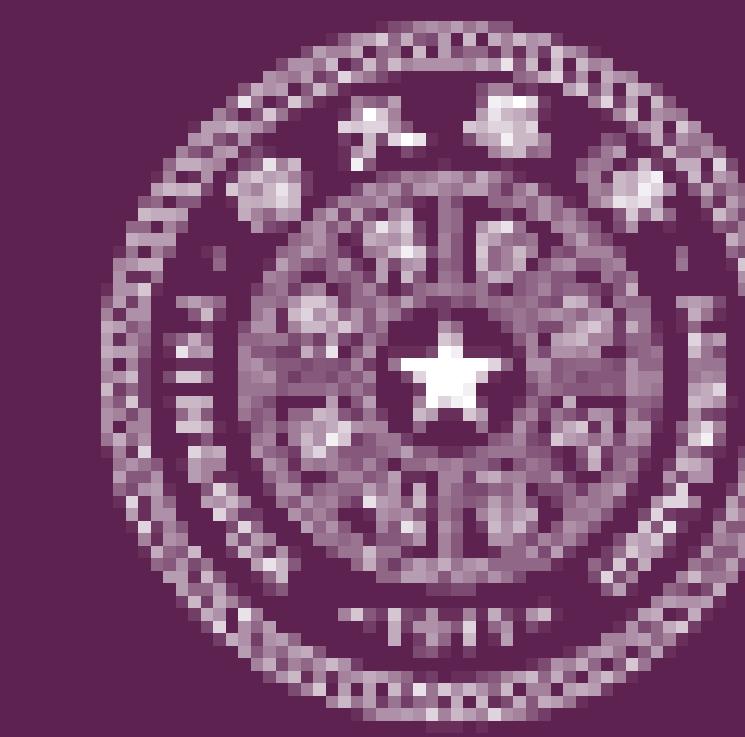


Regularized Anderson Acceleration for Off-Policy Deep Reinforcement Learning

Wenjie Shi, Shiji Song, Hui Wu, Ya-Chu Hsu, Cheng Wu, Gao Huang
Department of Automation, Tsinghua University, Beijing 100084, China



清华大学
Tsinghua University

MOTIVATION

- Sample inefficiency remains a major limitation of current RL algorithms for problems with continuous and high dimensional state spaces.
- Sample inefficiency makes learning in real physical systems impractical and severely prohibits the applicability of RL approaches in more challenging scenarios.

Existing methods

- To learn models of the underlying system dynamics.
- Off-policy training scheme aims to reuse past experience such as DQN, DDPG.

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

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

Due to this equality constraint, 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 coefficient vector α^k that minimizes the following objective function defined as the combined Bellman residuals among the entire state-action space $\mathcal{S} \times \mathcal{T}$,

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

Regularized variant with function approximation Challenges

- The Bellman residuals are calculated among the entire state-action space, but sweeping entire state-action space is intractable for continuous RL.
- Function approximation errors are unavoidable and lead to biased solution of Anderson acceleration.
- the solution may suffer from ill-conditioning when the squared Bellman residuals matrix is rank-deficient.

The coefficient vector is now adjusted to $\tilde{\alpha}^k$ that minimizes the perturbed objective function added with a Tikhonov regularization term,

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

where e_{k-m+i} represents the perturbation induced by function approximation errors. The solution to this regularized optimization problem can be obtained analytically

$$\tilde{\alpha}^k = \frac{(\tilde{\Delta}_k^T \tilde{\Delta}_k + \lambda I)^{-1} \mathbf{1}}{\mathbf{1}^T (\tilde{\Delta}_k^T \tilde{\Delta}_k + \lambda I)^{-1} \mathbf{1}}, \quad \tilde{\Delta}_k = [\tilde{\delta}_{k-m+1}, \dots, \tilde{\delta}_k] \in \mathbb{R}^{N_A \times m} \\ \tilde{\delta}_i = \mathcal{T} Q^{\pi_i} - Q^{\pi_i} + e_i \in \mathbb{R}^{N_A}$$

where λ is a positive scalar representing the scale of regularization.

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

$$\|\tilde{\alpha}^k\| \leq \sqrt{\frac{\lambda + \|\tilde{\Delta}_k\|^2}{m\lambda}}, \quad \|\tilde{\alpha}^k - \alpha^k\| \leq \frac{\|\tilde{\Delta}_k^T \tilde{\Delta}_k - \Delta_k^T \Delta_k\| + \lambda}{\lambda} \|\alpha^k\|.$$

Regularized Anderson acceleration for actor-critic

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

$$\text{RAA-Dueling-DQN} \quad Q_\theta(s_t, a_t) = \mathbb{E}_{s_{t+1}, r_t} \left[r_t + \gamma \sum_{i=1}^m \tilde{\alpha}_i \max_{a_{t+1}} Q_{\theta^i}(s_{t+1}, a_{t+1}) \right],$$

$$\text{RAA-TD3} \quad Q_\theta(s_t, a_t) = \mathbb{E}_{s_{t+1}, r_t} \left[r_t + \gamma \sum_{i=1}^m \tilde{\alpha}_i \hat{Q}_{\theta^i}(s_{t+1}, \pi_{\phi^i}(s_{t+1}) + \epsilon) \right], \quad \hat{Q}_{\theta^i}(s_t, a_t) = \min_{j=1,2} Q_{\theta_j^i}(s_t, a_t).$$

where θ^i is the parameters of target critic network.

EXPERIMENTS

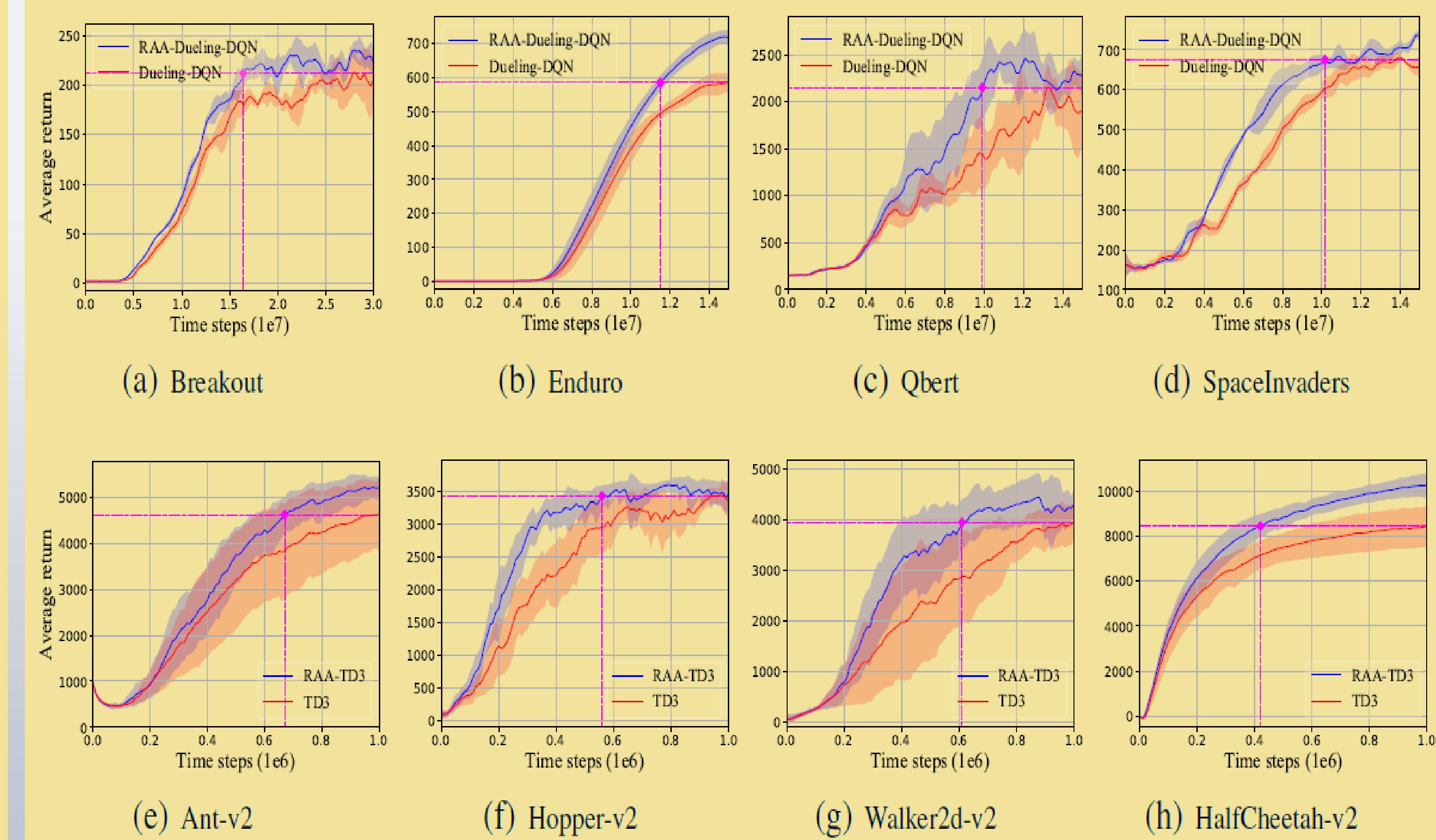


Figure 1. Learning Curves of Dueling-DQN, TD3 and their RAA variants on discrete and continuous control tasks.

