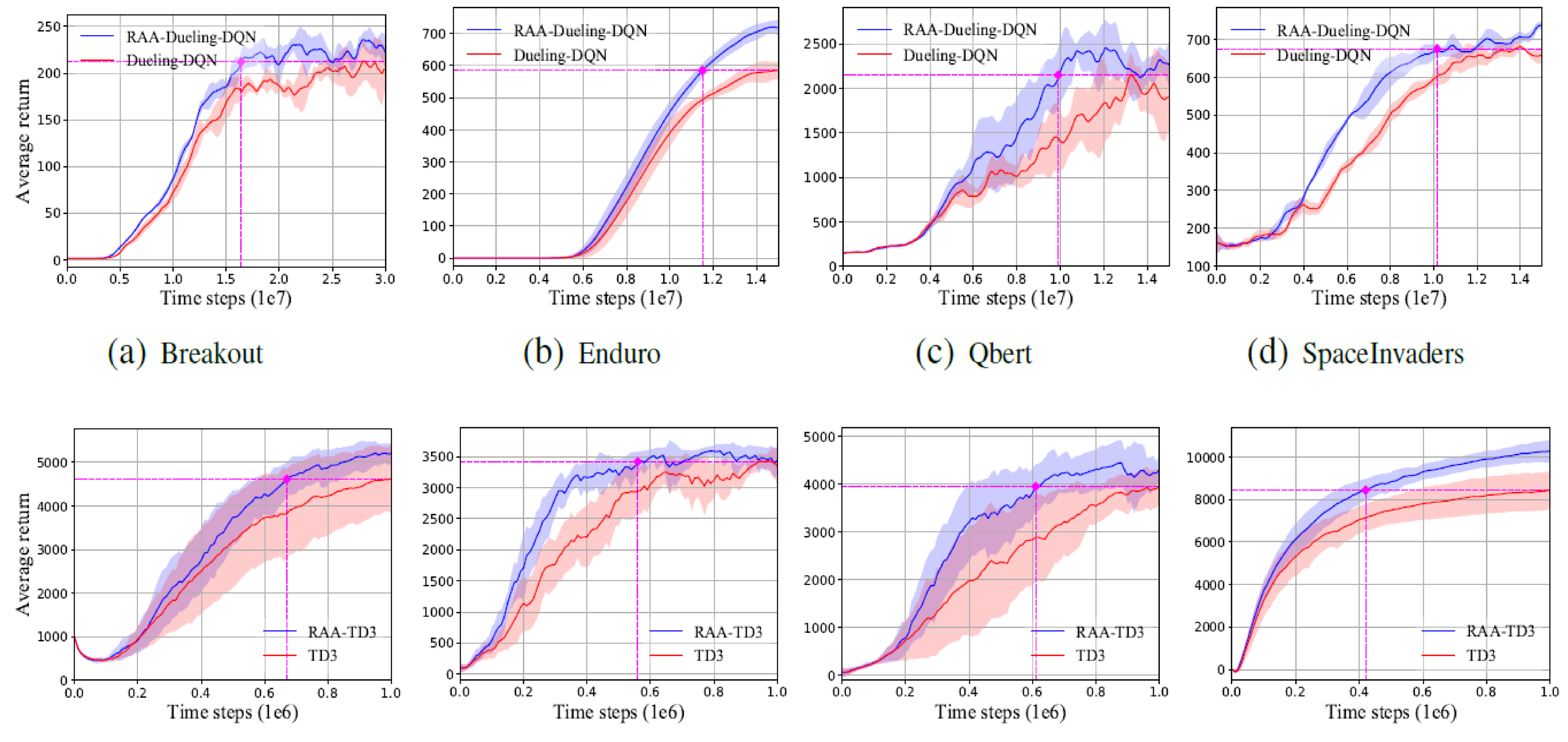
$$Q_\theta(s_t,a_t)=\mathbb{E}_{s_{t+1},r_t}\left[r_t+\gamma\sum_{i=1}^m\tilde\alpha_i\max_{a_{t+1}}Q_{\theta^i}(s_{t+1},a_{t+1})\right],$$

RAA-TD3:

$$Q_\theta(s_t,a_t)=\mathbb{E}_{s_{t+1},r_t}\left[r_t+\gamma\sum_{i=1}^m\tilde\alpha_i\hat Q_{\theta^i}(s_{t+1},\pi_{\phi'}(s_{t+1})+\epsilon)\right], \quad \hat Q_{\theta^i}(s_t,a_t)=\min_{j=1,2}Q_{\theta^i_j}(s_t,a_t).$$

EXPERIMENTS

Comparative evaluation

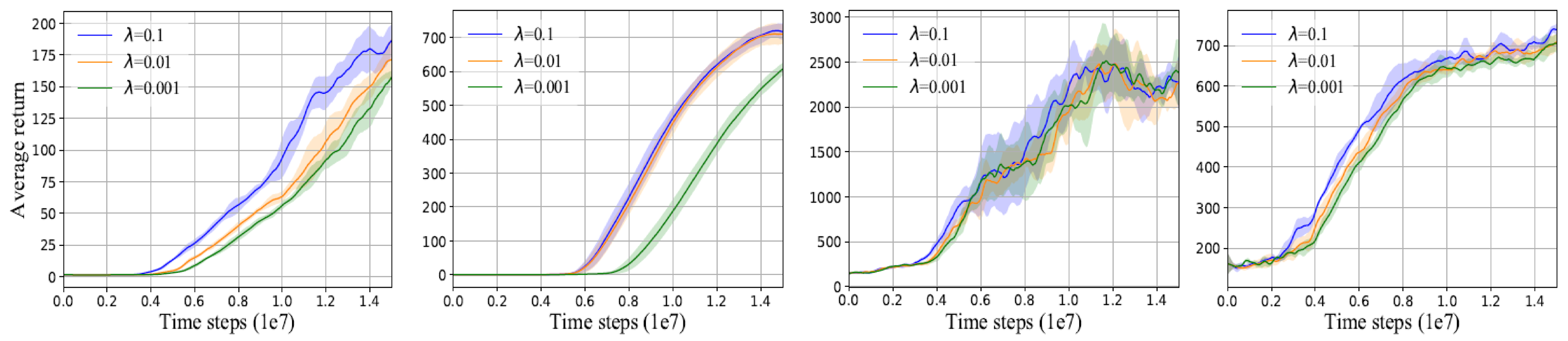


(a) Breakout (b) Enduro (c) Qbert (d) SpaceInvaders

(e) Ant-v2 (f) Hopper-v2 (g) Walker2d-v2 (h) HalfCheetah-v2

Figure 1: Learning Curves of Dueling-DQN, TD3 and their RAA variants on discrete and continuous control tasks.

Ablation studies



(a) Breakout (b) Enduro (c) Qbert (d) SpaceInvaders

Figure 2: Sensitivity of RAA-Dueling-DQN to the scaling of regularization on discrete control tasks.

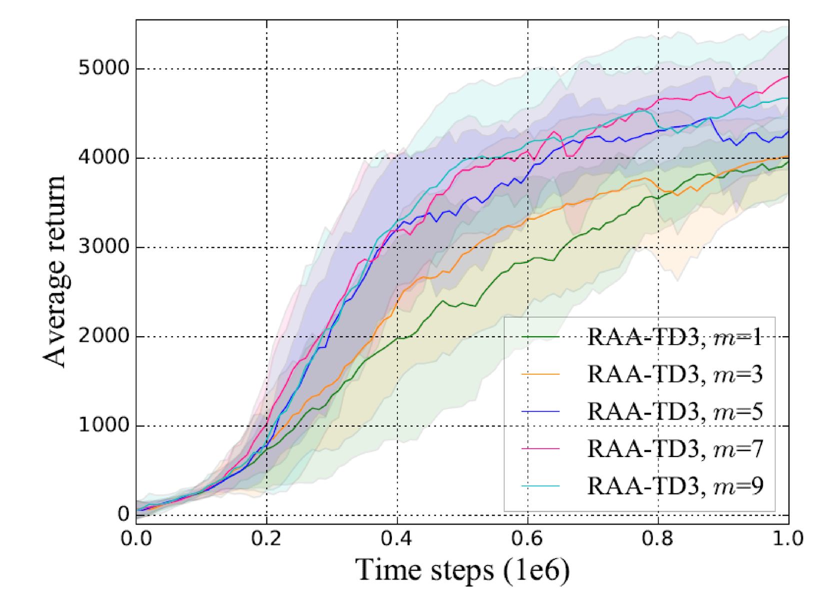


Figure 3: Learning Curves of RAA-TD3 on Walker2d-v2 with different m.

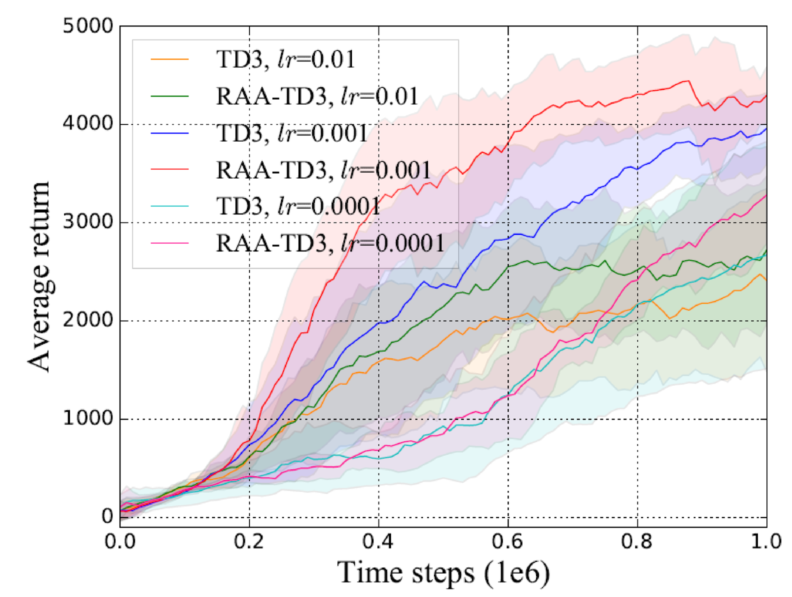


Figure 4: Performance comparison on Walker2d-v2 with different learning rates.

Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019)