Breaking the Deadly Triad with a Target Network

Shangtong Zhang[1], Hengshuai Yao[2,3], Shimon Whiteson[1]

[1] University of Oxford
[2] University of Alberta
[3] Huawei Technologies

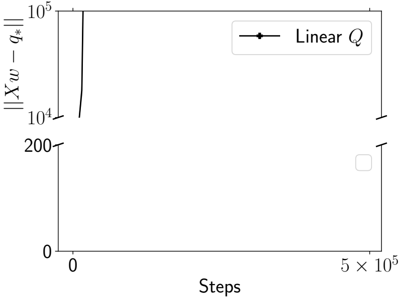
June 16, 2021

The deadly triad (Chapter 11.3 of Sutton and Barto (2018)) refers to the instability of an RL algorithm with function approximation, off-policy learning, and bootstrapping.

Linear Q-learning diverges in Barid's counterexample (Baird, 1995)

$$w_{t+1} \leftarrow w_t + \alpha \left(R_{t+1} + \gamma \max_a x(S_{t+1}, a)^\top w_t - x_t^\top w_t \right) x_t$$

$$x_t \doteq x(S_t, A_t)$$



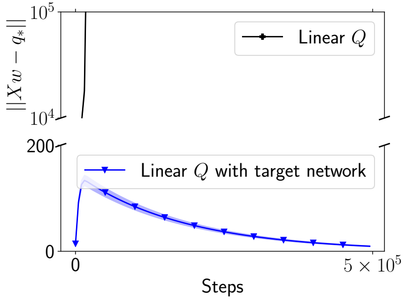
Surprisingly, linear Q-learning with a target network (Mnih et al., 2015) converges in Baird's counterexample

Linear Q-learning with a target network:

$$w_{t+1} \leftarrow w_t + \alpha \left(R_{t+1} + \gamma \max_a x(S_{t+1}, a)^\top \theta_t - x_t^\top w_t \right) x_t$$

$$\theta_{t+1} \leftarrow \theta_t + \beta(w_t - \theta_t)$$

Is this just by accident? No!



It is now proved that target network is an effective method to break the deadly triad in linear RL

$$w_{t+1} \leftarrow w_t + \alpha(R_{t+1} + \gamma \max_a x(S_{t+1}, a)^\top \theta_t - x_t^\top w_t)x_t - \alpha_t \eta w_t$$

$$\theta_{t+1} \leftarrow \Gamma_{B_1}(\theta_t + \beta(\Gamma_{B_2}(w_t) - \theta_t))$$

- η: ridge regularization
- Γ_{B_i}: projection to balls of radius B_i

A sufficient condition (not necessarily necessary): If $\|X\|$ is not too large, B_1 and B_2 are not too small, then $\{w_t\}$ converges to regularized TD fixed point.

The behavior policy can be w-dependent so it changes every step, and can be arbitrarily different from the target policy.

Why do we need two projections in updating the target network?

- With only Γ_{B_1}:

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = w^*(\theta(t)) - \theta(t) + \zeta(t),$$

  where $\zeta(t)$ is a reflection term.

- With both Γ_{B_1} and Γ_{B_2}:

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = w^*(\theta(t)) - \theta(t).$$

  Γ_{B_2} also ensures target network changes sufficiently slowly.

Our analysis of target network is widely applicable

(algorithms with linear per-step computational complexity)

- Policy Evaluation
  - Linear off-policy TD in discounted MDPs
  - Linear off-policy TD in average-reward MDPs (the first convergent linear off-policy policy evaluation algorithm for average-reward MDPs)
- Control
  - Linear Q-learning in discounted MDPs (the first convergent linear Q-learning with changing behavior policies and do not require behavior policies to be similar to target policies)
  - Improve Greedy GQ (Maei et al., 2010) to work with changing behavior policies
  - Linear Q-learning in average-reward MDPs (the first convergent linear off-policy control algorithm for average-reward MDPs)

Thanks

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. Machine Learning.

Maei, H. R., Szepesvári, C., Bhatnagar, S., and Sutton, R. S. (2010). Toward off-policy learning control with function approximation. In Proceedings of the 27th International Conference on Machine Learning.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature.

Sutton, R. S. and Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd Edition). MIT press.