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

- 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
Any question?

Send mail to: Olivier.Sigaud@upmc.fr