From Policy Gradient to Actor-Critic methods

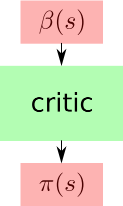
On-policy versus Off-policy

Olivier Sigaud

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

On-policy vs. off-policy

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: definitions

- "Off-policy learning": learning about one way of behaving, called the *target policy*, from data generated by another way of selecting actions, called the *behavior policy* (Maei et al.)
- "Off-policy data": training samples which were not generated using $\pi(s)$
- Two research topics:
  - Off-policy policy evaluation (not covered): how can we get the critic of a policy given data from another policy? (see Precup, Munos et al.)
  - Off-policy control: how can we get an optimal policy by training a policy given off-policy data?
- Ex: stochastic behavior policy, deterministic target policy.
- Training data can be more or less off-policy (close to data from $\pi(s)$)
- An algo. is said off-policy if it reaches the optimal policy using off-policy data.

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

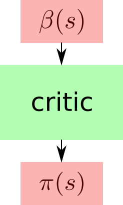
Precup, D. (2000) Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*

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

Why preferring 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

An illustrative study: two steps



- Step 1: 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
- Step 2: Close the loop:
  - Use the target policy + some exploration as behavior policy
  - If the target policy gets good, bias more towards good samples

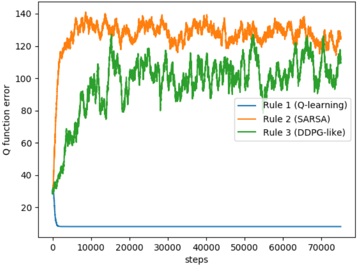
Learning a critic from samples

- We compare 3 algorithms: Q-LEARNING, SARSA, and a DDPG-like ACTOR-CRITIC
- The algorithms learn from uniformly generated samples
- Using a general format of samples S: $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_{t+1}, \mathbf{s}_{t+1}, \mathbf{a}')$ provides a unifying framework
- Makes it possible to apply a general update rule:

$$Q(\mathbf{s}_t, \mathbf{a}_t) \leftarrow Q(\mathbf{s}_t, \mathbf{a}_t) + \alpha[\mathbf{r}_{t+1} + \gamma Q(\mathbf{s}_{t+1}, \mathbf{a}') - Q(\mathbf{s}_t, \mathbf{a}_t)]$$

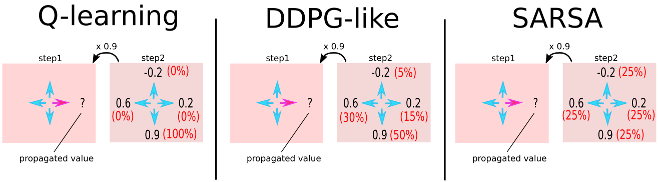
- There are three possible update rules:
  1. $a' = \text{argmax}\, a Q(\mathbf{s}_{t+1}, \mathbf{a})$ (corresponds to Q-LEARNING)
  2. $a' = \beta(\mathbf{s}_{t+1})$ (corresponds to SARSA)
  3. $a' = \pi(\mathbf{s}_{t+1})$ (corresponds e.g. to DDPG, an ACTOR-CRITIC algorithm)

Results



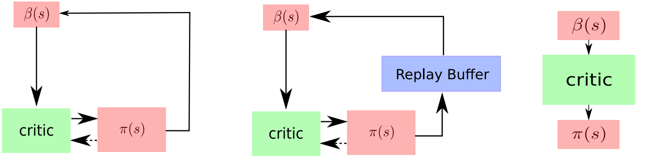
- Rule 1 learns an optimal critic (thus Q-LEARNING is truly off-policy)
- Rule 2 fails (thus SARSA is not off-policy)
- Rule 3 fails too (thus an algorithm like DDPG is not truly off-policy!)
- NB: different ACTOR-CRITIC implementations behave differently:
- If the critic estimates $V(\mathbf{s})$, ACTOR-CRITIC performs as Rule 1

Analysis



- Under uniform sampling of next action:
  - Q-LEARNING always propagates the value of the best action
  - The DDPG-like approach propagates a value depending on the current policy
  - SARSA propagates an average value

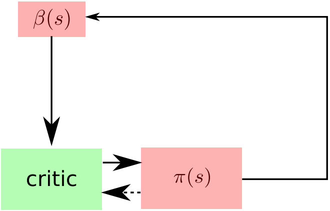
Three contexts (more details next slides)

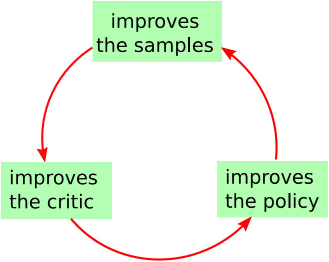


Closed-loop — Replay Buffer — Open-loop = offline

- Closed-loop case: data is on-policy
- Replay Buffer (RB) case: intermediate
- Open-loop case: offline RL

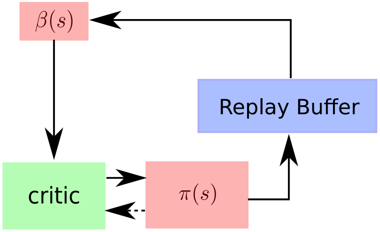
Closing the loop





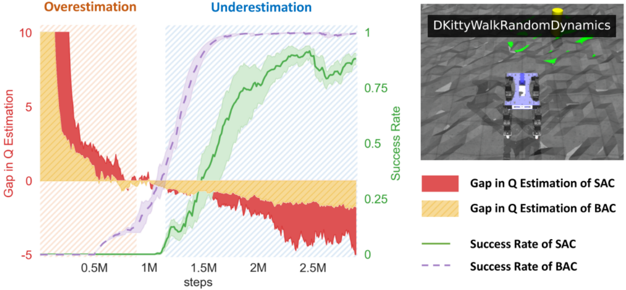
- If $\beta(\mathbf{s}) = \pi^*(\mathbf{s})$, then Rules 2 and 3 are equivalent,
- Furthermore, $Q(\mathbf{s}, \mathbf{a})$ will converge to $Q^*(\mathbf{s}, \mathbf{a})$, and Rule 1 will be equivalent too.
- Quite obviously, Q-LEARNING still works
- SARSA and ACTOR-CRITIC work too: $\beta(\mathbf{s})$ becomes "Greedy in the Limit of Infinite Exploration" (GLIE)
- In the closed-loop case, data is on-policy, on-policy algorithms can converge too.
- An on-policy algorithm can only converge if the data is on-policy.

Replay buffer case



- With a replay buffer, $\beta(\mathbf{s})$ is generally close enough to $\pi(\mathbf{s})$
- The bigger the RB, the more off-policy the data
- Being (at least partly) off-policy is a necessary condition for using a replay buffer

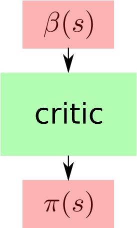
Off-policy RB algorithms: remark



- DDPG, TD3 and SAC use off-policy samples to update the critic
- To udpate the actor, they use $\delta_t = \mathbf{r}_{t+1} + \gamma \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_{t+1}, \pi_{\boldsymbol{\theta}}(\mathbf{s}_{t+1})) - \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_t, \mathbf{a}_t)$
- $\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_{t+1}, \pi_{\boldsymbol{\theta}}(\mathbf{s}_{t+1}))$ can be smaller than some $\hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\boldsymbol{\phi}}(\mathbf{s}_{t+1}, a))$ present in the replay buffer
- This can give rise to underestimation

Ji, T., Luo, Y., Sun, F., Zhan, X., Zhang, J., and Xu, H. (2023) Seizing serendipity: Exploiting the value of past success in off-policy actor-critic. *arXiv preprint arXiv:2306.02865*

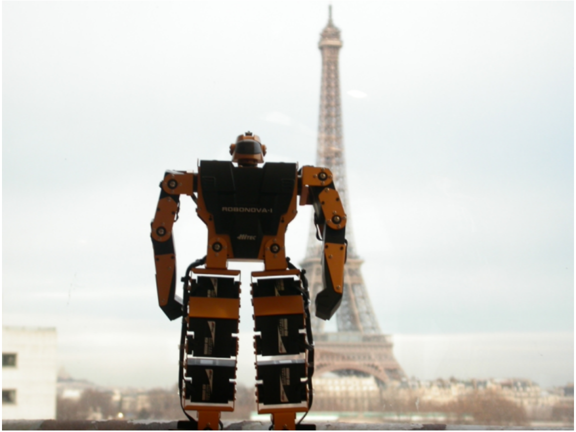
Offline RL case



- Q-LEARNING is the only truly off-policy algorithm that I know about
- Offline RL: train from a dataset without adding interaction data
- Central question: find the assumptions on the data so as to guarantee the optimal behavior can be found

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@isir.upmc.fr`

Ji, T., Luo, Y., Sun, F., Zhan, X., Zhang, J., and Xu, H. (2023).
Seizing serendipity: Exploiting the value of past success in off-policy actor-critic.
arXiv preprint arXiv:2306.02865.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020).
Offline reinforcement learning: Tutorial, review, and perspectives on open problems.
arXiv preprint arXiv:2005.01643.

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

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. (2016).
Safe and efficient off-policy reinforcement learning.
In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1046–1054.

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., and Szepesvári, C. (2000).
Convergence results for single-step on-policy reinforcement-learning algorithms.
Machine learning, 38(3):287–308.