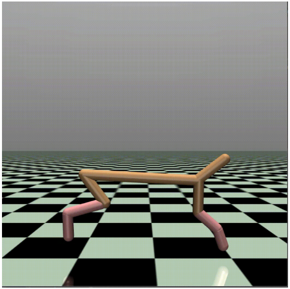
From Policy Gradient to Actor-Critic methods

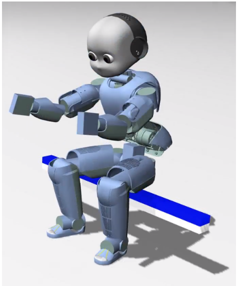
Wrap-up, Take Home Messages

Olivier Sigaud

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

Reinforcement learning over continuous actions



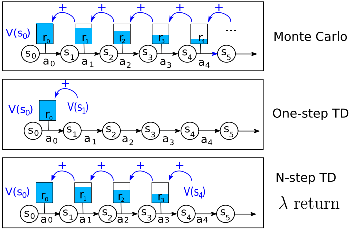


- In RL, you need a max over actions
- If the action space is continuous, this is a difficult optimization problem
- Policy gradient methods and actor-critic methods mitigate the problem by looking for a local optimum (Pontryagine methods vs Bellman methods)

Key Policy Gradient Steps

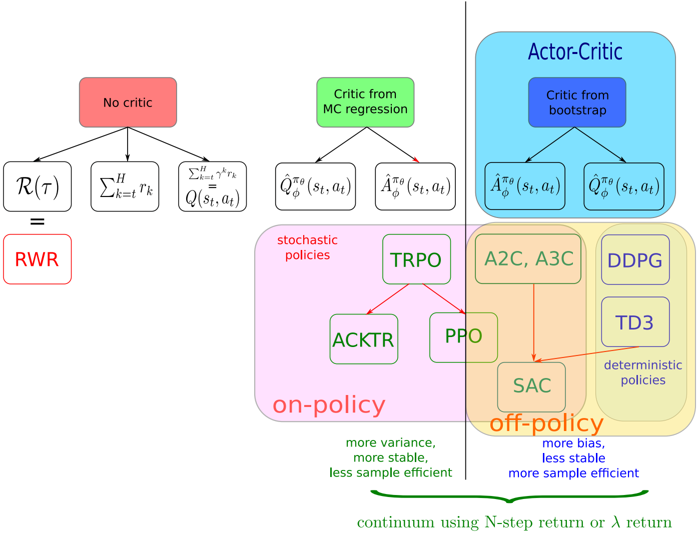
- 1. Splitting the trajectory into steps: Markov Hypothesis required
- Key difference to Direct Policy Search methods
- Makes it possible to optimize trajectories using a gradient over the policy parameters
- 2. Introducing the Q function
- Makes it possible to perform policy updates from a single step
- 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

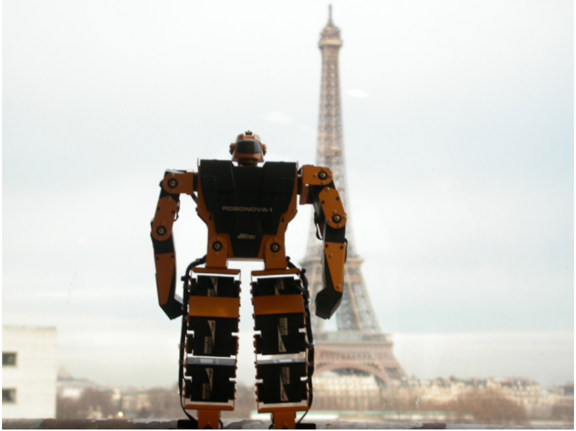


- Continuum between Monte Carlo methods and bootstrap methods
- Playing on the continuum helps finding the right bias-variance trade-off
- Being off-policy requires bootstrap
- No deep RL algorithm is truly off-policy, it's a matter of degree

Final view


Actor-Critic
No critic
Critic from
MC regression
Critic from
bootstrap
$\mathcal{R}(\tau)$
$\sum_{k=t}^{H} r_k$
$\sum_{k=t}^{H} \gamma^k r_k$
$=$
$Q(s_t, a_t)$
$\hat{Q}_\phi^{\pi_\theta}(s_t, a_t)$
$\hat{A}_\phi^{\pi_\theta}(s_t, a_t)$
$\hat{A}_\phi^{\pi_\theta}(s_t, a_t)$
$\hat{Q}_\phi^{\pi_\theta}(s_t, a_t)$
=
RWR
stochastic
policies
TRPO
A2C, A3C
DDPG
ACKTR
PPO
TD3
SAC
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`

ISIR
INSTITUT
DES SYSTÈMES
INTELLIGENTS
ET DE ROBOTIQUE