AFU is truly off-policy

Author Names Omitted for Anonymous Review. Paper-ID [add your ID here]

AFU is truly off-policy

Author Names Omitted for Anonymous Review. Paper-ID [add your ID here]

Abstract

1. Main idea

At the heart of many reinforcement learning algorithms, one performs temporal updates of an action value function $Q^\pi(s,a)$ of the general form:

$$Q^\pi(s_t,a_t) \leftarrow Q^\pi(s_t,a_t) + \alpha[r(s,a) + F^\pi(s_{t+1}) - Q^\pi(s_t,a_t)], \tag{1}$$

In the discrete state and discrete action case, the SARSA algorithm uses

$$F^\pi(s_{t+1}) = Q^\pi(s_{t+1},a_{t+1}), \tag{2}$$

where a_{t+1} is the action that the agent will take at the next time step. SARSA is considered an on-olicy algorithm and has been shown to converge provided that the behavior policy is "greedy in the limit of infinite exploration" [1].

In contrast, the Q-LEARNING algorithm uses

$$F^\pi(s_{t+1}) = \max_a Q^\pi(s_{t+1},a), \tag{3}$$

it is considered off-policy and it converges under a much wider range of behavior policies [2].

In the continuous state and action case, DDPG and TD3 are actor-critic approaches that use a deterministic policy. To update the critic, they also rely on (1), using

$$F^\pi(s_{t+1}) = Q^\pi(s_{t+1},\pi(s_{t+1})). \tag{4}$$

Finally, SAC uses a stochastic policy, resulting in:

$$F^\pi(s_{t+1}) = \mathbb{E}_{(a\sim\pi)}Q^\pi(s_{t+1},a) + entropy \tag{5}$$

with an additional entropy that we do not detail to focus on our point.

Now, coming back to SARSA, if we write with a slight abuse of notation $\pi(s_{t+1})$ the action that the agent will take in state s_{t+1} using its Q-table, we see that (2) can be rewritten as (4). Besides, (4) is just the deterministic special case of (5) without entropy, which is more general. This shows that the critic update in DDPG, TD3 and SAC inherits from SARSA rather than from Q-LEARNING. As a result, all these algorithms rely like SARSA on the fact that the policy π tends to output better and better actions in any state so that $\mathbb{E}_{(a\sim\pi)}Q^\pi(s_{t+1},a)$ approaches $max_a Q^\pi(s_{t+1},a)$. However, due to the

interdependency between π and Q^π and local optima in Q^π, convergence of the policy is far from guaranteed. Though they are characterized as off-policy, the inheritance of these algorithms from SARSA prevents them from being *truly off-policy*.

Designing an algorithm that inherits more directly from Q-LEARNING and benefits from its capability to converge from a wider set of behavior policies implies trying to get $F^\pi(s_{t+1}) = max_a Q^\pi(s_{t+1}, a)$ without relying on improvement from the policy, which means addressing the problem of maximizing Q values over the action space more frontally. This is the problem we address in AFU.

References

[1] Singh, S. P., Jaakkola, T., Littman, M. L., and Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38(3):287–308.

[2] Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292.