Reinforcement Learning

5. Off-policy versus on-policy RL

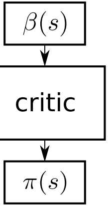
Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Introduction

- We have said that SARSA was on-policy and Q-LEARNING was off-policy.
- The goal of this class is to understand precisely what this means
- This distinction matters a lot for deep RL research
- Off-policy algorithms like DDPG are generally more sample efficient but less stable than on-policy algorithms like PPO

Basic concepts



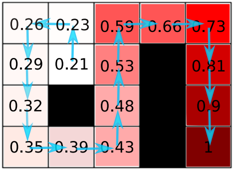
- To understand the distinction, one must consider three objects:
  - The behavior policy $\beta(s)$ used to generate samples.
  - The critic, which is generally $V(s)$ or $Q(s, a)$
  - The target policy $\pi(s)$ used to control the system in exploitation mode.

Off-policiness: definition

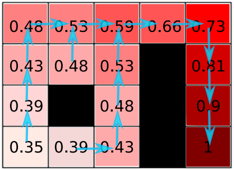
- "Off-policy learning" refers to learning about one way of behaving, called the *target policy*, from data generated by another way of selecting actions, called the *behavior policy*.
- Two notions:
  - Off-policy policy evaluation
  - Off-policy control

Maei, H. R., Szepesvári, C., Bhatnagar, S., & Sutton, R. S. (2010) Toward off-policy learning control with function approximation. *ICML*, pages 719–726.

Off-policy policy evaluation: Definition



$\beta(s)$



$\pi(s)$

- Can evaluate the critic of a target policy $\pi(s)$ from playing a different behavior policy $\beta(s)$?
- Obviously, $\beta(s)$ and $\pi(s)$ generate different values $V(s)$ or $Q(s, a)$
- The goal of "off-policy correction" is to correct for the sample mismatch
- The target policy does not need to be optimal
- This is a weak notion of off-policiness (not covered here)

Precup, D. (2000) Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*

Munos, R., Stepleton, T., Harutyunyan, A., & Bellemare, M. G. (2016) Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1054–1062

Off-policy control: Definition

- Whatever the behavior policy (as few assumptions as possible)
- The target policy should be an approximation to the optimal policy
- Ex: stochastic behavior policy, deterministic target policy
- An algorithm might be more or less off-policy depending on the assumptions on $\beta(s)$

Why prefering off-policy to on-policy control?

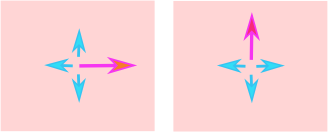
- More freedom for exploration
- Learning from human data (imitation)
- Reusing old data, e.g. from a replay buffer (sample efficiency)
- Transfer between policies in a multitask context

Approach

$$
\begin{tikzcd}
\boxed{\beta(s)} \arrow[d] \\
\boxed{\text{critic}} \arrow[d] \\
\boxed{\pi(s)}
\end{tikzcd}
$$

- Two steps: open-loop study then closed-loop study
  - Use uniform sampling as “behavior policy” (few assumptions)
  - No exploration issue, no bias towards good samples
  - NB: in uniform sampling, samples do not correspond to an agent trajectory
  - An alternative is random walk, but may raise exploration issues
  - Study critic learning from these samples
- Then close the loop:
  - Use the target policy + some exploration as behavior policy
  - If the target policy gets good, bias more towards good samples

Learning a critic from samples



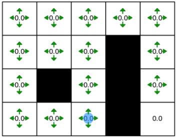
Two random actions

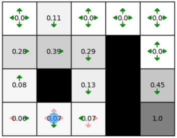
- General format of samples S: $(s_t, a_t, r_t, s_{t+1}, a')$
- Makes it possible to apply a general update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma Q(s_{t+1}, a') - Q(s_t, a_t)]$$

- There are three possible update rules:
  1. $a' = \text{argmax}\, a Q(s_{t+1}, a)$ (corresponds to Q-LEARNING)
  2. $a' = \beta(s_{t+1})$ (corresponds to SARSA)
  3. $a' = \pi(s_{t+1})$ (corresponds e.g. to DDPG, an ACTOR-CRITIC algorithm)

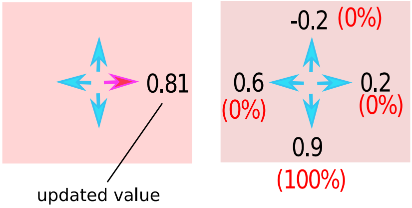
Off-policiness with uniform sampling





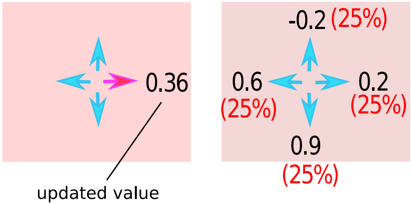
- We add a negative reward for hitting walls

Detailed mechanism: Rule 1 (Q-LEARNING)



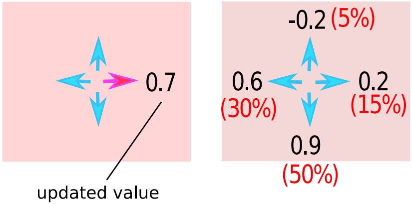
- $a' = \mathrm{argmax}\, aQ(s_{t+1}, a)$
- Always backpropagates the highest value
- Thus $Q(s, a)$ consistently converges to the value of acting optimally

Detailed mechanism: Rule 2 (SARSA)



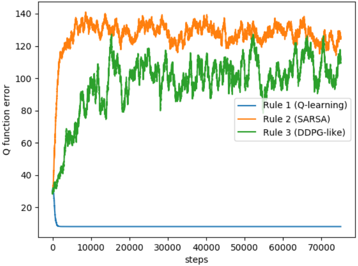
- $a' = \beta(s_{t+1})$
- Due to uniform sampling, the probabilities to take any a' are uniform
- Thus $Q(s, a)$ converges to some average, corresponding to performing random walk
- Being greedy wrt $Q(s, a)$ does not result in the optimal target policy

Detailed mechanism: Rule 3 (DDPG-like)



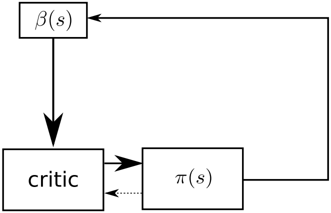
- $a' = \pi(s_{t+1})$
- $\pi(s_{t+1})$ evolve consistently with action values (update of $\alpha\delta$)
- Thus actions with higher values get sampled more often (vs uniformly in SARSA)
- If a good a' is sampled, increase the updated value
- If a bad a' is sampled, decrease it
- Results in more structured fluctuations

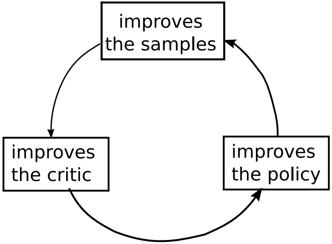
Results



- Rule 1 learns an optimal critic (thus Q-LEARNING is truly off-policy)
- Rule 2 fails (thus SARSA is not off-policy)
- Rule 3 fails too (thus an algorithm like DDPG is not truly off-policy!)
- NB: different ACTOR-CRITIC implementations behave differently
- E.g. if the critic estimates $V(s)$, then equivalent to Rule 1

Closing the loop





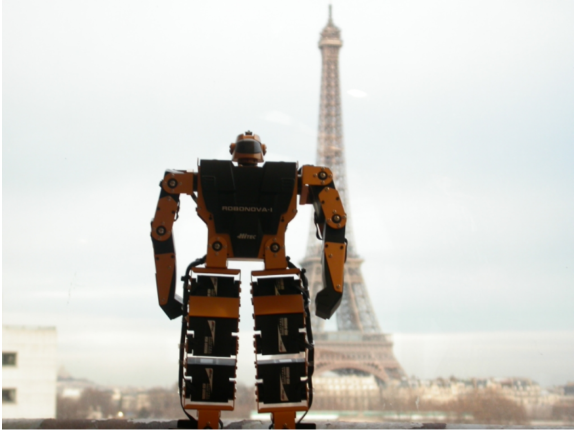
- If $\beta(s) = \pi^*(s)$, then Rules 2 and 3 are equivalent,
- Furthermore, $Q(s, a)$ will converge to $Q^*(s, a)$, and Rule 1 will be equivalent too.
- Quite obviously, Q-LEARNING still works
- SARSA and ACTOR-CRITIC work too: the $\beta(s)$ becomes "Greedy in the limit of infinite exploration" (GLIE)

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

Corresponding labs

- See https://github.com/osigaud/rl_labs_notebooks
- One notebook about off-policy versus on-policy learning
- Check the three rules convergence properties when $\beta(s)$ is uniform random sampling

Any question?



Send mail to: `Olivier.Sigaud@upmc.fr`

Maei, H. R., Szepesvári, C., Bhatnagar, S., & Sutton, R. S. (2010).
Toward off-policy learning control with function approximation.
Edité dans *ICML*, pages 719–726.

Munos, R., Stepleton, T., Harutyunyan, A., & Bellemare, M. G. (2016).
Safe and efficient off-policy reinforcement learning.
Edité dans *Advances in Neural Information Processing Systems*, pages 1054–1062.

Precup, D. (2000).
Eligibility traces for off-policy policy evaluation.
Computer Science Department Faculty Publication Series, page 80.

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