From Policy Gradient to Actor-Critic methods

Wrap-up, Take Home Messages

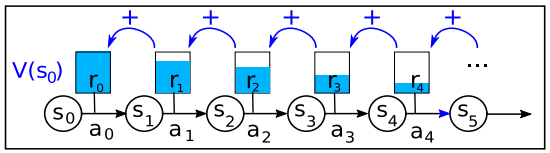
Olivier Sigaud

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

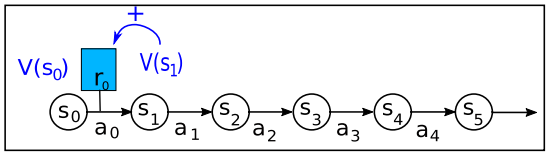
Key Policy Gradient Steps

- 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

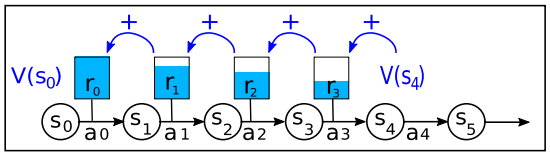
Bias-variance, Being Off-policy



Monte Carlo

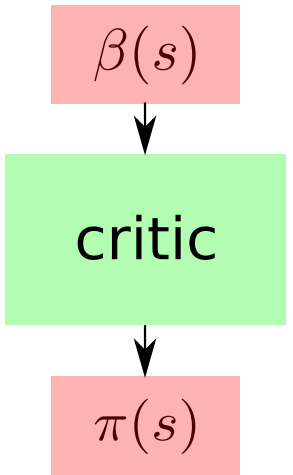


One-step TD



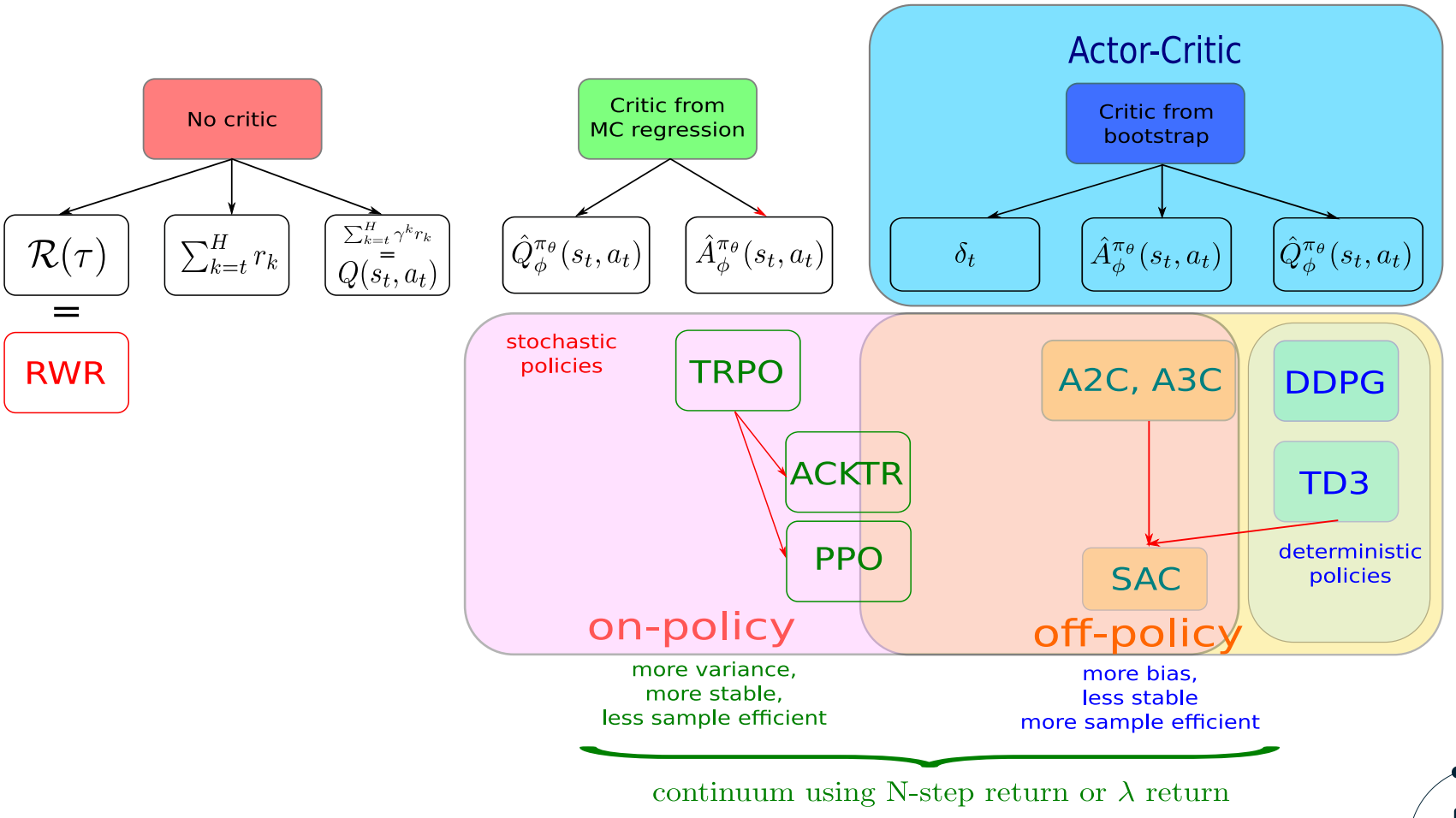
N-step TD

λ return

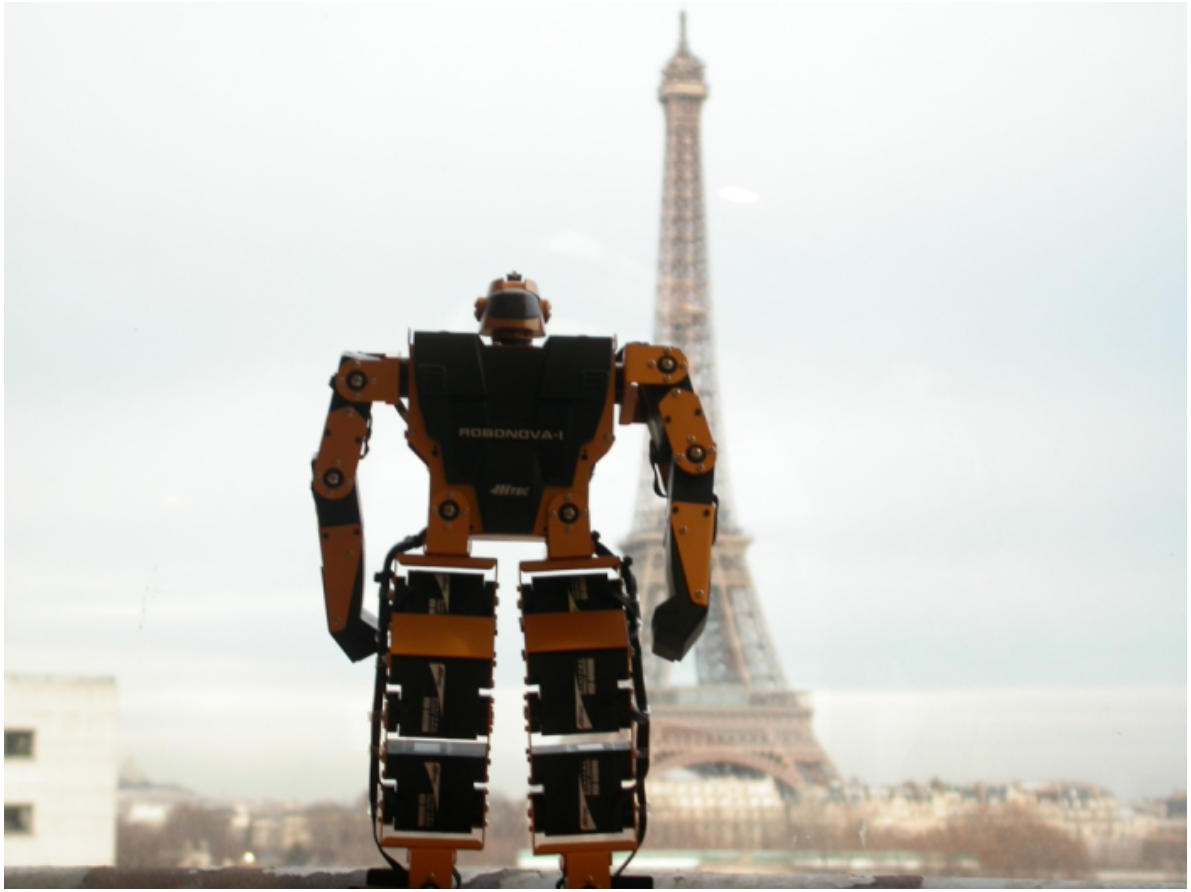


- Continuum between Monte Carlo methods and bootstrap methods
- Playing on the continuum helps finding the right bias-variance trade-off
- Being off-policy requires bootstrap
- No deep RL algorithm is truly off-policy, it's a matter of degree

Final view


No critic
$\mathcal{R}(\tau)$
=
RWR
$\sum_{k=t}^{H} r_k$
$\sum_{k=t}^{H} \gamma^k r_k$
=
$Q(s_t, a_t)$
Critic from MC regression
$\hat{Q}_\phi^{\pi_\theta}(s_t, a_t)$
$\hat{A}_\phi^{\pi_\theta}(s_t, a_t)$
Actor-Critic
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
on-policy
off-policy
more variance, more stable, less sample efficient
more bias, less stable more sample efficient
continuum using N-step return or λ return

Any question?



Send mail to: Olivier.Sigaud@upmc.fr