

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

Lingheng Meng

Electrical and Computer Engineering
University of Waterloo
Waterloo, Canada
lingheng.meng@uwaterloo.ca

Rob Gorbet

Knowledge Integration
University of Waterloo
Waterloo, Canada
rbgorbet@uwaterloo.ca

Dana Kulić

Monash University
Melbourne, Australia
dana.kulic@monash.edu

Abstract—Multi-step (also called n -step) methods in reinforcement learning (RL) have been shown to be more efficient than the 1-step method due to faster propagation of the reward signal, both theoretically and empirically, in tasks exploiting tabular representation of the value-function. Recently, research in Deep Reinforcement Learning (DRL) also shows that multi-step methods improve learning speed and final performance in applications where the value-function and policy are represented with deep neural networks. However, there is a lack of understanding about what is actually contributing to the boost of performance. In this work, we analyze the effect of multi-step methods on alleviating the overestimation problem in DRL, where multi-step experiences are sampled from a replay buffer. Specifically building on top of Deep Deterministic Policy Gradient (DDPG), we propose Multi-step DDPG (MDDPG), where different step sizes are manually set, and its variant called Mixed Multi-step DDPG (MMDDPG) where an average over different multi-step backups is used as update target of Q-value function. Empirically, we show that both MDDPG and MMDDPG are significantly less affected by the overestimation problem than DDPG with 1-step backup, which consequently results in better final performance and learning speed. We also discuss the advantages and disadvantages of different ways to do multi-step expansion in order to reduce approximation error, and expose the tradeoff between overestimation and underestimation that underlies offline multi-step methods. Finally, we compare the computational resource needs of Twin Delayed Deep Deterministic Policy Gradient (TD3), a state-of-art algorithm proposed to address overestimation in actor-critic methods, and our proposed methods, since they show comparable final performance and learning speed.

I. INTRODUCTION

Deep Reinforcement Learning (DRL) [1] incorporates the powerful representation capacity of nonlinear Deep Neural Networks (DNNs) into classic Reinforcement Learning (RL), but also results in low data efficiency and instability, i.e., drastic fluctuations in accumulated reward increase rather than a smooth increase [2], [3], [4]. Deep Q-Networks (DQNs) [1] represent the Q-value function with DNNs, but is only designed for discrete action spaces, while Deep Deterministic Policy Gradient (DDPG) [5] was proposed to tackle continuous control tasks, where the actor and critic are both represented with DNNs. The move from low dimensional state and action spaces to high-dimensional state and action spaces by employing DNNs comes with the cost of low data efficiency and non-monotonic learning progress, where policy diverges from optimal due to inaccurate approximation of the

value function. These disadvantages hinder DRL from broad use in applications where interactive data collection is time-consuming and smooth adaption of behavior is crucial to maintain engagement, e.g. interactive robots [6].

Low data efficiency corresponds to slow learning speed, assuming the learning algorithm is capable of learning an optimal policy given sufficient data, and can be caused by two reasons: 1) lack of data, and 2) lack of training. If slow learning is caused by lack of data, the environment is under-explored. In this case, an efficient exploration strategy, e.g. parameter space noise [7], or a complementary source of experiences, e.g. World Models [8], can be helpful for generating additional training data. On the other hand, if slow learning is caused by lack of training, since enough experiences have been collected but not coded into policy, a more effective way to use the collected data is necessary such as prioritized replay buffer [9], [10] and hindsight experience replay [11].

Instability is partially related to the catastrophic forgetting problem of Deep Learning (DL) [12] which is inherent in the continuously evolving nature of policy learning in Reinforcement Learning (RL). In addition, inaccurate estimation and continuous tuning of the Q-value function might lead the learned policy in directions far away from optimal or cause it to fluctuate around a local optimum. The overestimation problem [13] in Q-learning is a typical example of inaccurate estimation in which the maximization of an inaccurate Q-value estimate induces a consistent overestimation. As the result of overestimation, the estimated Q-value of a given (state, action) pair might explode and drive the corresponding policy away from optimal. Therefore, DRL algorithms based on DNNs should strive to alleviate overestimation problem if it cannot be completely overcome.

Building on top of DDPG [5], we experiment with Multi-step DDPG (MDDPG), where different step sizes are manually set, and with a variant called Mixed Multi-step DDPG (MMDDPG) where a mixture of different multi-step backups is used as target Q-value. In this paper, we refer backup target Q-values with more than 1 immediate reward as multi-step backups. We first experimentally show that MDDPG and MMDDPG outperform DDPG, in terms of final performance and learning speed, mainly because of their effect helps alle-

viate the overestimation problem. Then, we compare MMD-DPG with other state-of-the-art approaches to show that the proposed method can achieve comparable performance to TD3 [14] which is a state-of-art DRL algorithm and dedicated to addressing overestimation problem. After that, we discuss the underestimation and overestimation underlying offline multi-step method. At the end, we conclude this work and provide prospects for future research.

II. RELATED WORK

Multi-step methods have been studied in traditional RL for both on- and off-policy learning considering both the forward view, i.e., updating each state by looking forward to future rewards and states, and the backward view, i.e., updating each state by combining the current Temporal Difference (TD) error with eligibility traces of past events [15]. Recently, a new multi-step action-value algorithm $Q(\sigma)$ was proposed to allow the degree of sampling performed by the algorithm at each step during its backup to be continuously varied, with Sarsa at one extreme, and Expected Sarsa at the other [16]. The results show that an intermediate value of σ performs better than either extreme. However, a systematic way to adjust σ still needs to be studied, and the learning tasks in [16] are relatively simple with small state and action spaces, avoiding the need for DL methods. Multi-step TD learning for non-linear function approximation was studied in [17] where forward TD(λ) was investigated on simple discrete control tasks and λ is a hyper-parameter controlling the weight of different multi-step update targets. Although a neural network was used for function approximation in this work, only simple discrete control tasks were examined. Rainbow [18], an integrated learning agent combining many extensions including multi-step learning, found that multi-step not only helps speed up early learning but also improves final performance on Atari 2600 games. However, Rainbow is built on top of DQN and only discrete action space tasks were examined. In our work, we focus on continuous control tasks.

Multi-step methods have also been investigated in asynchronous methods [19], which rely on parallel actors employing different exploration policies in parallel instances of environments, to decorrelate consecutive updating experiences and to stabilize policy learning without using a replay buffer. However, such a parallel paradigm can only work practically in simulated environments, and is unfeasible for real applications where multiple instances of physical systems (e.g., robots) are too expensive.

The overestimation problem [13] is cited as the reason non-linear function approximation fails in RL. Based on Double Q-learning [20], Double DQN [21] was shown to be effective in alleviating this problem for discrete action spaces. Although TD3 [14] proposed to take the minimum of two bootstrapped Q-values of a state-action pair which are separately estimated by two critics to overcome the overestimation problem, the extra neural network for the second critic also introduces additional computation cost, especially when the state and action spaces are large.

Model-based Value expansion (MVE) [22] is a multi-step method that expands multi-step on a learned environment model, whose performance tends to degrade in complex environments. As an improvement of MVE, STochastic Ensemble Value Expansion (STEVE) [23] expands various multi-steps on an ensemble of learned environment models, including transition dynamics and reward function, then adds these to an ensemble of Q-value functions. The final target Q-value is a weighted mean of these generated target Q-values. Both MVE and STEVE suffer from modeling error, and more importantly they both introduce vast extra computation, especially STEVE.

Averaged-DQN [24] is proposed to reduce variance and stabilize learning by exploiting an average over a set of target Q-values calculated from a set of past Q-value functions. It is shown that Averaged-DQN also helps in alleviating the overestimation problem. Different from that, our method MMDDPG takes the average over a set of target Q-values calculated with different step sizes, which is also shown to be more stable than MDDPG.

III. BACKGROUND

A. Reinforcement Learning

An RL agent learns an optimal policy through interactions with the external environment. Formally, RL can be formulated as a Markov Decision Process (MDP), where optimal actions are chosen by the agent to maximize the expected reward over discrete time-steps in an environment. An MDP is defined by a tuple (S, A, p, r) with state space S , action space A , transition probability p , and reward function r . At each discrete time step t , an agent selects an action $a_t \in A$ at state $s_t \in S$ following its policy $\pi(a_t|s_t)$. As a result of the action, the agent transitions to the next state $s_{t+1} \in S$ according to transition probability $p(s_{t+1}|s_t, a_t) \in [0, 1]$ and receives a reward $r(s_t, a_t, s_{t+1}) \in [r_{min}, r_{max}]$. The goal of the agent is to learn an optimal policy π^* which maximizes the expectation of a discount accumulated reward $\mathbb{E}_{\pi^*, s_0 \sim \rho_0} [\sum_{i=0}^{\infty} \gamma^i r_i | s_0]$ where γ is the discount factor and ρ_0 is the distribution of the initial state s_0 .

Almost all RL algorithms estimate the state-value function V or the action value function Q . A V-value function of a state s under a policy π is the expected return when starting in s and following π thereafter, which can be formally defined by $V^\pi(s_t) = \mathbb{E}_\pi [\sum_{i=t}^{\infty} \gamma^{i-t} r_i | s_t]$. A Q-value function represents the expected accumulated future reward when taking action a_t in state s_t and following a policy π afterwards, and can be formally expressed as $Q^\pi(s_t, a_t) = \mathbb{E}_\pi [\sum_{i=t}^{\infty} \gamma^{i-t} r_i | s_t, a_t]$.

B. Multi-step Bootstrapping

The motivation of Multi-step methods [15] is to facilitate fast propagation of knowledge about the outcomes of selected actions, by taking into account rewards by looking n steps forward. Formally, n -step discounted accumulated reward can be expressed as

$$R_t^{(n)} = \sum_{i=0}^{n-1} \gamma^i r_{t+i}, \quad (1)$$

where when $n = 1$ only one reward is considered, while for $n > 1$ all rewards received between t and $t + n - 1$ are considered. Based on this n -step discounted accumulated reward, the corresponding Bellman optimality equation for Q^* and V^* [25] can be defined as

$$Q^*(s_t, a_t) = \mathbb{E}_{\pi^*} \left[R_t^{*(n)} + \gamma^n \max_a Q^*(s_{t+n}, a) \right], \quad (2)$$

and

$$V^*(s_t) = \mathbb{E}_{\pi^*} \left[R_t^{*(n)} + \gamma^n V^*(s_{t+n}) \right], \quad (3)$$

where $R_t^{*(n)}$ is the n -step discounted accumulated reward following optimal policy π^* .

C. Deep Deterministic Policy Gradient (DDPG)

DDPG uses an actor-critic framework, implemented by deep neural networks, where the policy update is based on the Deterministic Policy Gradient [26]. In DDPG, the critic, $Q_{\theta^Q}(s_t, a_t)$, is a Q-value function approximated by a deep neural network with weights θ^Q and is updated following Eq. 4:

$$L_{\theta^Q} = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim U(D)} \left[\left(\hat{Q}_t - Q_{\theta^Q}(s_t, a_t) \right)^2 \right], \quad (4)$$

where $\hat{Q}_t = r_t + \gamma Q_{\theta^{Q-}}(s_{t+1}, \mu_{\theta^{\mu-}}(s_{t+1}))$, $Q_{\theta^{Q-}}$ and $\mu_{\theta^{\mu-}}$ are the target critic and actor, parameterized by θ^{Q-} and $\theta^{\mu-}$ respectively, which are soft-updated following Eq. 5:

$$\theta^{Q-} \leftarrow (1 - \tau) \theta^{Q-} + \tau \theta^Q, \quad \theta^{\mu-} \leftarrow (1 - \tau) \theta^{\mu-} + \tau \theta^\mu \quad (5)$$

with $0 < \tau \ll 1$ to stabilize the learning target, and experiences (s_t, a_t, r_t, s_{t+1}) are randomly sampled from replay buffer D as indicated as a uniform distribution $U(\cdot)$. In DDPG, the actor, $\mu_{\theta^\mu}(s_t)$, is a deterministic policy represented by another deep neural network with weights θ^μ and is updated following Eq. 6:

$$\begin{aligned} \nabla_{\theta^\mu} J &\approx \mathbb{E}_{s_t \sim U(D)} [\nabla_{\theta^\mu} Q_{\theta^Q}(s_t, a)] \\ &= \mathbb{E}_{s_t \sim U(D)} [\nabla_a Q_{\theta^Q}(s_t, a) \nabla_{\theta^\mu} \mu_{\theta^\mu}(s_t)], \end{aligned} \quad (6)$$

where $a = \mu_{\theta^\mu}(s_t)$. All experiences are collected by adding exploratory noise to the deterministic policy, $a' = \mu_{\theta^\mu}(s_t) + \epsilon_a$.

IV. PROPOSED METHODS

In this section, we will progressively introduce MDDPG and MMDDPG.

A. Multi-step DDPG (MDDPG)

MDDPG is a variant of DDPG where multi-step experiences sampled from the replay buffer are used to calculate the direct accumulated reward, which is then added to the bootstrapped Q-value after these experiences.

Based on the n -step discounted accumulated reward in Eq. 1, we can easily realize n -step bootstrapped return using consecutively stored experiences in the replay buffer. Assuming that past experiences of an agent are consecutively stored in the replay buffer D and the experience at time step t is sampled into a training mini-batch, the n consecutive

experiences from t to $t+n-1$ are treated as a single multi-time step sample. Then, for each sample in the n -step mini-batch $\{(s_t, a_t, r_t, \dots, s_{t+n}, d_{t+n})^{(i)}\}_{i=1}^N$ with size N , the n -step bootstrapped estimated action value function can be defined as Eq. 7:

$$\hat{Q}_t^{(n)} = \begin{cases} \sum_{i=0}^{n-1} \gamma^i r_{t+i} + \gamma^n \max_a Q_{\theta^{Q-}}(s_{t+n}, a), & \text{if } \forall k \in [1, \dots, n] \text{ and } d_{t+k} \neq 1; \\ \sum_{i=0}^{k-1} \gamma^i r_{t+i}, & \text{if } \exists k \in [1, \dots, n] \text{ and } d_{t+k} = 1. \end{cases} \quad (7)$$

with $d_{t+k} = 1$ if the episode is done, otherwise $d_{t+k} = 0$.

The value function Q of MDDPG is updated by minimizing the objective given in Eq. 8:

$$L_{\theta^Q} = \mathbb{E}_{(s_t, a_t, r_t, \dots, s_{t+n}, d_{t+n}) \sim U(D)} \left[\left(\hat{Q}_t^{(n)} - Q_{\theta^Q}(s_t, a_t) \right)^2 \right], \quad (8)$$

where $\hat{Q}_t^{(n)}$ is Eq. 7, while the policy update remains the same as DDPG. For convenience, in this paper we will denote MDDPG with step size n as MDDPG(n). Specifically, when $n = 1$, MDDPG(1) is equivalent to DDPG.

B. Mixed Multi-step DDPG (MMDDPG)

MMDDPG is a variant of MDDPG based on the observation that for different tasks the best choice of step size n may differ. MMDDPG mixes target Q-values calculated with different step sizes. This also helps to reduce the bias of the target Q-value by mixing a small set of target Q-values. The mixture can be an average over target Q-values with different step sizes from 1 to n as $\hat{Q}_t^{(n_{avg})}$ in Eq. 9, or the minimum of such a set of target Q-values as $\hat{Q}_t^{(n_{min})}$ in Eq. 9. Or considering $n = 1$ is the most prone to overestimation, MMDDPG could take the average over target Q-values with step size from 2 to n , as $\hat{Q}_t^{(n_{avg-1})}$ in Eq. 9:

$$\begin{aligned} \hat{Q}_t^{(n_{avg})} &= \frac{1}{n} \sum_{i=1}^n \hat{Q}_t^{(i)} \quad \text{or} \quad \hat{Q}_t^{(n_{min})} = \min_{i \in [1, n]} \hat{Q}_t^{(i)} \\ \text{or} \quad \hat{Q}_t^{(n_{avg-1})} &= \frac{1}{n-1} \sum_{i=2}^n \hat{Q}_t^{(i)}. \end{aligned} \quad (9)$$

Similar to MDDPG(n), we will denote MMDDPG with different mixture methods introduced in Eq. 9 as MMDDPG(n -avg), MMDDPG(n -min), and MMDDPG(n -avg-1), respectively.

V. EXPERIMENTS

Tasks used for evaluation are all continuous control tasks from PyBulletGym¹ which provides OpenAI Gym [27] compatible environments based on open-source Bullet Physics engine rather than non-free MuJoCo engine. We compare MDDPG and MMDDPG with vanilla DDPG, TD3, SAC [28], MVE and STEVE.² All algorithms use policy and value

¹<https://github.com/benelot/pybullet-gym>

²DDPG, TD3 and SAC use implementations in <https://github.com/openai/spinningup>, and MVE and STEVE use the implementations in <https://github.com/tensorflow/models/tree/master/research/steve>

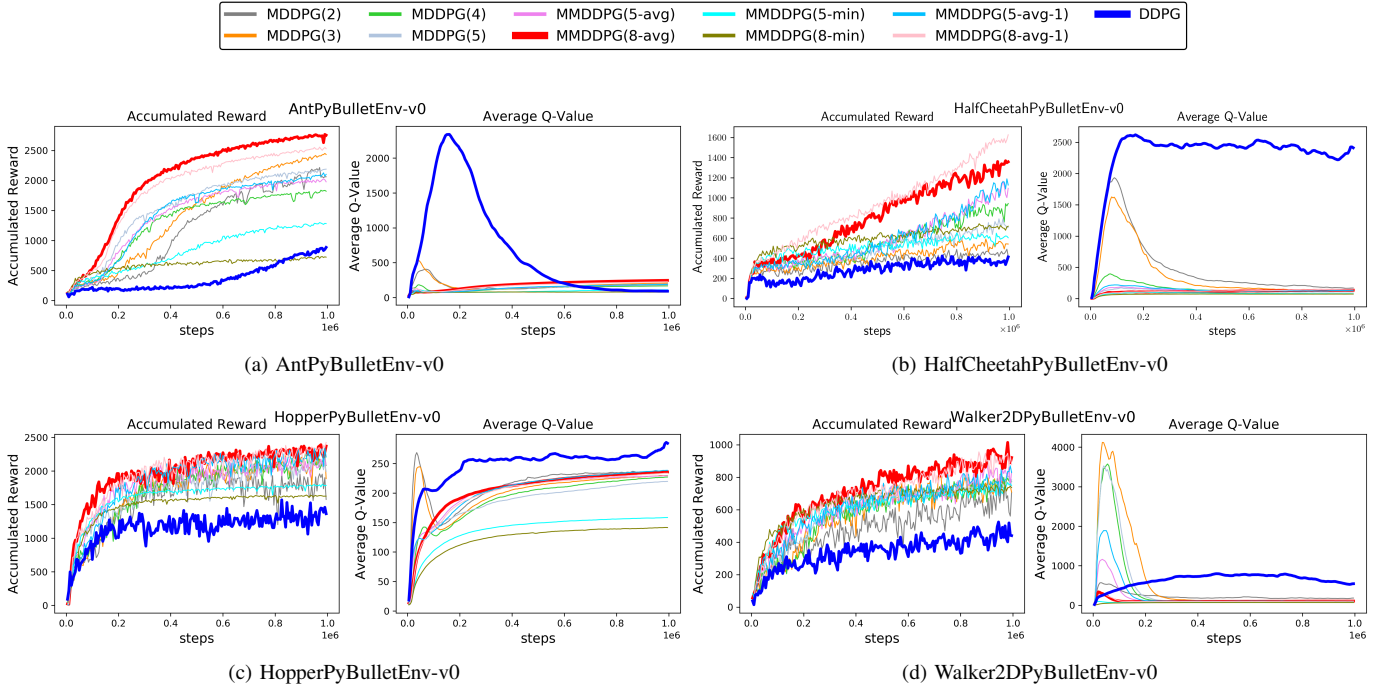


Fig. 1. Comparison Among MDDPG, MMDDPG and DDPG, where for each task accumulated reward and average Q-value are shown side-by-side correspondingly to demonstrate the relationship between the overestimation of Q-value and performance.

functions with 2 hidden layers and each layer has 300 hidden units. Other hyper-parameters are set to the default values. All experiments are run five times for five different random seeds.

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

Fig. 1 compares DDPG with its variants MDDPG and MMDDPG with different step size n . This figure illustrates that all MDDPG(n) with $n > 1$ outperform DDPG, and especially for MMDDPG(8-avg) the improvement, in terms of final performance and learning speed, is significant as highlighted with the red line. To illustrate the underlying relationship between the performance and learned Q-value, the average Q-value is shown in parallel, from which we can see that the bad performance of DDPG always corresponds to an extremely large Q-value. Note that even though multi-step methods help to relieve the overestimation problem, they cannot completely overcome this problem, as shown by the drastic increase and followed by the sharp decrease in Q-value within the first few epochs in Fig. 1, whereas DDPG takes more time to decrease its Q-value, and in some cases never does. The initial overestimation is caused by approximation error on most (state, action) pairs, because at the beginning stage of the learning only a small set of (state, action) pairs are encountered, causing a high error in (state, action) pairs without training data. However, as more and more experiences are collected in the replay buffer, the approximation error is reduced. From the average Q-values in Fig. 1, we can see that none of the examined approaches avoid this initial explosion in the average Q-values.

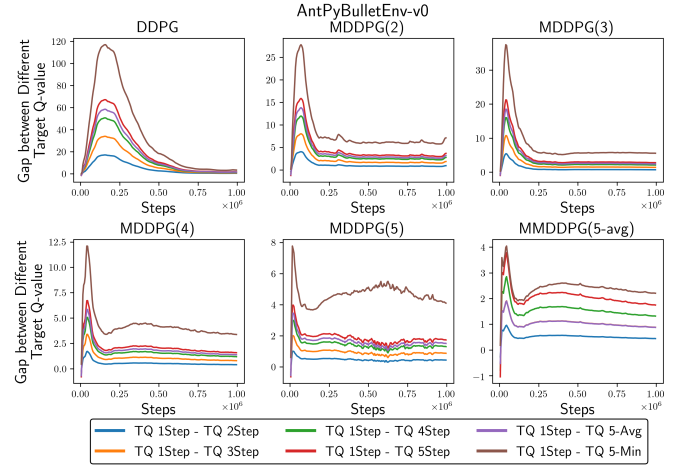


Fig. 2. The Difference in Estimated Target Q-values Between 1-step and Multi-step Methods, where the larger the value, the bigger the difference. (MDDPG(n) is a multi-step DDPG with step size n , and MMDDPG(n -avg) is a mixture of 1- to n -step DDPG.)

To investigate why multi-step methods help to alleviate the overestimation problem, we record backups of sampled experiences $\hat{Q}_t^{(1)}$, $\hat{Q}_t^{(2)}$, $\hat{Q}_t^{(3)}$, $\hat{Q}_t^{(4)}$, $\hat{Q}_t^{(5)}$, $\hat{Q}_t^{(n_{avg})}$, and $\hat{Q}_t^{(n_{min})}$ for DDPG, MDDPG(n) and MMDDPG(n -avg) to depict the gap between 1-step and multi-step backups. For example, $\hat{Q}_t^{(1)} - \hat{Q}_t^{(2)}$, indicated as “TQ 1Step- TQ 2Step” in Fig. 2 shows the difference between 1-step and 2-step backups. Four key characteristics can be observed in this figure: **(1)** within a specific algorithm all gaps are positive which means

multi-step methods provide smaller estimated target Q-values than that of the 1-step method; (2) **the larger the step, the smaller the corresponding estimated target Q-value**, e.g. the blue line underneath the yellow line in each sub-figure; (3) the difference becomes smaller with increased interactions; and (4) among the different algorithms, the magnitude of the estimated Q-value decreases as the step size n increases. These findings provide insight into multi-step methods' effect on alleviating the overestimation problem.

B. Performance Comparison

This section focuses on comparing MMDDPG(8-avg) with other baselines namely DDPG, TD3, SAC, MVE and STEVE. Special attention is given to MVE and STEVE, because these two algorithms are very similar to MDDPG and MMDDPG with the difference that they expand multi-steps in a learned environment model. Fig. 3 shows the learning curves of these algorithms on various tasks, and Table I summarizes the maximum average return. Obviously, MMDDPG(8-avg) significantly outperforms DDPG on all examined tasks. Surprisingly, MMDDPG(8-avg) performs comparably and even better on some tasks than TD3 which is currently one of the state-of-the-art approaches, specifically designed to address function approximation error in DDPG. Considering all examined tasks in this work have a dense reward signal, where multi-step methods' effect on enabling fast propagation of reward will be less important, we speculate that multi-step plays a similar role in alleviating the overestimation problem as does TD3, but using a different mechanism. Compared with MDDPG and MMDDPG, the disadvantage of TD3 is it introduces more computation for training its critics, because it maintains two separate critics and at each training step these two critics are updated to the minimum estimated Q-value of their target critic networks. Detailed comparison in terms of computation cost among DDPG, MDDPG, MMDDPG and TD3 will be discussed in Section VI-2.

Counterintuitively, SAC performs worse than TD3 on tasks from PyBulletGym, and for some tasks SAC is even worse than DDPG. This is unexpected, as it is shown in [28] that SAC outperforms TD3 on some difficult continuous control tasks from OpenAi gym which use the MuJoCo [29] physics engine. One possible explanation is that PyBulletGym uses Bullet physics [30], at the same time environments in PyBulletGym are ported from Roboschool environments which are harder than MuJoCo gym, as the robot's body is heavier than that in MuJoCo tasks and termination states are added if robot flips over. This needs further investigation. Especially if SAC is going to be employed in a real robot, the belief that SAC is the best choice might be misleading.

MVE performs the worst on most tasks. STEVE is shown to be sample efficient on AntPyBulletEnv-v0 and is better than MVE, which is consistent with the results in [23]. However, STEVE is worse than MMDDPG(8-avg) and TD3.

VI. DISCUSSION

In this section, we discuss the advantages and disadvantages of different ways to do multi-step expansion, and expose the tradeoff between overestimation and underestimation that underlies offline multi-step methods. Then, we compare the computation resource consumption between TD3 and our proposed methods, since they show comparable final performance and learning speed.

1) *Comparison of Multi-step Expansion Methods:* As shown in Eq. 10, there are three ways to calculate $\hat{Q}_t^{(n)}(s_t, a_t)$ depending on how the $n - 1$ experiences after (s_t, a_t, r_t, d_t) are acquired: (1) **offline expansion**, sampled from the replay buffer, e.g. MDDPG and MMDDPG; (2) **online expansion**, sampled from the environment according to an online policy, e.g. $Q(\sigma)$ [16]; (3) **model-based expansion**, sampled from a learned environment model according to an online policy, e.g. MVE and STEVE.

$$\hat{Q}_t^{(n)}(s_t, a_t) = r_t + \underbrace{\sum_{k=1}^{n-1} \gamma^k r_{t+k}}_{\text{underestimation prone}} + \underbrace{\gamma^n \max_a Q_{\theta Q-}(s_{t+n}, a)}_{\text{overestimation prone}} \quad (10)$$

Theoretically, online expansion is the best as the multi-step experiences are directly sampled from the environment. However, this is unrealistic, because expanding multi-step for each experience (s_t, a_t, r_t, s_{t+1}) in a mini-batch is time-consuming, especially when running multiple parallel environments (e.g. in simulation) is impossible.

A compromise is learning an environment model, then doing multi-step expansion on the learned environment model, as is done in MVE and STEVE. The challenge with this approach is that learning an environment, including transition dynamics and reward function, might be as hard or even harder than learning a policy, even without considering the extra cost for computation resources. It is also not clear to what extent the error introduced by the learned environment model will harm the learning of a policy. As shown in Fig. 3 and Table I, MVE and STEVE do not provide significant benefit, compared with MMDDPG(8-avg) and TD3.

Offline expansion is a solution somewhat in between. On one hand, it does not need to learn an environment model or to expand according to current policy on the environment, but uses past experiences after (s_t, a_t, r_t, s_{t+1}) as an expansion of the current online policy on the environment, with only negligible extra computation required. On the other hand, it is not an exact expansion of online policy, which introduces error in the estimated target Q-value and tends to be an underestimation of Q-value following current online policy. Nevertheless, offline expansion gradually approaches online expansion as the replay buffer fills with experiences following a stable optimal policy. This is seen in Fig. 4 where the initial gap between online and offline multi-step expansion is big,

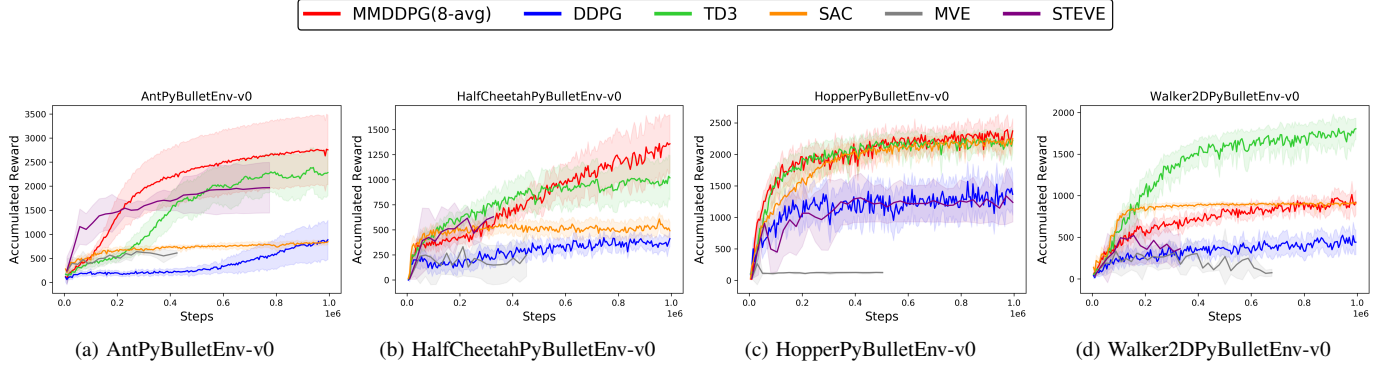


Fig. 3. Learning curves for PyBulletGym tasks. The shaded area shows half of standard deviation of the average accumulated return over 10 trials. MVE and STEVE are not run for full 1 million steps as they take more than 58 hours even on a machine with 2 x NVIDIA P100 Pascal and 3 CPUs @ 2.1GHz.

TABLE I
MAXIMUM AVERAGE RETURN OVER 10 TRIALS OF 1 MILLION STEPS OF MMDDPG(8-avg), DDPG, TD3, SAC, MVE AND STEVE. THE MAXIMUM VALUE FOR EACH TASK IS **BOLDED**.

Tasks	MMDDPG(8-avg)	DDPG	TD3	SAC	MVE	STEVE
AntPB	2767.1±1461.4	885.4±811.7	2388.0±832.6	845.5±103.7	639.5±33.9	1969.0±525.7
HalfCheetahPB	1368.9±527.3	422.2±137.6	1033.9±429.6	608.3±131.1	331.9±290.6	630.5±132.2
Walker2dPB	1014.0 ±316.3	524.0±227.5	1806.8±270.0	918.9±33.4	332.4±261.8	522.7±368.3
HopperPB	2391.9±473.3	1570.7±626.8	2253.9±295.2	2249.9±207.9	263.4±332.4	1338.6±449.6
AntMJC	3042.8±1038.5	2014.8±1371.6	3495.2±725.9	1680.5±414.6	—	—
HalfCheetahMJC	2242.2±338.6	1311.8±1367.6	2201.1±692.8	1977.7±180.4	—	—
Walker2dMJC	1365.6±409.9	844.7±521.1	1583.3±670.1	779.1±178.7	—	—

indicating large underestimation, but gradually decreases with the increase of interactions.

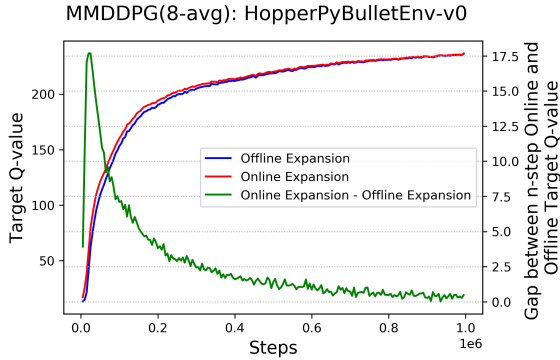


Fig. 4. Comparison between Online and Offline Multi-step Expansion, where the blue and the red line correspond to average of offline and online multi-step expansion over a mini-batch sampled from replay buffer, and the green line is the gap between them.

Obviously, multi-step expansion cannot completely overcome the overestimation problem, because the bootstrapped Q-value after n -steps is still prone to overestimation as shown in Eq. 10. But since $n > 1$, the bootstrapped part is weighted less than in the 1-step method. Overall, offline multi-step expansion tends to be an underestimation of the online multi-step expansion, while the bootstrapped Q-value after n -step tends to be an overestimation of the value in state s_{t+n} . Therefore, the step size n balances the tradeoff between

overestimation and underestimation.

2) *Computation Resource Consumption Comparison with TD3*: Similar to MVE and STEVE, MDDPG and MMDDPG proposed in this paper employ multi-step expansion to provide a more accurate target Q-value estimation for the critic in DDPG. As discussed in Section V-B and VI-1, MMDDPG outperforms MVE and STEVE in terms of learning speed, final performance and computation resource consumption. However, unlike multi-step expansion, TD3 takes the minimum of target Q-values estimated from two critics as the final target Q-value to update these two critics, to avoid value approximation error. MDDPG and MMDDPG show comparable final performance and learning speed as TD3 on most tasks. Here we focus on comparing the computation resource consumption of these three methods.

TABLE II
COMPARISON OF FORWARD AND BACKWARD PROPAGATION ON A MINI-BATCH FOR UPDATING THE CRITIC

	DDPG	TD3	MDDPG(n)	MMDDPG(n)
FP	1	2	1	n
BP	1	2	1	1
Total	2	4	2	$n+1$

FP: Forward Propagation, BP: Backward Propagation.

Table II summarizes the time of forward and backward propagation needed for training the critic on a mini-batch of transitions. As shown in the table, DDPG needs 1 forward propagation to estimate the target Q-value and 1 backward

propagation to update the critic. TD3 performs 2 forward and 2 backwards propagations, one for each critic. MDDPG with a specific step size n does not introduce extra propagations compared with DDPG. However, the number of forward propagations needed for MMDDPG with a specific choice of step size n is n , while only 1 backward propagation is needed for the updating critic. Therefore, MDDPG consumes less computation resource than TD3, while MMDDPG consumes more computational resource than TD3 only when $n \geq 4$, assuming the forward and backward propagation are equally demanding in terms of computational resources.

VII. CONCLUSIONS

In this paper, we empirically revealed multi-step methods' effect on alleviating overestimation in DRL, by proposing MDDPG and MMDDPG which are a combination of DDPG and multi-step methods, and discussed the underlying underestimation and overestimation tradeoff. Results show that employing multi-step methods in DRL helps to alleviate the overestimation problem by exploiting bootstrapping. This paper also discussed the advantages and disadvantages of three ways to implement multi-step methods from the point of view of extra computation cost and modeling error.

However, a principled way for choosing step size n is still needed. Perhaps dynamically tuning n during the course of learning is more suitable as at different stages of learning the trade-off between overestimation and underestimation needs to be balanced differently. The most important future direction arising from this work is to find a more effective way to overcome overestimation since this is key to improving DRL algorithms' sample efficiency, while still retaining a simple exploration method in order to limit computational needs.

ACKNOWLEDGMENT

This research was enabled in part by support provided by SHARCNET (www.sharcnet.ca) and Compute Canada (www.computeCanada.ca). The authors would like to thank Daiwei Lin for fruitful discussion. This work is supported by a SSHRC Partnership Grant in collaboration with Philip Beesley Architect, Inc.

REFERENCES

- [1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [2] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [3] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, *et al.*, "The limits and potentials of deep learning for robotics," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 405-420, 2018.
- [4] A. Irpan, "Deep reinforcement learning doesn't work yet," <https://www.alexirpan.com/2018/02/14/rl-hard.html>, 2018.
- [5] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
- [6] L. Meng, D. Lin, A. Francey, R. Gorbet, P. Beesley, and D. Kulić, "Learning to engage with interactive systems: A field study," *arXiv preprint arXiv:1904.06764*, 2019.
- [7] M. Plappert, R. Houthoofd, P. Dhariwal, S. Sidor, R. Y. Chen, X. Chen, T. Asfour, P. Abbeel, and M. Andrychowicz, "Parameter space noise for exploration," *arXiv preprint arXiv:1706.01905*, 2017.
- [8] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.
- [9] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.
- [10] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. Van Hasselt, and D. Silver, "Distributed prioritized experience replay," *arXiv preprint arXiv:1803.00933*, 2018.
- [11] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba, "Hindsight experience replay," in *Advances in Neural Information Processing Systems*, 2017, pp. 5048-5058.
- [12] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521-3526, 2017.
- [13] S. Thrun and A. Schwartz, "Issues in using function approximation for reinforcement learning," in *Proceedings of the 1993 Connectionist Models Summer School Hillsdale, NJ. Lawrence Erlbaum*, 1993.
- [14] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *arXiv preprint arXiv:1802.09477*, 2018.
- [15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [16] K. De Asis, J. F. Hernandez-Garcia, G. Z. Holland, and R. S. Sutton, "Multi-step reinforcement learning: A unifying algorithm," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [17] H. van Seijen, "Effective multi-step temporal-difference learning for non-linear function approximation," *arXiv preprint arXiv:1608.05151*, 2016.
- [18] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [19] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, 2016, pp. 1928-1937.
- [20] H. V. Hasselt, "Double q-learning," in *Advances in Neural Information Processing Systems*, 2010, pp. 2613-2621.
- [21] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Thirtieth AAAI conference on artificial intelligence*, 2016.
- [22] V. Feinberg, A. Wan, I. Stoica, M. I. Jordan, J. E. Gonzalez, and S. Levine, "Model-based value estimation for efficient model-free reinforcement learning," *arXiv preprint arXiv:1803.00101*, 2018.
- [23] J. Buckman, D. Hafner, G. Tucker, E. Brevdo, and H. Lee, "Sample-efficient reinforcement learning with stochastic ensemble value expansion," in *Advances in Neural Information Processing Systems*, 2018, pp. 8224-8234.
- [24] O. Anschel, N. Baram, and N. Shimkin, "Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org*, 2017, pp. 176-185.
- [25] J. F. Hernandez-Garcia and R. S. Sutton, "Understanding multi-step deep reinforcement learning: A systematic study of the dqn target," *arXiv preprint arXiv:1901.07510*, 2019.
- [26] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," 2014.
- [27] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.
- [28] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *arXiv preprint arXiv:1801.01290*, 2018.
- [29] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026-5033.
- [30] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," <http://pybullet.org>, 2016-2019.