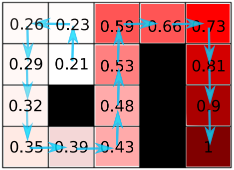
Reinforcement Learning

5bis. Off-policy policy evaluation

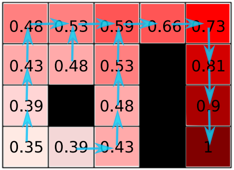
Olivier Sigaud

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

Off-policy policy evaluation: Definition



$\beta(s)$



$\pi(s)$

- Can we evaluate the critic of a target policy $\pi(s)$ from playing a different behavior policy $\beta(s)$?
- The target policy does not need to be optimal
- This is a weak notion of off-policiness
- 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

Off-policy correction: link to off-policy control

- We consider two **arbitrary** behavior $\beta(s)$ and target $\pi(s)$ policies
- We want to evaluate $\pi(s)$ from samples coming from $\beta(s)$, by correcting the samples based on the difference between $\beta(s)$ and $\pi(s)$
- The resulting critic $Q^\pi(s,a)$ will be closer to $Q^*(s,a)$ only if $\pi(s)$ is better than $\beta(s)$
- If $\beta(s)$ and $\pi(s)$ are two consecutive policies $\pi_k(s)$ and $\pi_{k+1}(s)$ from an iterative policy improvement method, applying off-policy correction only makes sense if policy improvement is **monotonous**
- In the above, I'm assuming the successive critics are used to derive the successive policies, which is not explicit in the off-policy policy evaluation setting
- General idea: applying off-policy correction can help converge to the optimal policy in an iterative policy improvement setting (perspective of TRPO and PPO)

Correction through importance sampling

- Importance sampling: given two distributions $d(x)$ and $d'(x)$
- $\mathbb{E}_d\{x\} = \mathbb{E}_{d'}\{x\frac{d(x)}{d'(x)}\}$
- Illustrate
- Explain how it applies to off-policy policy evaluation

Off-policy correction: assumptions

- To apply importance sampling to $\beta(s)$ and $\pi(s)$, we need $\beta(s)$ to be known and stochastic with non-null probabilities
- $\delta_t = r_{t+1} + \gamma \frac{\pi(s_{t+1}, a_{t+1})}{\beta(s_{t+1}, a_{t+1})} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$
- In the policy evaluation setting, we may know $\pi(s_{t+1}, a_{t+1})$ and $\beta(s_{t+1}, a_{t+1})$, but in the control setting:
  - We are looking for π^*, we generally don't know it
  - β might be an external process which we don't know (e.g. human demonstrations)

Tree backup

- The constraints on $\beta(s)$ are not realistic
- $\delta_t = r_{t+1} + \gamma \sum_{a \in A} \pi(a|s_{t+1}) Q(s_{t+1}, a) - Q(s_t, a_t)$
- Tree backup: different formulation remove the constraints
- Note: in Q-LEARNING, $\sum_{a \in A} \pi(a|s_{t+1}) Q(s_{t+1}, a) = max_a Q(s_{t+1}, a)$, thus 1-step Q-LEARNING does not need off-policy correction
- We still need to know about π, does not apply to the control setting
- Retrace: improvement over Tree Backup, applies to control, but constraints again...

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

Retrace

- Retrace: improvement over Tree Backup, applies to control, but constraints again...

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

Reactor

- On-policy: using samples from the target policy
- Off-policy: using samples from any behaviour policy
- Can the $\beta - LOO$ policy gradient in Reactor be applied to the continuous action case?

Gruslys, A., Azar, M. G., Bellemare, M. G., & Munos, R. (2017) The reactor: A sample-efficient actor-critic architecture. *arXiv preprint arXiv:1704.04651*

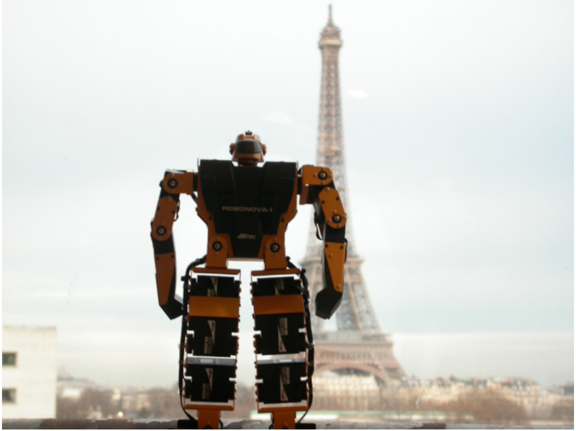
Summary

- Table from Matthieu Zimmer

TODO

- Explain why Q-LEARNING and DQN do not need off-policy correction: they are truly off-policy
- Explain why some n-step return schemes need it, and some don't

Any question?



Send mail to: Olivier.Sigaud@upmc.fr