Towards Truly Off-policy Continuous Action Actor-Critic Learning

(Towards Truly Off-policy DDPG)

Olivier Sigaud

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

Being truly off-policy

- We have seen (class 5) that Q-LEARNING was truly off-policy
- The update rule is

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

- In DDPG, we rather use

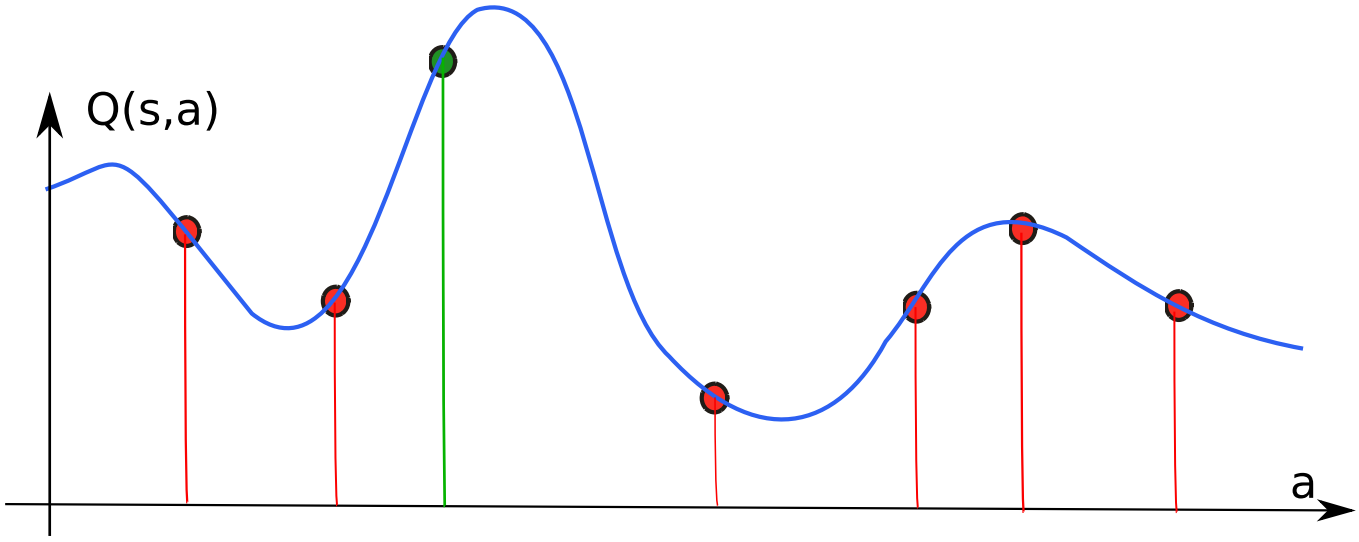
$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma Q(s_{t+1}, \pi(s_{t+1})) - Q(s_t, a_t)]$$

- This is not truly off-policy
- Over a continuous action space, finding $a' = \text{argmax}\, a Q(s_{t+1}, a)$ is intractable
- How can we be closer to doing so?

General approach

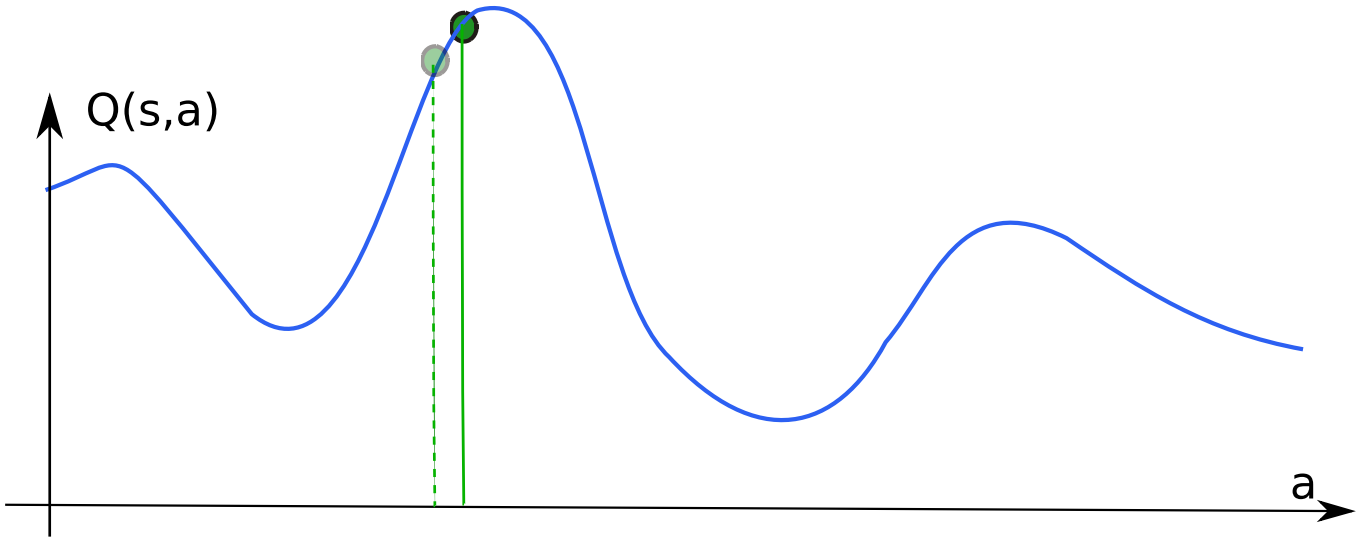
- Find a way to sample $a' \sim \operatorname{argmax} aQ(s_{t+1}, a)$ for two things:
  - To estimate $Q^*(s, a)$ rather than $Q^\pi(s, a)$, so as to be truly off-policy
  - To eventually replace the actor with an action selection which is greedy wrt the critic
- Study both points separately:
- 1. Learning the critic from uniform sampling, in an open-loop approach
- 2. Once we get a good enough estimate of $Q^*(s, a)$, we close the loop, with one of three options:
  - With a standard actor learning from the gradient of $Q^*(s, a)$
  - With an actor taken as greedy with respect to $Q^*(s, a)$ (see above)
  - With a regression-based approach (see other document)

Naive sampling approach



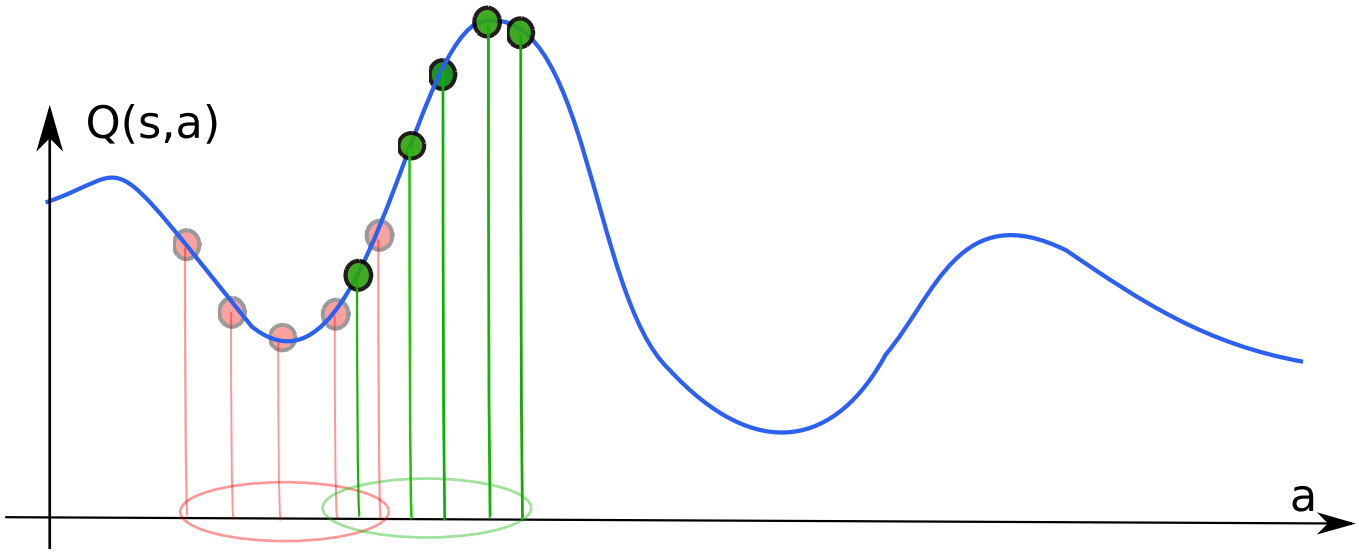
- We consider a fixed state s
- We look for $a' \sim \operatorname{argmax} aQ(s, a)$
- Most naive sampling approach: uniform sampling of action space, with K samples
- Take the best a' over the sample
- How big is K for this approach to work (as a function of the action space)?

Naive optimization approach



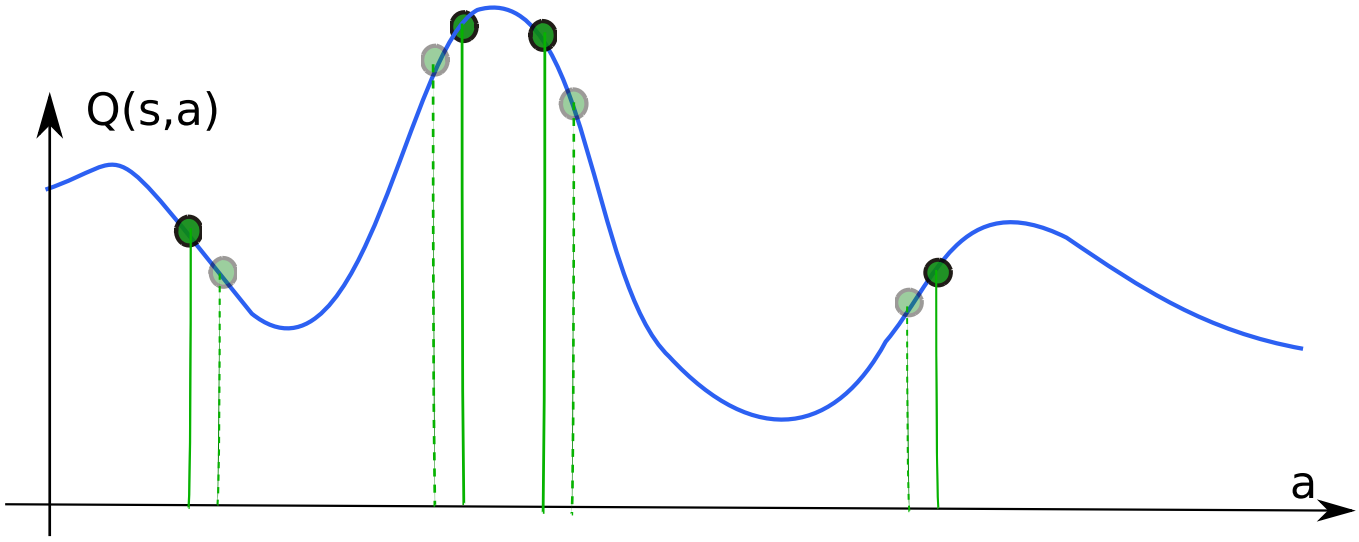
- In DDPG, we sample a unique action $a' = \pi(s)$
- This sample is good if the policy π is good
- We rely on DDPG policy improvement to sample better and better a'
- The non-naive approaches below combine both naive approaches (sampling + optimization)

Use a Cross-Entropy method



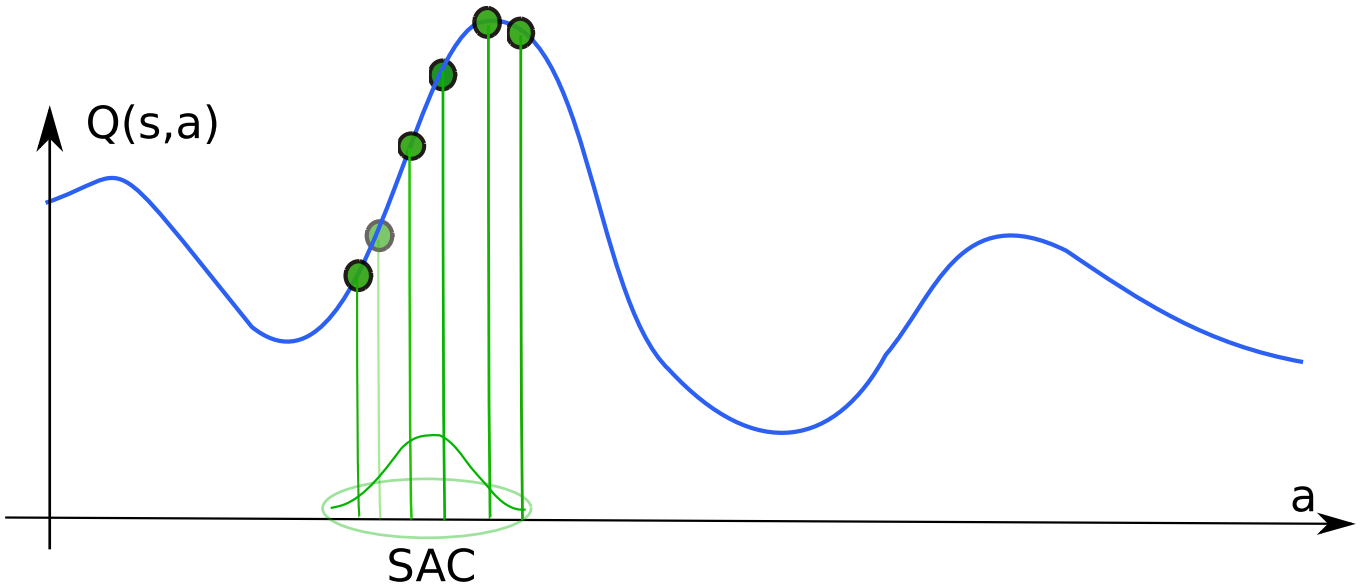
- Instead of uniform sampling, bias sampling based on the performance of the locally sampled actions
- Typical methods: CEM, CMA-ES
- Set a population and generations budget accounting for K samples

Sample K DDPG actors



- Take a DDPG with K actors (=D3PG?)
- Take the best a' over their proposed actions

Use SAC



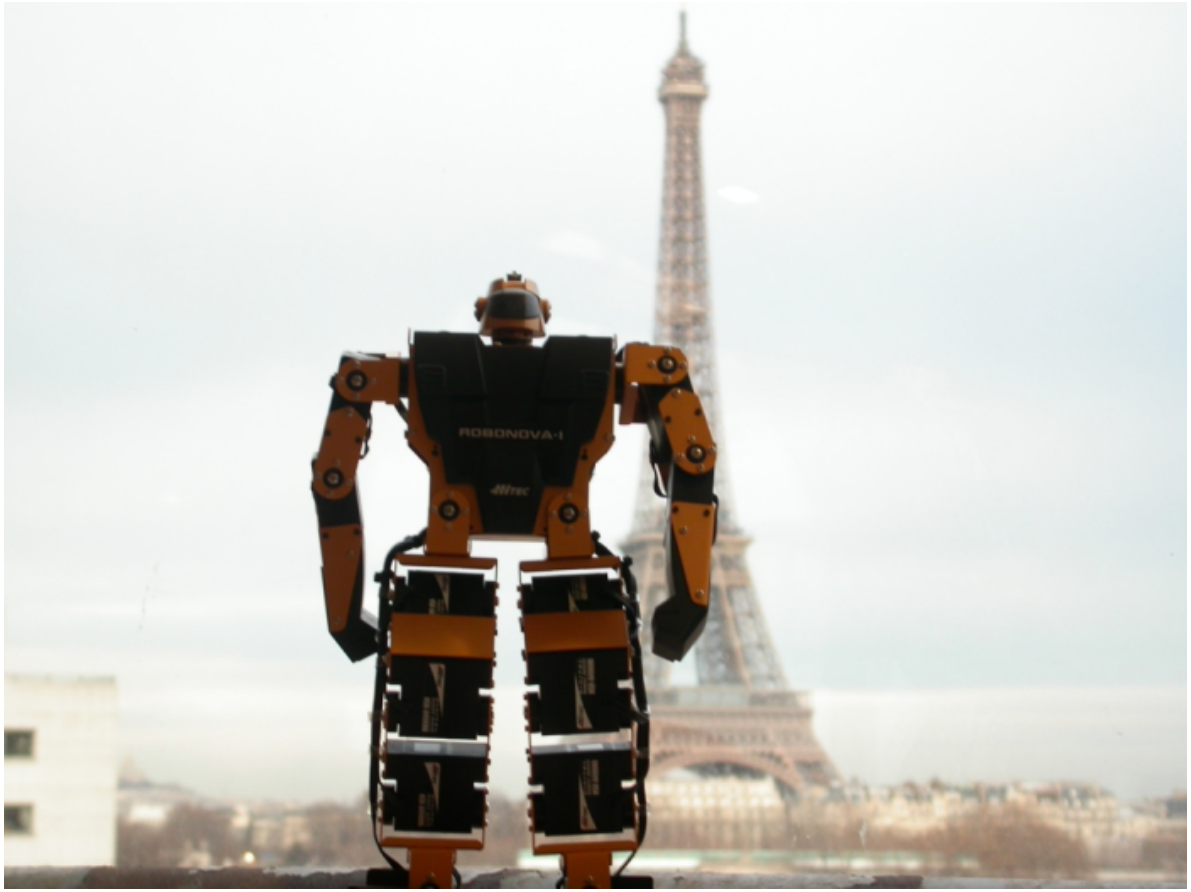
- Illustrate
- Instead of K DDPG actors, sample K actions from SAC
- Take the best a' over these samples
- Study the differences (entropy regularization, what else?)

Todo

- Take a simple continuous action benchmark where the optimal Q^* function can be determined (e.g. LQG, see Ben Recht?)
- Compare empirically the three approaches suggested above in terms of how well they estimate Q^*, using uniform sampling as training samples
- Once done, compare the three possible implementations of the actor:
  - On the simple benchmark
  - On various mujoco benchmarks

Mania, H., Guy, A., & Recht, B. (2018) Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*

Any question?



Send mail to: Olivier.Sigaud@upmc.fr

Mania, H., Guy, A., & Recht, B. (2018).

Simple random search provides a competitive approach to reinforcement learning.
arXiv preprint arXiv:1803.07055.