Twin-Delayed Deep Deterministic Policy Gradient (TD3)

Anonymous Authors[1]

Abstract

blar

1. Paper Critiques

A summary and critical review is provided for each of the three papers selected for this project. I have chosen to reproduce results form the first paper, (Fujimoto et al., 2018), and, hence, a more detailed summary and review is included Paper 1.

1.1. Paper 1: (Fujimoto et al., 2018)

1.1.1. Summary

Fujimoto et al. propose improvements to the well-known deep Reinforcement Learning (RL) scheme Deep Deterministic Policy Gradient (DDPG) originally detailed by (Lillicrap et al., 2015). The proposed algorithm is referred to as either "Twin-Delayed Deep Deterministic Policy Gradient", "Twin-Delayed DDPG", or most concisely, TD3. The emergence of the TD3 algorithm can be directly traced back to early RL contributions through a series of major contributions. Various RL approaches use the state-action value function $Q(s, a)$, updated by iterating on the Bellman equation, to formulate an effective policy for off-policy control. An early breakthrough for problems with discrete state and action spaces occurred in 1989 with *Q-Learning*: a temporal difference learning algorithm that allows powerful agents to be trained in an off-policy manor (off-policy indicates that data may be collected from any policy, not necessarily from the agent being optimized) (Sutton & Barto, 2018). While effective, this approach, along with other RL approaches at the time, was limited by the assumption of discrete spaces. A major breakthrough in Reinforcement Learning came in 2015 when the Q-Learning algorithm was extended to continuous state space problems with the Deep Q-Network (DQN) algorithm (Mnih et al., 2015). The DQN approach involves the same update rule as Q-Learning, but rather than direct computation, is able to effectively estimate the Q function using a neural network: alleviating the need for tabulated results. However, similar to Q-Learning, DQN still suffered from the discrete action space assumption. This was okay for Atari game tasks, where actions were limited to available Atari controls, but limited the applicability for continuous control tasks, such as robotics. Addressing this step, (Lillicrap et al., 2015) introduced Deep Deterministic Policy Gradient (DDPG): an extension of DQN that uses a second "actor" neural network to construct a parameterized action function, and performed the $Q(s, a)$ update with the new actor network. A key contribution with DDPG is the inclusion of "target" networks for both the actor and value (i.e. *critic*) networks. By including duplicate networks, the second "target" networks may be held constant while the actual networks are updated. This is a more stable process because it doesn't involve continuously updating estimates to be used in the minibatch updates. While powerful, however, DDPG contained several notable limitation. In particular, the value function is often over-estimated, which leads to a significant bias in learning.

Twin-Delated DDPG addresses the instability of DDPG in several ways. First, two separate value networks are included. By taking the minimum value estimate of the two networks mitigates the effect of the value function over-estimation bias present in DDPG (along with other actor-critic schemes). Next, DDPG involves *bootstrapping*, i.e., performing updates based on estimates rather than true values. Noisy estimates cause noisy gradients which pose significant difficulty in network optimization. (Fujimoto et al., 2018) seek to mitigate this error by delaying updates to the actor network in hopes that additional updates to the critic network will provide more accurate estimates for the actor updates. Finally, to avoid peak overfitting in the policy network, policy smoothing is included, which introduces a small amount of clipped random noise to the output of the target policy. Together, these improvements form the TD3 variant of DDPG.

1.1.2. Critical Review

Fujimoto et. al. provide an extremely practical and concise account of their algorithm, contributions, and overall approach. Results from experiments are very compelling: outperforming state-of-the-art algorithms in virtually every

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

task (including their improved version of DDPG). While I understand this is not possible for a conference paper with length limitations, one drawback of this paper is that I don't think it would really be possible to fully understand the algorithm from this paper as a standalone document. In particular, understanding DDPG is vital to understanding TD3's improvements. In turn, an thorough understanding DDPG requires knowing the DQN algorithm, which itself builds off a base-knowledge of Q-Learning. I saw these as the most important 'steps' leading up to TD3, but this is by no means an exhaustive literature background for the TD3 approach. While this necessitated a significant amount of additional reading to fully grasp the approach, I don't fault the authors because they had a very helpful "Related Work" section that directed me to various sources to understand more of the underlying theory that enables their approach.

Literature review aside, there are several additional specific points worth mentioning. First, the authors did an excellent job in their discussion of the overestimation bias of value functions in actor-critic approaches. In addition to a theoretical understanding, the numerical results enhanced the discussion, and made it easily understandable for the reader. If I had to single out one part that could perhaps be clarified, it would be the section on smoothing regularization. I understand how the noise helps negate negative effects of peaks in the value estimate, however I don't think it is very clear how this noise should be produced, what scale it should be, and why it must come from a normal distribution. The appendices address how the policy must be clipped to avoid impossible actions from the noise, but I think a more thorough theoretical analysis of the underlying statistics could reveal a more rigorous way to choose the random process. One final concern I have about the approach itself is the overall complexity. Although results demonstrate its performance, the complexity of having six total neural networks (DQN originally only had one!) indicates to me that there may be unintended consequences of these complex interactions.

1.2. Paper 2: (Haarnoja et al., 2018)

1.2.1. Summary

(Haarnoja et al., 2018)

1.2.2. Critical Review

1.3. Paper 3: (Nachum et al., 2018)

1.3.1. Summary

1.3.2. Critical Review

2. Implementation Results

References

Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning (ICML)*, 2018. URL https://arxiv.org/pdf/1802.09477.pdf.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcementlearning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018. URL https://arxiv.org/pdf/1801.01290.pdf.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning, 2015. URL https://arxiv.org/pdf/1509.02971.pdf.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518, Feb 2015.

Nachum, O., Gu, S. S., Lee, H., and Levine, S. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 3303–3313, 2018. URL https://proceedings.neurips.cc/paper/2018/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf.

Sutton, R. S. and Barto, A. G. *Reinfrocement Learning: An Introduction*. The MIT Press, second edition, 2018.