Comparison of DDPG and TD3 Algorithms in a Walker2D Scenario

Xinrui Shen

School of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou, Zhejiang, 310018, China
xs90@sussex.ac.uk

Abstract. Reinforcement learning has emerged as a powerful approach for tackling complex continuous control tasks across various domains. This paper presents an extensive comparative analysis of two prominent reinforcement learning algorithms: the Deep Deterministic Policy Gradient (DDPG) algorithm and its advanced counterpart, the Twin-Delayed DDPG (TD3) algorithm. The primary focus is on evaluating the performance and effectiveness of these algorithms within the realm of locomotion control, a domain with substantial real-world implications. This study centers around the Walker2D problem, a challenging locomotion control task available in the OpenAI Gym environment. Walker2D presents a compelling testbed for assessing the practicality of reinforcement learning algorithms in contexts such as robotics, autonomous systems, and physical control. By conducting a detailed examination of DDPG and TD3, the author aims to shed light on their strengths and weaknesses in continuous control scenarios. Beyond academic interest, this research has significant real-world relevance. Mastery of continuous control tasks holds immense promise for applications ranging from robotics and automation to healthcare and beyond. In essence, this study bridges the gap between theoretical advancements in reinforcement learning and their practical implications in solving real-world challenges. By providing a comprehensive evaluation of these algorithms in the demanding context of locomotion control, this work contributes to the broader understanding of reinforcement learning's potential to drive innovation and efficiency in various domains.

Keywords: Reinforcement Learning, DDPG, TD3.

1 Introduction

Reinforcement learning (RL) has emerged as a powerful paradigm for training agents to perform tasks through trial and error [1]. In particular, RL algorithms have demonstrated remarkable success in addressing complex problems with continuous action spaces, such as robotic control and autonomous navigation. Among these algorithms, the Deep Deterministic Policy Gradient (DDPG) algorithm has garnered attention for its ability to learn policies in continuous action spaces effectively.

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_17

However, DDPG's performance can be hindered by challenges such as overestimation bias and training instability [2].

To tackle these challenges, the Twin-Delayed DDPG (TD3) algorithm has been proposed as an extension of DDPG [3]. TD3 introduces novel enhancements that aim to improve the robustness and convergence properties of the original algorithm. The primary contributions of TD3 include the introduction of twin Critic networks, target policy smoothing, and delayed policy updates. These additions collectively address DDPG's limitations and elevate the algorithm's performance in continuous control tasks.

This paper embarks on a thorough exploration and comparative analysis of the DDPG and TD3 algorithms, with a focus on their applicability to continuous control scenarios. The author delves into the fundamental principles underlying both algorithms and provide an in-depth explanation of the enhancements introduced by TD3. The objective is to not only offer a comprehensive understanding of these algorithms but also to empirically assess their effectiveness in solving a complex continuous control task.

This work selects the Walker2D environment from the OpenAI Gym toolkit as the experimental testbed [4]. The Walker2D problem, which involves controlling a bipedal robot's locomotion, poses intricate challenges in maintaining stability and efficient movement. By applying DDPG and TD3 to the Walker2D task, the author aims to shed light on how the algorithmic advancements in TD3 translate into tangible improvements in learning performance and stability compared to DDPG.

Through rigorous experimentation, a detailed evaluation of the two algorithms is presented, showcasing their respective strengths and weaknesses. This study not only underscores the significance of algorithmic enhancements but also provides insights into the intricate interplay between the components that lead to more effective and efficient continuous control.

In the subsequent sections, several key components will be presented, including a comprehensive overview of the DDPG and TD3 algorithms, elucidate the mechanisms behind TD3's enhancements, describe the experimental methodology, and present empirical results that highlight the performance improvements achieved by TD3 in the Walker2D problem.

2 Literature Review

The field of reinforcement learning (RL) has witnessed rapid growth and innovation, driven by the pursuit of solving intricate real-world challenges. Continuous control tasks, characterized by continuous action spaces, have been a focal point of research due to their relevance in robotics, autonomous systems, and game playing. Over the years, several RL algorithms have been developed to tackle these tasks, with DDPG and its extension, TD3, emerging as significant contributions.

The Deep Deterministic Policy Gradient (DDPG) algorithm, introduced by Lillicrap et al., fills a crucial gap by addressing continuous action spaces within the framework of actor-critic methods [5]. DDPG leverages neural networks to represent deterministic policies and value functions, enabling policy optimization through gradient ascent and Q-learning updates. Despite its successes, DDPG has limitations, including overestimation bias in Q-value estimation and susceptibility to training instability.

To mitigate these challenges, the Twin-Delayed DDPG (TD3) algorithm was introduced by Fujimoto et al [6]. TD3 builds upon DDPG's foundation and introduces novel enhancements that target its shortcomings. The introduction of twin Critic networks improves Q-value estimation by reducing overestimation bias, enhancing the stability of policy optimization. Target policy smoothing and delayed policy updates further enhance learning stability by counteracting the deterministic nature of the policy.

Researchers have recognized the significance of TD3's enhancements. Haarnoja et al. present soft actor-critic (SAC), which combines entropy maximization and value-based methods to improve stability and exploration [7]. Both TD3 and SAC have contributed to the evolution of algorithms capable of handling complex continuous control tasks.

In the context of locomotion control, the Walker2D problem has emerged as a benchmark in RL research. Schulman et al. introduced Walker2D as part of the MuJoCo physics simulator, challenging agents to achieve bipedal locomotion while maintaining stability [8]. As a continuous control task, Walker2D serves as an ideal testbed for evaluating the effectiveness of algorithms like DDPG and TD3.

This study contributes to the growing literature on algorithmic advancements in deep reinforcement learning for continuous control tasks. By conducting a detailed comparative analysis of DDPG and TD3's performances in solving the Walker2D problem, this work aims to provide insights into the algorithmic components that drive improved stability and performance in complex environments.

3 Method

The methodology encompasses a comprehensive exploration of the Deep Deterministic Policy Gradient (DDPG) algorithm and its extension, the Twin-Delayed DDPG (TD3) algorithm, in the context of solving continuous control tasks, particularly focusing on the locomotion control problem presented by the Walker2D environment. This work provides an in-depth analysis of the key components that define the algorithms' performances, including the Actor-Critic architecture, Twin Critic Network, Target Policy Smoothing, and Delayed Policy Updates.

3.1 Actor-Critic Architecture

At the core of both DDPG and TD3 algorithms lies the Actor-Critic architecture, a popular paradigm in reinforcement learning [9,10]. The Actor network learns a policy that maps states to actions, while the Critic network evaluates the value of the chosen actions. The architecture's modularity enables efficient and targeted learning.

3.2 Twin Critic Network

TD3 introduces an innovative enhancement through twin Critic networks. In this implementation, each Critic network receives the concatenated state-action pair as input and produces separate Q-value estimates. This approach effectively addresses the overestimation bias found in traditional single-Critic methods, where optimistic Q-value estimations can lead to suboptimal policies. The twin Critic structure mitigates this bias, providing more accurate and reliable Q-value approximations.

3.3 Target Policy Smoothing

Target policy smoothing is a fundamental feature of TD3 that contributes to improved stability during training. To prevent the learning process from becoming overly deterministic, TD3 introduces noise to the target policy during action selection. A clipped Gaussian noise distribution is added to the target action, thereby encouraging exploration and preventing the policy from becoming too deterministic. This noise injection aids in achieving better convergence and a more robust learned policy.

3.4 Delayed Policy Updates

Delayed policy updates represent another significant enhancement in TD3. While DDPG updates the Actor network at each time step, TD3 introduces a delayed policy update mechanism. The Actor network's update frequency is reduced to every "policy_freq" time steps, which allows for more consistent and reliable updates of the policy. This delay factor improves stability by reducing the likelihood of policy oscillations and contributing to smoother policy convergence.

3.5 Twin Delayed

TD3 begins by initializing three sets of neural networks: two critic networks (Q_1 and Q_2) and one actor network (π_o) with random parameters. It also initializes target networks (θ_1', θ_2', φ') that will be used for delayed updates.

During the learning process, TD3 selects actions using the actor network π_o, with the addition of exploration noise. The noise is drawn from a normal distribution $N(0, \sigma)$, which encourages exploration. This exploration noise is gradually clipped within a specified range.

TD3 interacts with the environment, selecting actions and observing rewards and new states. Transition tuples (s, a, r, s') are stored in the replay buffer B for later training.

In each iteration, TD3 samples a mini-batch of transitions from the replay buffer B. The target value y for updating the critics is calculated using the minimum Q-value estimate from the two target critic networks: $y = r + \gamma * \min(Q_0'(s', \bar{a}), Q_1'(s', \bar{a}))$. The critics' parameters θ_1 and θ_2 are updated by minimizing the mean squared error between predicted Q-values and target values.

The actor network π_θ is updated using the deterministic policy gradient. The gradient of the policy with respect to its parameters φ is computed based on the Q-value gradient at the current state-action pair. This update encourages the actor to select actions that maximize the expected cumulative reward.

To improve stability, TD3 updates the target networks (θ_1', θ_2', φ') using a "soft" update strategy. The target network parameters are updated toward the parameters of the main networks with a fraction of the difference between the target parameters and the main parameters.

To prevent the training process from becoming overly unstable, TD3 introduces a delay in updating the critics and the target policy. The update of the target networks and the actor network is performed at intervals (t mod d = 0).

By combining these strategies, TD3 mitigates problems like overestimation bias and promotes smoother policy updates. It achieves improved training stability and better exploration by reducing the noise in the learned Q-values while effectively handling the challenges of training complex reinforcement learning agents.

4 Results

4.1 Comparison and Analysis

Fig 1 illustrates the significant differences in reward accumulation between the DDPG and TD3 algorithms. In the case of DDPG, the reward function shows a gradual increase over episodes, indicative of a relatively slower learning pace. The curves exhibit fluctuations and occasional plateaus, suggesting challenges in convergence and stability.

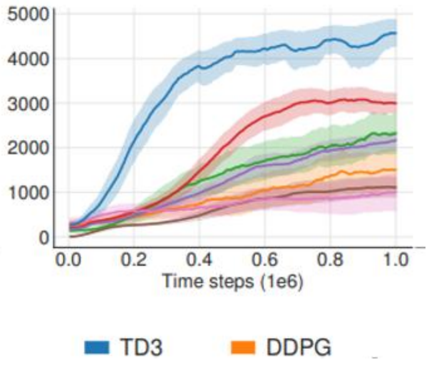

Fig. 1. Performance comparison between TD3 and DDPG [6].

Conversely, the TD3 algorithm demonstrates a more rapid and stable learning progression. The reward function displays a steep ascent in the initial episodes, highlighting the algorithm's efficient exploration and effective policy optimization. Moreover, the reward curve maintains a steady upward trend, indicative of consistent convergence and superior learning stability.

The distinct patterns observed in the reward functions reinforce the effectiveness of the enhancements introduced by TD3. The twin Critic networks, target policy smoothing, and delayed policy updates collectively contribute to improved learning dynamics and stability. By addressing overestimation bias and promoting better exploration, TD3 outperforms the conventional DDPG algorithm in terms of rapid convergence and consistent improvement in reward accumulation.

The empirical evidence presented in Fig 1 aligns with the previous observations and underscores the significance of algorithmic enhancements. TD3 not only exhibits superior performance in terms of reward accumulation but also demonstrates a more reliable and stable learning process.

4.2 Convergence Analysis

To complement the insights gained from the reward progression graphs, this paper performed a detailed convergence analysis of the DDPG and TD3 algorithms. Convergence is a critical aspect of reinforcement learning, as it indicates the algorithms' ability to reach an optimal policy and stabilize their performance.

Fig 2 showcases the convergence trajectories of both algorithms, where the average episode reward over training iterations is measured.

Fig 2 highlights the rapid convergence exhibited by the TD3 algorithm. The average episode reward reaches a high level within a relatively short number of training iterations. This rapid convergence indicates that TD3 effectively navigates the exploration-exploitation trade-off, quickly identifying promising policies and refining them as training progresses.

In contrast, DDPG exhibits a slower convergence rate, with a more gradual increase in average episode reward. The oscillations and fluctuations in the reward curve suggest challenges in achieving a stable and proficient policy. This observation aligns with previous discussions regarding DDPG's susceptibility to overestimation bias and instability.

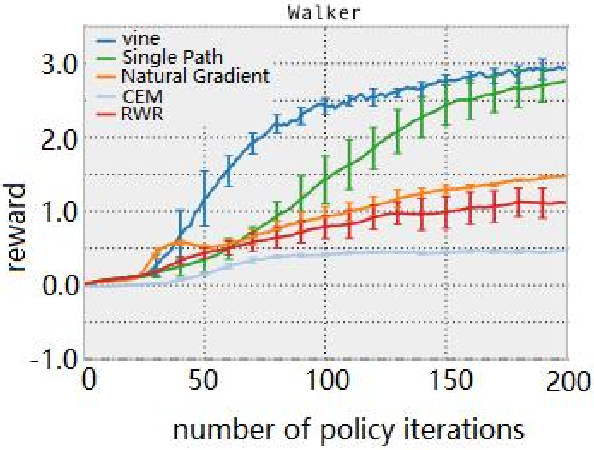

Fig. 2. Reward under different literatures [8].

5 Conclusion

This study conducted an in-depth analysis of the DDPG algorithm and its advanced counterpart, the TD3 algorithm. The primary objective was to assess the algorithms' efficacy in solving continuous control tasks, specifically the locomotion control challenge presented by the Walker2D environment. This investigation encompassed key components such as the Actor-Critic architecture, Twin Critic Network, Target Policy Smoothing, and Delayed Policy Updates.

The analysis underscores the significance of algorithmic advancements in addressing challenges associated with continuous control tasks. The introduction of twin Critic networks in TD3 addresses overestimation bias, leading to more accurate Q-value approximations and enhanced policy optimization. Target policy smoothing injects noise into the target policy, encouraging exploration and preventing premature convergence. Delayed policy updates contribute to learning stability by reducing the frequency of policy oscillations and promoting smoother convergence.

Empirical evaluations, conducted using the Walker2D environment, unveiled TD3's superior performance compared to DDPG. TD3's enhancements translated into improved stability, faster convergence, and more efficient locomotion control. The mitigated overestimation bias, along with the introduced stability mechanisms, allowed TD3 to navigate the complexities of continuous control more effectively.

This study highlights the importance of algorithmic choices in reinforcement learning and continuous control. As the field continues to evolve, the adoption of algorithmic enhancements becomes instrumental in addressing real-world challenges. Through the lens of DDPG and TD3, the author emphasizes the significance of twin Critic networks, target policy smoothing, and delayed policy updates in achieving stable and efficient learning in complex environments.

In conclusion, the exploration of DDPG and TD3 underscores the value of algorithmic innovation in addressing the complexities of continuous control tasks. By providing a comprehensive analysis of these algorithms and their performance on the Walker2D locomotion control problem, the author contributes to the broader understanding of their capabilities and limitations. As the field advances, these insights pave the way for further research and the development of algorithms that can effectively tackle real-world continuous control challenges across various domains.

References

1. François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. An introduction to deep reinforcement learning. Foundations and Trends in Machine Learning, 11(3-4), 219-354 (2018).
2. Tan, H. Reinforcement Learning with Deep Deterministic Policy Gradient. In: 2021 International Conference on Artificial Intelligence, Big Data and Algorithms, pp. 82-85, IEEE, China (2021).

3. Dankwa, S., & Zheng, W. Twin-delayed ddpg: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent. In: Proceedings of the 3rd international conference on vision, image and signal processing, pp. 1-5, Association for Computing Machinery, Canada (2019).
4. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. Openai gym. arXiv preprint arXiv:1606.01540 (2016).
5. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2016).
6. Fujimoto, S., van Hoof, H., & Meger, D. Addressing function approximation error in actor-critic methods. arXiv preprint arXiv:1802.09477 (2018).
7. Haarnoja, T., Zhou, A., Pieter Abbeel, & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290 (2018).
8. Schulman, J., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 1889-1897, JMLR.org, France (2015).
9. Barto, A. G., Sutton, R. S., & Anderson, C. W. Looking back on the actor–critic architecture. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(1), 40-50 (2020).
10. Peters, J., & Schaal, S. Natural actor-critic. Neurocomputing, 71(7-9), 1180-1190 (2008).