The Effect of Multi-step Methods on Overestimation in Deep Reinforcement Learning

Lingheng Meng
Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada

Rob Gorbet
Knowledge Integration, University of Waterloo, Waterloo, Canada

Dana Kulić
Monash University, Melbourne, Australia

Presented by: Lingheng Meng, Adaptive Systems Lab

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Motivation

- Multi-step (also called *n*-step) methods in Reinforcement Learning (RL), with tabular representation of the value-function, have been shown to be more efficient than the 1-step method due to faster propagation of the reward signal.
- Research in Deep Reinforcement Learning (DRL), with value-function and policy approximated by deep neural networks, shows that multi-step methods improve learning speed and final performance.
- However, there is a lack of understanding about what is contributing to the boost of performance of multi-step methods in DRL.

Background

Overestimation Problem

Assume Q^{true} is represented by a function approximator Q^{approx} with noise $E(s', a')$:

$$Q^{approx}(s', a') = Q^{true}(s', a') + E(s', a')$$

Then, for Q-Learning technique

$$Q^{approx}(s, a) \leftarrow r(s, a) + \max_{a'} Q^{approx}(s', a')$$

zero-mean noise may easily result in overestimation problem because

$$\max_{a'} Q^{approx}(s', a') > \max_{a'} Q^{true}(s', a')$$

S. Thrun, and A. Schwartz. "Issues in using function approximation for reinforcement learning." 1993.

Background

Overestimation Problem

E.g., if

$$Q^{true}(s', a') = 0 \quad \text{and} \quad \mathbb{E}\left[E(s', a')\right] = 0$$

then

$$\begin{aligned}\max_{a'} Q^{approx}(s', a') &= \max_{a'} \left[Q^{true}(s', a') + E(s', a')\right] \\ &= \max_{a'} \left[0 + E(s', a')\right] > 0\end{aligned}$$

while

$$\max_{a'} Q^{true}(s', a') = 0$$

Background

Deep Deterministic Policy Gradient (DDPG)

Critic, i.e. Q-value, is optimized by minimizing

$$L_{\theta^Q} = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim U(D)} \left[\left(\hat{Q}_t - Q_{\theta^Q}(s_t, a_t) \right)^2 \right]$$

where $\hat{Q}_t = r_t + \gamma \max_{a \leftarrow \mu_{\theta^{\mu-}}(s_{t+n})} Q_{\theta^{Q-}}(s_{t+1}, a)$, $Q_{\theta^{Q-}}$ is target critic, and $\mu_{\theta^{\mu-}}$ is target actor representing the optimal policy.

Actor, i.e. policy, is optimized by maximizing

$$J_{\theta^\mu} = \mathbb{E}_{s_t \sim U(D)} \left[Q_{\theta^Q}(s_t, \mu_{\theta^\mu}(s_t)) \right]$$

where Q_{θ^Q} and μ_{θ^μ} are online critic and actor, respectively.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, et al. "Continuous control with deep reinforcement learning." 2015.

Proposed Methods

Multi-step Deep Deterministic Policy Gradient (MDDPG)

Bootstrapped target Q is based on multi-step immediate rewards

$$\hat{Q}_t^{(n)} = \begin{cases} \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \max_a Q_{\theta^{Q^-}}(s_{t+n}, a), & if\ \forall\ k \in [1, \cdots, n]\ \ and\ d_{t+k} \neq 1; \\ \sum_{i=0}^{k-1} \gamma^i r_{t+i}, & if\ \exists\ k \in [1, \cdots, n]\ \ and\ d_{t+k} = 1. \end{cases}$$

where n indicates n immediate rewards are used.

Then, Q is optimized by minimizing

$$L_{\theta^Q} = \mathbb{E}_{(s_t, a_t, r_t, \cdots, s_{t+n}, d_{t+n}) \sim U(D)} \left[\left(\hat{Q}_t^{(n)} - Q_{\theta^Q}(s_t, a_t) \right)^2 \right]$$

Mixed Multi-step DDPG (MMDDPG)

- An average over target Q-values with different step sizes from 1 to *n*

$$\hat{Q}_t^{(n_{avg})} = \frac{1}{n}\sum_{i=1}^{n}\hat{Q}_t^{(i)}$$

- The minimum of a set of target Q-values

$$\hat{Q}_t^{(n_{min})} = \min_{i\sim[1,n]}\hat{Q}_t^{(i)}$$

- An average over target Q-values with step size from 2 to n, considering n= 1 is the most prone to overestimation:

$$\hat{Q}_t^{(n_{avg-1})} = \frac{1}{n-1}\sum_{i=2}^{n}\hat{Q}_t^{(i)}$$

Experiment Results

Experimental Evidence of Multi-step Methods' Effect on Alleviating Overestimation

- Almost all MDDPG(n) with n >1 outperform DDPG
- Bad performance of DDPG corresponds to an extremely overestimated Q-value

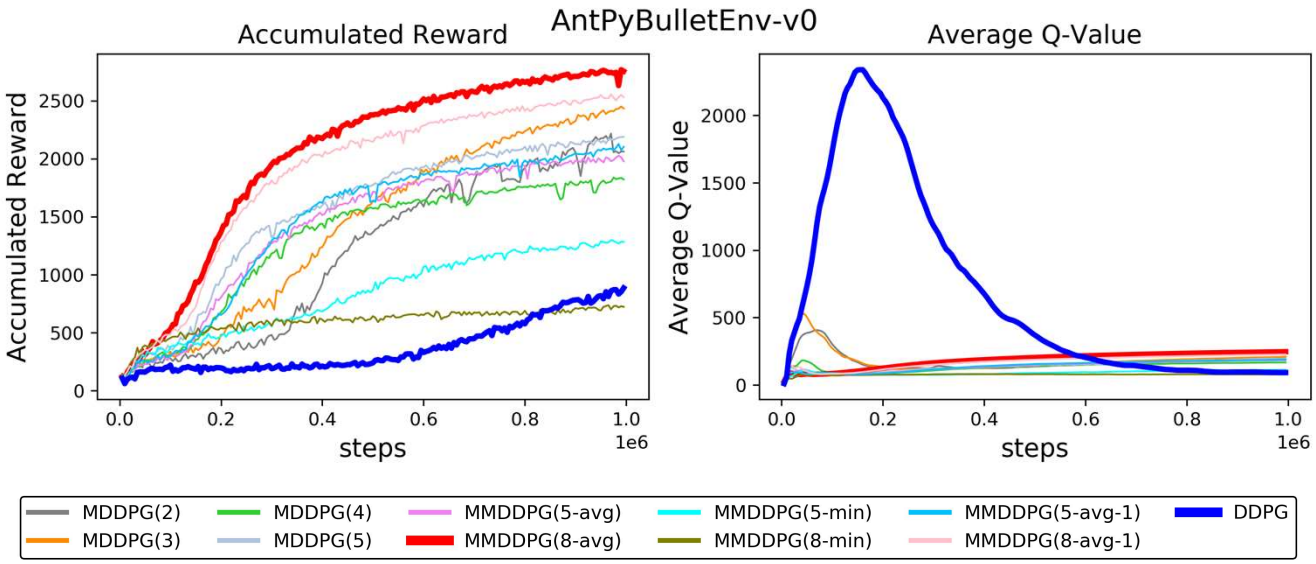

Fig. 1 Comparison among MDDPG, MMDDGP and DDPG on AntPyBulletEnv-v0

Experiment Results

Experimental Evidence of Multi-step Methods' Effect on Alleviating Overestimation

- All positive gaps means multi-step methods provide smaller estimated target Q-values than that of the 1-step method.
- The larger the step, the smaller the corresponding estimated target Q-value.
- The difference becomes smaller with increased interactions.
- The magnitude of the estimated Q-value decreases as the step size *n* increases.

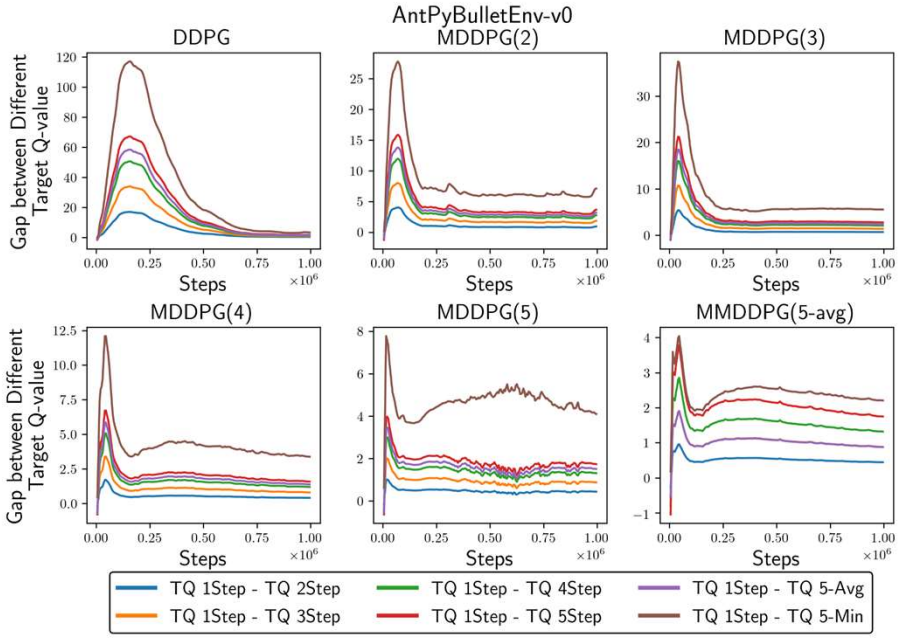

Fig. 2 The Difference in Estimated Target Q-values Between 1-step and Multi-step Methods, where the larger the value, the bigger the difference.

Meng, Lingheng, Rob Gorbet, and Dana Kulić

UNIVERSITY OF WATERLOO

FACULTY OF ENGINEERING

Questions? Comments?

lingheng.meng@uwaterloo.ca