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
Email: shiwj16@mails.tsinghua.edu.cn

NEURAL INFORMATION PROCESSING SYSTEMS

Tsinghua University

MOTIVATION

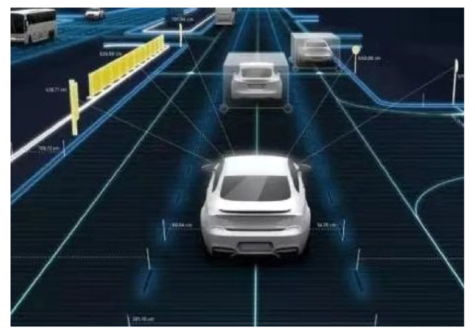
Sample inefficiency

AlphaGo Zero

millions of self-play

StarCraft

several days



Autonomous Driving

safety-critical scenarios

- Millions of trials are required to learn.
- Leading to long training time.
- Making the learning in real physical systems impractical.

Existing methods

- To learn the model of system dynamics.
- To reuse past experience (off-policy).

Main observations

- RL is closely linked to fixed-point iteration.
- Anderson acceleration is a method capable of speeding up the computation of fixed point iterations.

METHOD

Fixed-point problem:
Given g: $\mathbb{R}^n \to \mathbb{R}^n$, solve $x = g(x)$.

RL problem:
Given the optimality Bellman operator $\mathcal{T}$: solve $Q^\pi = \mathcal{T}Q^\pi$.

Algorithm FPI. Fixed-Point Iteration:
Given x_0,
For $k = 0,1, \ldots$
Set $x_{k+1} = g(x_k)$.

Algorithm PI. Policy Iteration:
Given π_0,
For $k = 0,1, \ldots$
Set $Q^{\pi_{k+1}} = \mathcal{T}Q^{\pi_k}$.

Algorithm AA. Anderson Acceleration:
Given x_0 and $m \geq 1$,
For $k = 0,1, \ldots$
Set $F_k = (f_{k-m+1}, \ldots, f_k)$, where $f_i = g(x_i) - x_i$.
Solve $\alpha^k = (\alpha_1^k, \ldots, \alpha_m^k)^T$:
$min_\alpha \, ||F_k\alpha||_2$ s.t. $\sum \alpha_i = 1$.
Set $x_{k+1} = \sum_{i=1}^m \alpha_i^k g(x_{k-m+i})$.

Algorithm AA for PI.
Given π_0 and $m \geq 1$,
For $k = 0,1, \ldots$
Set $\Delta_k = (\delta_{k-m+1}, \ldots, \delta_k)$, where $\delta_i = \mathcal{T}Q^{\pi_i} - Q^{\pi_i}$.
Solve $\alpha^k = (\alpha_1^k, \ldots, \alpha_m^k)^T$:
$\boxed{min_\alpha \, ||\Delta_k\alpha||_2 \text{ s.t. } \sum \alpha_i = 1.}$
Set $Q^{\pi_{k+1}} = \sum_{i=1}^m \alpha_i^k \mathcal{T}Q^{\pi_{k-m+i}}$.

Challenges

- Sweeping entire state-action space is intractable.
- Function approximation errors are unavoidable.
- The solution may suffer from ill-conditioning.

Algorithm RAA. Regularized AA:

$\boxed{min_\alpha \, ||\widetilde{\Delta}_k\alpha||_2 + \lambda||\alpha||^2}$

EXPERIMENTS

Comparative evaluation

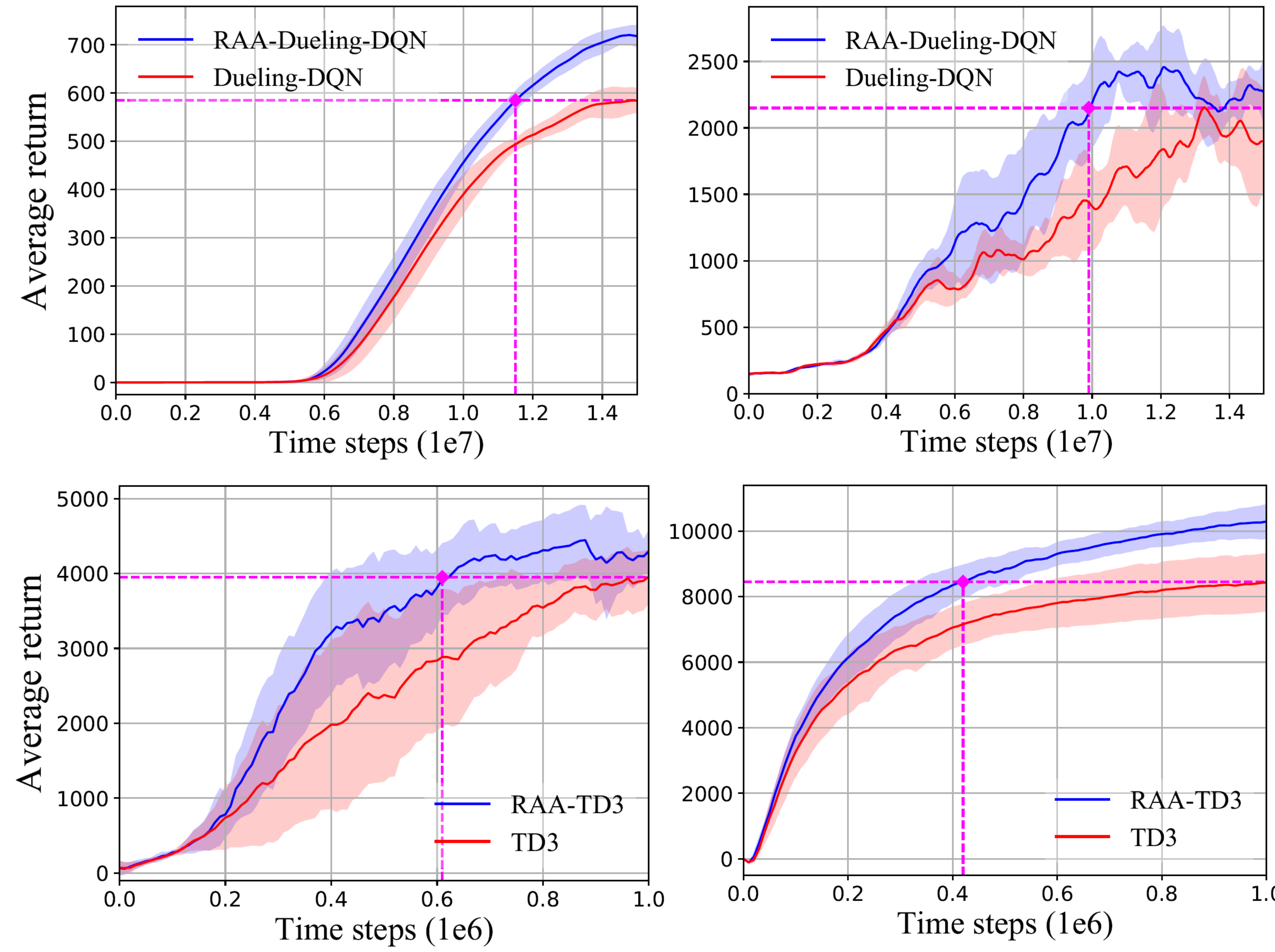
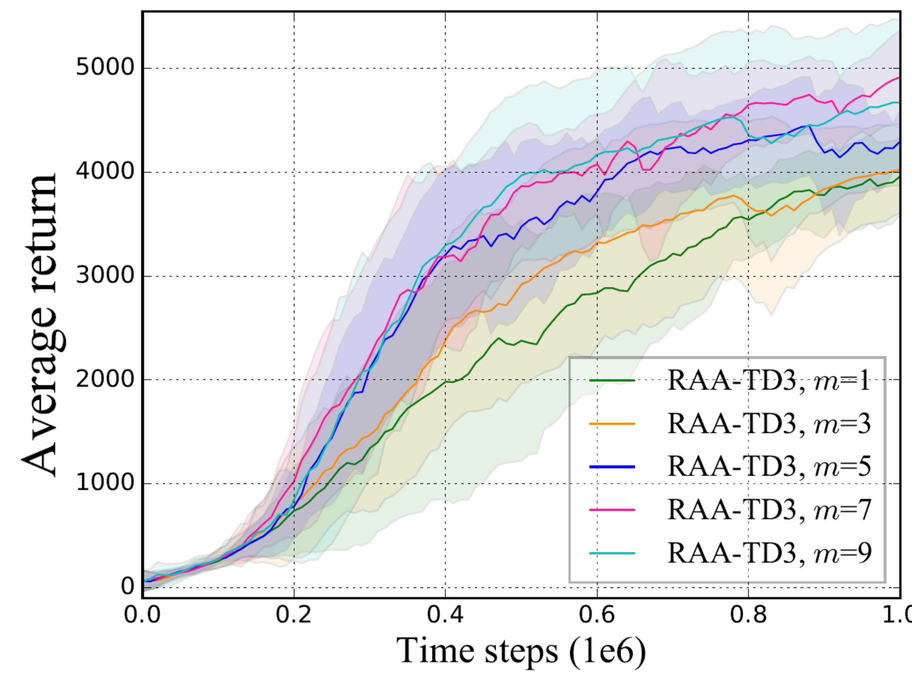


Figure 1: Learning Curves of Dueling-DQN, TD3 and their RAA variants on discrete and continuous control tasks.

Ablation studies



Larger m leads to faster convergence and better final performance.

Code

Arxiv