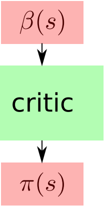
From Policy Gradient to Actor-Critic methods

On-policy versus Off-policy

Olivier Sigaud

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

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.

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

Off-policy learning: definition

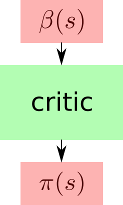
- "Off-policy learning" refers to learning about one way of behaving, called the *target policy*, from data generated by another way of selecting actions, called the *behavior policy*.
- Two notions:
  - Off-policy policy evaluation (not covered)
  - Off-policy control:
    - Whatever the behavior policy (as few assumptions as possible)
    - The target policy should be an approximation to the optimal policy
    - Ex: stochastic behavior policy, deterministic target policy

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

Why prefering off-policy to on-policy control?

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

Approach: two steps



- Open-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
  - 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

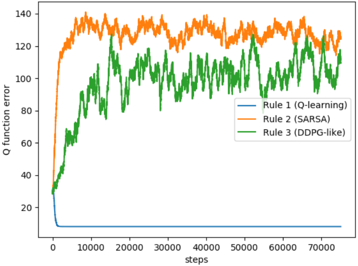
Learning a critic from samples

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

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

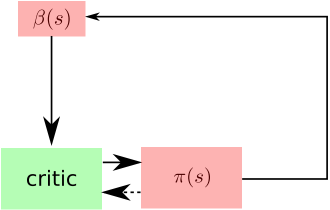
- There are three possible update rules:
  1. $a' = \mathrm{argmax}\, a Q(s_{t+1}, a)$ (corresponds to Q-LEARNING)
  2. $a' = \beta(s_{t+1})$ (corresponds to SARSA)
  3. $a' = \pi(s_{t+1})$ (corresponds e.g. to DDPG, an ACTOR-CRITIC algorithm)

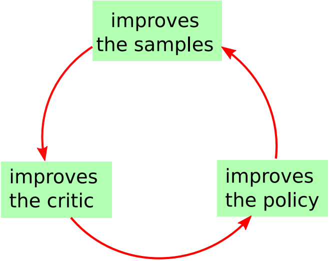
Results



- 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

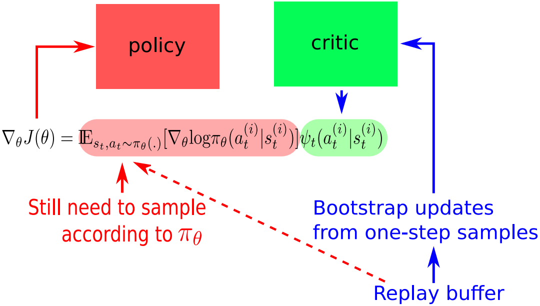
Closing the loop





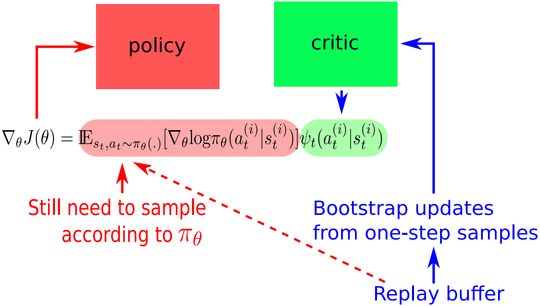
- If $\beta(s) = \pi^*(s)$, then Rules 2 and 3 are equivalent,
- Furthermore, $Q(s,a)$ will converge to $Q^*(s,a)$, and Rule 1 will be equivalent too.
- Quite obviously, Q-LEARNING still works
- SARSA and ACTOR-CRITIC work too: $\beta(s)$ becomes "Greedy in the Limit of Infinite Exploration" (GLIE)

Policy search case



- Q-LEARNING is the only truly off-policy algorithm that I know about
- With continuous action, you cannot compute $\max_a Q_\phi^\pi(\mathbf{s}_{t+1}, \mathbf{a})$
- An algorithm is more or less off-policy depending on assumptions on $\beta(\mathbf{s})$
- With a replay buffer, $\beta(\mathbf{s})$ is generally close enough to $\pi(\mathbf{s})$
- DDPG, TD3, SAC are said off-policy because they use a replay buffer

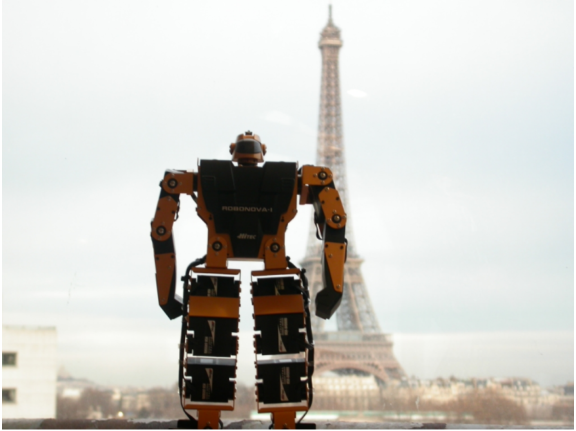
Limits to being off-policy



- DDPG, TD3, SAC use the same off-policy samples to update both the critic and the actor
- OK for the critic, not for the actor
- Does it make sense to sample differently for actor and critic?
- Yes, if several actors share one critic
- Towards offline reinforcement learning

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020

Any question?



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

ISIR
INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu.
Offline reinforcement learning: Tutorial, review, and perspectives on open problems.
arXiv preprint arXiv:2005.01643, 2020.

Hamid Reza Maei, Csaba Szepesvári, Shalabh Bhatnagar, and Richard S. Sutton.
Toward off-policy learning control with function approximation.
In *ICML*, pp. 719–726, 2010.

Satinder P. Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvári.
Convergence results for single-step on-policy reinforcement-learning algorithms.
Machine learning, 38(3):287–308, 2000.