Wrap-up, Take Home Messages

Olivier Sigaud

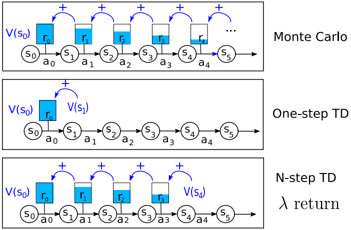
Sorbonne Université
http://people.isir.upmc.fr/sigaud

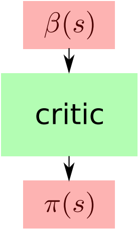
Wrap-up

Key Policy Gradient Steps

- 1. Splitting the trajectory into steps: Markov Hypothesis required
- Key difference to Direct Policy Search methods
- Makes it possible to optimize trajectories using a gradient over policy params
- 2. Introducing the Q function
- Makes it possible to perform policy updates from a single step
- Opens the way to the replay buffer, critic networks, partly off-policy methods
- 3. Using baselines
- Makes it possible to reduce variance
- When learning critics from bootstrap, becomes actor-critic

Bias-variance, Being Off-policy





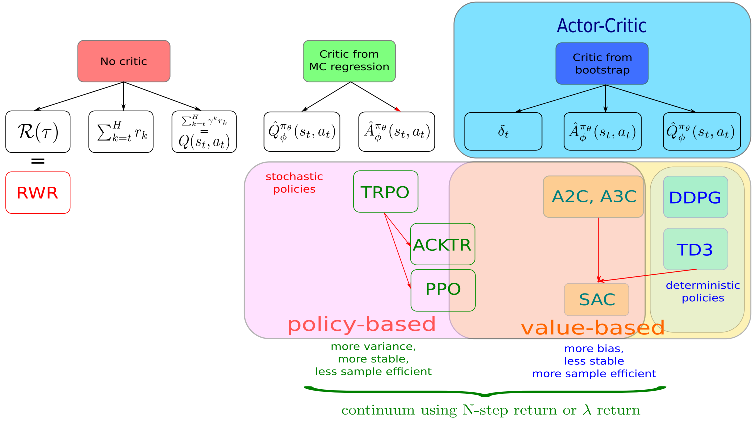
- Continuum between Monte Carlo methods and bootstrap methods
- Playing on the continuum helps finding the right bias-variance trade-off
- Being off-policy requires bootstrap
- No deep RL algorithm is truly off-policy, it's a matter of degree

The right distinction

- Off-policy versus on-policy is not so clear, as being off-policy is a matter of degree
- Being actor-critic, using a replay buffer does not lead to a clear-cut distinction (A2C blurs the classification)
- The right distinction is between value-based approaches (start from a critic) and policy-based approaches (start from the gradient on the policy)
- A2C, DQN, DDPG, TD3, SAC, TQC, DROQ are value-based
- REINFORCE, TRPO, ACKTR, PPO are policy-based

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017) Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30

Final view



Actor-Critic

No critic

$\mathcal{R}(\tau)$

$\sum_{k=t}^{H} r_k$

$\sum_{k=t}^{H} \gamma^k r_k = Q(s_t, a_t)$

=

RWR

Critic from MC regression

$\hat{Q}_{\phi}^{\pi_\theta}(s_t, a_t)$

$\hat{A}_{\phi}^{\pi_\theta}(s_t, a_t)$

Critic from bootstrap

δ_t

$\hat{A}_{\phi}^{\pi_\theta}(s_t, a_t)$

$\hat{Q}_{\phi}^{\pi_\theta}(s_t, a_t)$

stochastic policies

TRPO

ACKTR

PPO

A2C, A3C

SAC

DDPG

TD3

deterministic policies

policy-based

value-based

more variance, more stable, less sample efficient

more bias, less stable more sample efficient

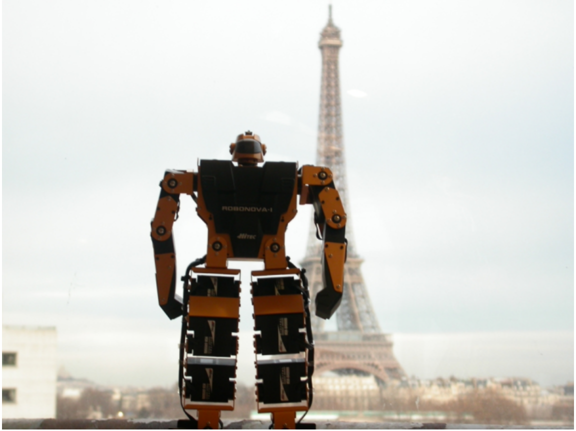
continuum using N-step return or λ return

- Even more recent: RLPD...

Chen, X., Wang, C., Zhou, Z., & Ross, K. (2021) Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*

Hiraoka, T., Imagawa, T., Hashimoto, T., Onishi, T., & Tsuruoka, Y. (2021) Dropout Q-functions for doubly efficient reinforcement learning. *arXiv preprint arXiv:2110.02034*

Any question?



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

Chen, X., Wang, C., Zhou, Z., and Ross, K. (2021).
Randomized ensembled double Q-learning: Learning fast without a model.
arXiv preprint arXiv:2101.05982.

Hiraoka, T., Imagawa, T., Hashimoto, T., Onishi, T., and Tsuruoka, Y. (2021).
Dropout Q-functions for doubly efficient reinforcement learning.
arXiv preprint arXiv:2110.02034.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017).
Bridging the gap between value and policy based reinforcement learning.
Advances in neural information processing systems, 30.