COMPARATIVE ANALYSIS OF TD3 AND SAC ALGORITHMS IN CONTINUOUS CONTROL ENVIRONMENTS

RIJU DATTA [DATTARIJ@SEAS]

ABSTRACT. In this study, we evaluate the performance of two prominent off-policy deep reinforcement learning algorithms, Twin Delayed Deep Deterministic Policy Gradient (TD3) and Soft Actor-Critic (SAC), within the MuJoCo-based DeepMind Control Suite environment, focusing on the "Walk" and "Run" tasks of the Walker domain. Our comparative analysis reveals that TD3 not only achieves higher rewards but also demonstrates significantly more robust performance across these tasks compared to SAC. These findings underscore the efficacy of the TD3 algorithm, particularly its advantage in environments requiring precise and stable control, which is critical for real-time applications such as robotic locomotion. The results suggest that the architectural and operational differences between TD3 and SAC, including TD3's targeted noise reduction techniques and delayed policy updates, contribute to its superior performance. This study highlights the importance of algorithm selection in reinforcement learning applications and suggests areas for further research into optimizing algorithmic components for specific tasks within continuous control spaces.

1. INTRODUCTION

The burgeoning field of reinforcement learning (RL) offers transformative potential for automating complex decision-making and control tasks across a spectrum of applications, from robotics to autonomous vehicles. Central to realizing this potential is the ability to evaluate and compare the performance of various RL algorithms reliably within simulated environments that approximate real-world conditions. The DeepMind Control Suite, a sophisticated platform for such simulations, provides a structured environment to benchmark the efficacy of different continuous control space algorithms. However, despite these advancements, establishing robust and generalizable performance benchmarks remains a challenge, primarily due to the high variability in algorithmic behavior across different tasks and settings. This project aims to extend the foundational work of the original authors by systematically assessing a range of RL algorithms on a diverse set of physical control tasks. By doing so, we not only contribute to the development of more stable and effective learning agents but also provide critical insights necessary for transitioning these agents from simulated scenarios to real-world applications, where robustness and reliability are paramount.

1.1. **Contributions.**

Riju - training, algorithm implementation, and reward/walker visualization code for both TD3 and SAC algorithms (ensuring compatibility with Deepmind Control Suite library); full report write-up

2. BACKGROUND

Deep reinforcement learning (Deep RL) represents a significant advancement in the field of artificial intelligence, merging the principles of classical reinforcement learning with the power of deep neural networks to solve complex decision-making and control tasks. Originating from the need to handle high-dimensional state and action spaces, Deep RL has evolved through the integration of deep learning, enabling agents to learn optimal policies directly from raw sensory inputs. Key principles in this field involve learning a policy function π that dictates the action a_t an agent takes in a given state s_t, with the goal of maximizing cumulative future rewards.

Among the notable developments in Deep RL, the Twin Delayed Deep Deterministic Policy Gradient (TD3) and Soft Actor-Critic (SAC) algorithms stand out for their contributions to addressing critical challenges like sample efficiency and training stability. TD3 improves upon the Deep Deterministic Policy Gradient (DDPG) by introducing techniques such as policy delay and dual critic architecture to reduce overestimation bias and variance, enhancing the robustness and reliability of learning in continuous action domains. In contrast, SAC incorporates the maximum entropy framework to systematically encourage exploration by rewarding the agent for trying diverse actions, thus ensuring better coverage of the state space and preventing premature convergence to suboptimal policies. These algorithms represent pivotal shifts in designing more efficient and stable strategies for deploying RL in real-world scenarios.

3. Related Work

In the domain of reinforcement learning, various algorithms have been rigorously evaluated using the DeepMind Control Suite to establish robust performance benchmarks [1]. Notable among these are the A3C algorithm, an asynchronous version of the Advantage Actor Critic, which facilitates distributed training but may suffer from stability issues due to asynchronous updates. The DDPG algorithm, recognized for enabling continuous action space learning through deterministic policy gradients, is often critiqued for its hyperparameter sensitivity and instability. Extending DDPG, the D4PG introduces a distributional approach to the critic's value function and prioritized experience replay, targeting improvements in data efficiency and stability.

In our work, we further explore the Twin Delayed DDPG (TD3) and Soft Actor-Critic (SAC) models, which are advancements over the aforementioned agents. TD3 addresses the overestimation bias and variance issues prevalent in DDPG by employing dual critic networks and policy delay updates, enhancing learning stability and performance. SAC, leveraging the maximum entropy framework, focuses on optimizing both the reward and entropy to encourage exploration and prevent premature convergence, thus offering a significant improvement in sample efficiency and policy robustness. These enhancements make TD3 and SAC more suited to complex, real-world applications requiring robust and efficient decision-making.

4. Approach

DeepMind Control Suite Library

The DeepMind Control Suite [1] is a comprehensive collection of continuous control tasks designed to serve as standardized performance benchmarks for reinforcement learning (RL) agents. The suite, implemented in Python and powered by the MuJoCo physics engine, offers an array of tasks that are straightforward to use and modify, with standardized action, observation, and reward structures that facilitate easy benchmarking and interpretation of learning curves.

Key features of the DeepMind Control Suite include:

Stable and Interpretable Tasks: The tasks are well-tested and structured, ensuring stability and interpretability of rewards.

Ease of Use and Modification: Written in Python with physical models defined using MJCF, the tasks are easily accessible and modifiable.

Benchmarking and Evaluation: The suite includes benchmarks for various RL algorithms, enabling consistent performance comparisons.

Walker Domain: Walk and Run Tasks One of the domains in the suite is the Walker domain, which features an improved planar walker model. The tasks in this domain are "walk" and "run".

Dimensions: The state, control, and observation spaces are defined as (18, 6, 24). Walk and Run Tasks: Rewards include components that incentivize forward velocity, challenging the RL agents to optimize for speed and stability.

Description of TD3 and SAC Deep RL algorithms

TD3 Algorithm

Summary of the TD3 Algorithm: The Twin Delayed Deep Deterministic policy gradient (TD3) algorithm is an advanced reinforcement learning method designed to address overestimation bias and function approximation errors in actor-critic settings, particularly within continuous control domains [2]. This algorithm extends the Deep Deterministic Policy Gradient (DDPG) framework by incorporating strategies from Double Q-learning, using dual critics for more accurate value estimation and delayed policy updates to stabilize training.

Key Features of TD3:

(1) Clipped Double Q-learning: TD3 modifies the traditional Double Q-learning approach for actor-critic methods by maintaining two separate critics. Each critic independently estimates the action values, and the smaller of

the two estimates is used for the policy update. This technique mitigates the overestimation bias commonly seen in single-critic architectures like DDPG by providing a more conservative estimate of the action values.
(2) Delayed Policy Updates: To prevent the policy from chasing noisy or inaccurate value estimates too rapidly, TD3 introduces delayed policy updates. The policy is updated less frequently than the value networks, which allows the value estimates to stabilize before they are used to adjust the policy. This reduces the impact of any single erroneous value update on the overall policy performance.
(3) Target Policy Smoothing: TD3 employs a regularization strategy called target policy smoothing, which involves adding noise to the target policy. This smoothing reduces the variance in the policy updates, leading to more stable learning. It mimics the off-policy learning update used in Expected SARSA, thereby enforcing that similar actions should have similar values, further stabilizing the learning process.

Comparison with Other Deep RL Algorithms:

Vanilla Policy Gradient and PPO Both Vanilla Policy Gradient and Proximal Policy Optimization (PPO) are policy-based methods that directly optimize the policy distribution. They are prone to high variance and instability due to their dependence on the quality of the policy gradient estimates. TD3's actor-critic architecture, enhanced by Double Q-learning and delayed updates, offers a more stable alternative by decoupling policy and value estimation, which is less susceptible to variance from individual gradients.

DDPG DDPG is a precursor to TD3 and shares its actor-critic framework. However, DDPG often suffers from overestimation bias due to the use of a single critic that updates its policy based on possibly noisy value estimates. TD3 addresses these flaws by integrating dual critics and smoothing techniques, significantly enhancing the reliability and stability of the training process.

Enhancements Over DDPG TD3's methodology for handling the inherent problems of function approximation and temporal difference learning in actor-critic methods leads to more stable and reliable policy learning, especially in environments with continuous action spaces. The use of dual critics helps balance the learning process by providing more grounded value estimates, while delayed policy updates ensure that these estimates are reliable before they influence the policy direction. Additionally, the target policy smoothing technique in TD3 helps to maintain a stable learning trajectory, unlike DDPG which can suffer from erratic policy updates due to its single critic design.

Overall, TD3 presents a robust framework for dealing with the complexities of continuous control tasks in deep reinforcement learning, outperforming older methods like DDPG, PPO, and Vanilla Policy Gradient by addressing their critical weaknesses in estimating and updating policies.

SAC Algorithm

The Soft Actor-Critic (SAC) is a model-free, off-policy actor-critic deep reinforcement learning algorithm that incorporates principles from maximum entropy reinforcement learning [3]. It is designed to address the major challenges in applying deep RL to real-world tasks, particularly high sample complexity and brittle convergence properties.

Key Features of SAC:

(1) **Entropy Maximization:** SAC integrates an entropy term into the reward objective, which encourages the policy to explore more widely and avoid premature convergence to suboptimal deterministic policies. This entropy maximization leads to more robust exploration and a more diverse range of behaviors, enhancing the algorithm's ability to escape local optima.
(2) **Actor-Critic Architecture:** The algorithm uses a standard actor-critic setup with separate policy (actor) and value function (critic) networks. This separation allows the critic to independently evaluate the policy dictated by the actor, which provides a stable and unbiased gradient for updating the actor.
(3) **Off-Policy Learning:** SAC leverages off-policy data using experience replay, which improves sample efficiency by reusing past experience for multiple updates. This feature allows SAC to learn effective policies with fewer interactions with the environment, a significant advantage in complex domains.
(4) **Stochastic Policy Framework:** Unlike algorithms that utilize deterministic policies, SAC employs a stochastic policy framework. This approach inherently incorporates exploration into the policy itself, avoiding the need for an external exploration mechanism like epsilon-greedy in DDPG.

Performance and Advantages: SAC has demonstrated state-of-the-art performance on a range of continuous control benchmark tasks. Its key advantages include:

- **Sample Efficiency:** Due to its off-policy nature and efficient use of experience replay, SAC significantly outperforms on-policy methods like PPO and traditional actor-critic algorithms in terms of the number of samples needed to achieve similar performance.
- **Stability:** The inclusion of the entropy term and the use of a stochastic policy lead to more stable learning dynamics compared to methods like DDPG, which can be sensitive to hyperparameters and suffer from stability issues.
- **Robustness:** Entropy maximization not only aids exploration but also makes the policy more robust to perturbations and model inaccuracies, which are common in real-world scenarios.

Shortcomings: While SAC presents numerous advantages, it also has some limitations:

- **Computational Complexity:** The need to calculate and optimize the entropy term can add computational overhead compared to simpler methods.
- **Tuning of Temperature Parameter:** The performance of SAC can be sensitive to the choice of the temperature parameter that balances the entropy and reward components of the objective. Finding the right balance requires careful tuning, which can be challenging.
- **Potential Over-exploration:** While the entropy term helps in exploration, excessive entropy can lead to overly stochastic policies that may not always converge to the most effective behaviors, especially in tasks where precise actions are crucial.

Soft Actor-Critic represents a significant advancement in the development of deep reinforcement learning algorithms for continuous action spaces. Its integration of maximum entropy principles with a robust actor-critic architecture makes it a powerful tool for tackling complex real-world problems. Despite its shortcomings, such as computational demands and sensitivity to hyperparameters, SAC's advantages in stability, efficiency, and robustness make it a preferred choice in many applications. Further research into optimizing its components could alleviate some of its limitations, potentially broadening its applicability and performance in even more challenging domains.

In both cases, we referenced the OpenAI Spinning Up implementations of the TD3 and SAC algorithms for a reference about how to design continuous control-space deep RL algorithms ([4] and [5]).

5. Experimental Results

In this study, we evaluated the performance of the Twin Delayed Deep Deterministic Policy Gradient (TD3) and Soft Actor-Critic (SAC) algorithms on the DeepMind Control Suite's Walker domain, specifically focusing on the 'Walk' and 'Run' tasks. Both algorithms were tested over a substantial span of 500,000 timesteps, a setup aimed at assessing their capacity for learning and stabilization over prolonged training periods. The experiments utilized a consistent network architecture, hyperparameters, and training configurations for each algorithm, ensuring that the evaluations focused solely on algorithmic efficacy without external optimization variances.

The test returns, plotted for all timesteps, reveal distinct performance characteristics between the two algorithms. TD3 shows a steady increase in performance in both tasks, indicated by the ascending trend in average test returns, suggesting an effective learning strategy that progressively improves through exploration and exploitation. In contrast, SAC exhibits more variability in its test returns, with notable spikes in performance but a less consistent upward trend. This suggests that while SAC can achieve high performance, it may require additional tuning to stabilize learning over time.

These findings highlight TD3's robustness and reliability in environments requiring continuous control, making it particularly suited for tasks where consistent performance is critical. Conversely, the high variability in SAC's results underscores its sensitivity to specific task dynamics, suggesting that its entropy-based exploration strategy, while beneficial in some scenarios, may lead to less predictable outcomes in others. Overall, these results contribute to our understanding of how different deep reinforcement learning strategies manifest in complex simulated environments, informing future research and application in real-world settings.

Links to Sample Videos of Walker-Walk and Walker-Run Tasks (using the TD3 algorithm):

(1) `https://drive.google.com/file/d/1G05ZrZgSyx6SgkieHOd_WPc4yqg5H334/view?usp=sharing` (Walker - Walk)
(2) `https://drive.google.com/file/d/1QvaJ_Wv7a9XPhVIUzn4yFX4k87L0gcv-/view?usp=sharing` (Walker - Run)

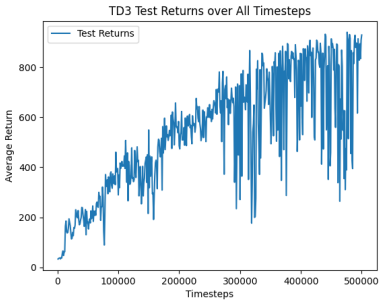

FIGURE 1. TD3 Test Return Plot for Walker-Walk Task

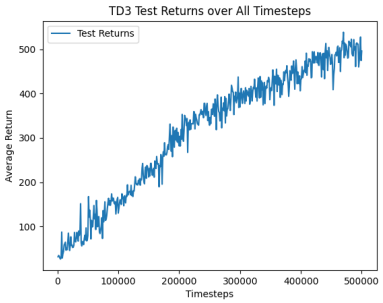

FIGURE 2. TD3 Test Return Plot for Walker-Run Task

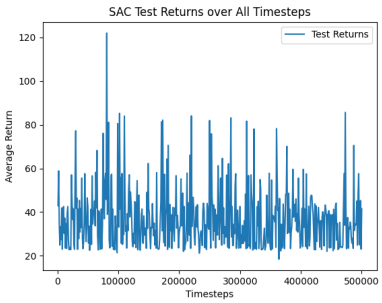

FIGURE 3. SAC Test Return Plot for Walker-Walk Task

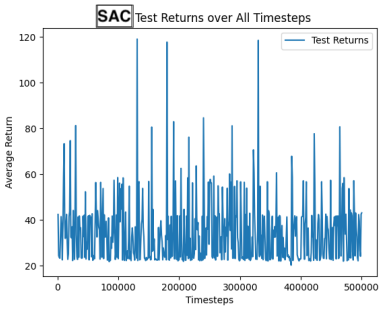

FIGURE 4. SAC Test Return Plot for Walker-Run Task

6. DISCUSSION

The experimental results reveal that the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm consistently outperforms the Soft Actor-Critic (SAC) in the Walker domain of the DeepMind Control Suite, particularly in tasks requiring stable and continuous control. One potential reason for TD3's superior performance could be its dual critic design and policy update delays, which help mitigate the overestimation bias often observed in value-based methods like SAC. These features likely contribute to a more stable learning curve and robust handling of the simulation's dynamics. Looking forward, to build upon the current findings and further improve the applicability of reinforcement learning in complex real-world scenarios, several steps could be considered. First, experimenting with hybrid models that integrate the strengths of both TD3 and SAC might provide a balanced approach to managing exploration and exploitation, potentially leading to improved performance across a broader range of tasks. Additionally, expanding the evaluation framework to include newer and more diverse simulation environments could help in understanding the limitations and capabilities of these algorithms under different conditions. Moreover, incorporating adaptive hyperparameter tuning mechanisms could enhance the algorithms' ability to self-optimize in response to the specific characteristics of the task and environment. Long-term investments in developing more sophisticated simulation models that closely mimic real-world physics could also be crucial. Such advancements would not only refine the training process but also ensure that the transition from simulation to real-world application is more seamless and reliable. These steps would collectively push the boundaries of current reinforcement learning applications, paving the way for more robust and universally applicable solutions.

REFERENCES

(1) https://arxiv.org/pdf/1801.00690 (Original DeepMind Control Suite paper)
(2) https://arxiv.org/pdf/1802.09477 (Original TD3 paper)
(3) https://arxiv.org/pdf/1801.01290 (Original SAC paper)
(4) https://spinningup.openai.com/en/latest/algorithms/td3.html (OpenAI Spinning Up Implementation of the TD3 algorithm)
(5) https://spinningup.openai.com/en/latest/algorithms/sac.html (OpenAI Spinning Up Implementation of the SAC algorithm)