Addressing Function Approximation Error in Actor-Critic Methods

Niharika Shrivastava
School of Computing
National University of Singapore
Singapore, 119077
niharika@comp.nus.edu.sg

Abstract

The authors of [1] have a threefold objective - to prove that overestimation bias and the accumulation of error over time are present in an actor-critic setting; to propose a novel actor-critic algorithm Twin Delayed Deep Deterministic Policy Gradient (TD3) to minimize overestimation bias; and to introduce a regularization method which bootstraps similar actions to further reduce variance.

1 Clipped Double Q-Learning for Actor-Critic

It is shown that an overestimation bias accumulates significantly over time leading to inaccurate value estimates which results in poor policy updates. Since, in an actor-critic setting, a feedback loop is present between the actor and critic networks, a sub-optimal action may get highly favoured by a sub-optimal critic, reinforcing the sub-optimal action in the next policy update.

To curb this problem, a clipped variant of Double Q-Learning is used with a pair of actors (π_{θ_1}, π_{θ_2}) and critics (Q_{θ_1}, Q_{θ_2}), where the target update value is computed by taking the minimum of the less biased estimate Q_{θ_2} and the biased estimate Q_{θ_1}. Thus, no additional overestimation in the target over the standard Q-learning target is introduced and any underestimation bias is inherently never propagated through the policy update.

2 Reduce Variance

Further, the use of a stable target network to reduce the growth of errors is established by showing that the occurrence of divergence without target networks is because of policy updates with a high variance value estimate. Therefore, policy and target networks are updated only after a delay thereby limiting repeating updates with respect to an unchanged critic. The final updates will have lower variance and higher quality.

Regularization is used to reduce inaccuracies during critic updation by adding clipped noise around the target action to smooth the value estimate by bootstrapping off of similar state-action value estimates.

3 Twin Delayed Deep Deterministic Policy Gradient (TD3)

TD3 combines these concepts by maintaining a pair of critics that update at each time step according to Clipped Double Q-Learning with regularization, along with a single actor having delayed policy updates. This resulted in improved learning speed and performance in the continuous domain.

Preprint. Under review.

References

[1] Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing Function Approximation Error in Actor-Critic Methods. ArXiv. https://doi.org/10.48550/arXiv.1802.09477