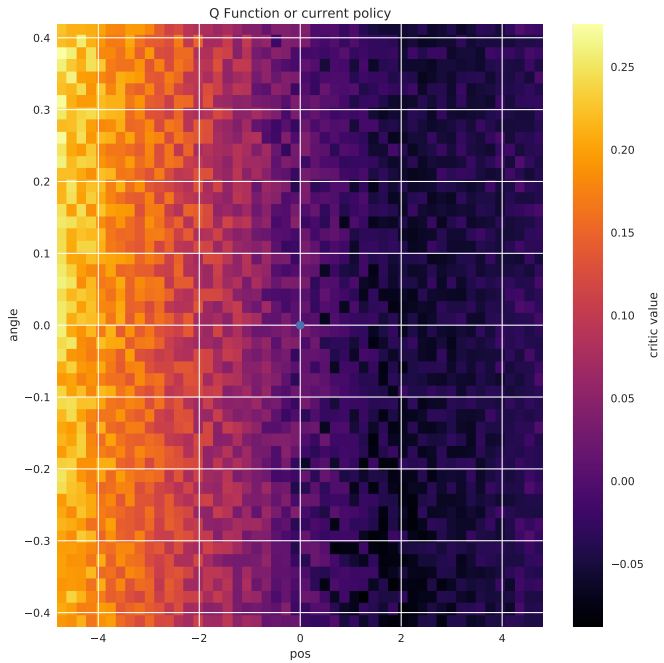
Policy Gradient in practice

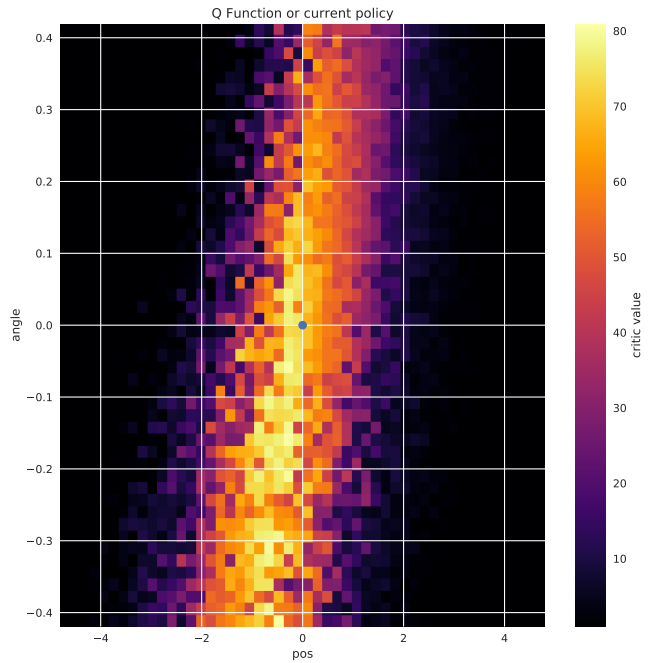
Don't become an alchemist :)

Olivier Sigaud

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

Initial/Final critic

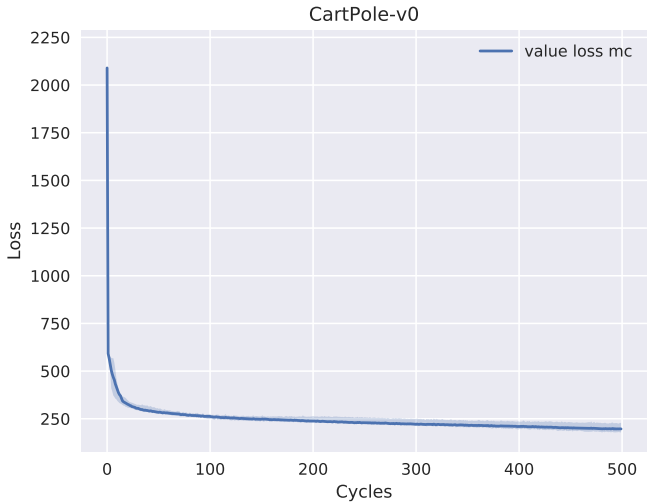


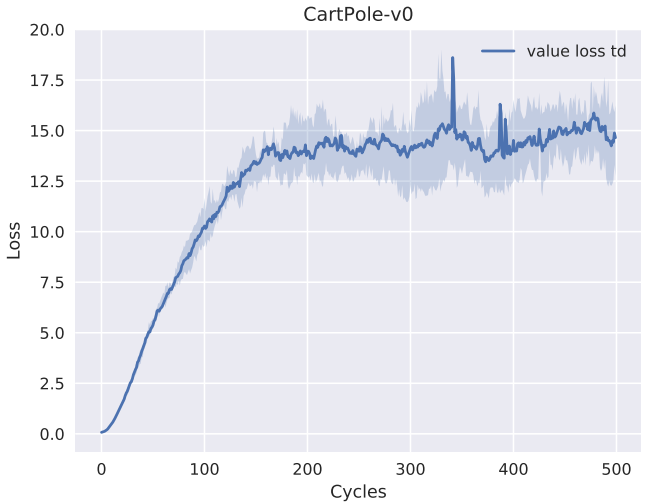


- Obtained from Bernoulli policy training and Monte Carlo evaluation method
- Batches obtained from policies along training
- General idea: it is better to be with null angle and position

MC vs TD estimation

- Obtained from Monte Carlo batches from a top policy with low variance





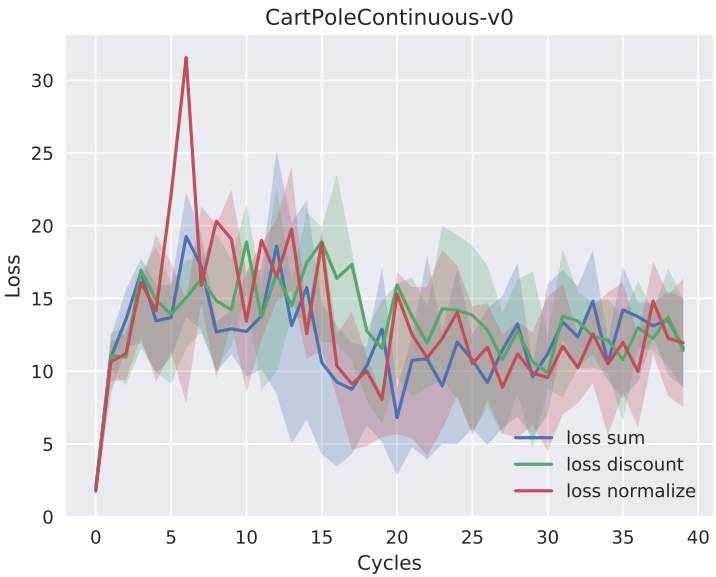
MC: $(\sum_{k=t}^{H} \gamma^{k-t} r(\mathbf{s}_k, \mathbf{a}_k) - \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\phi_j}(\mathbf{s}_t, \mathbf{a}_t))^2$

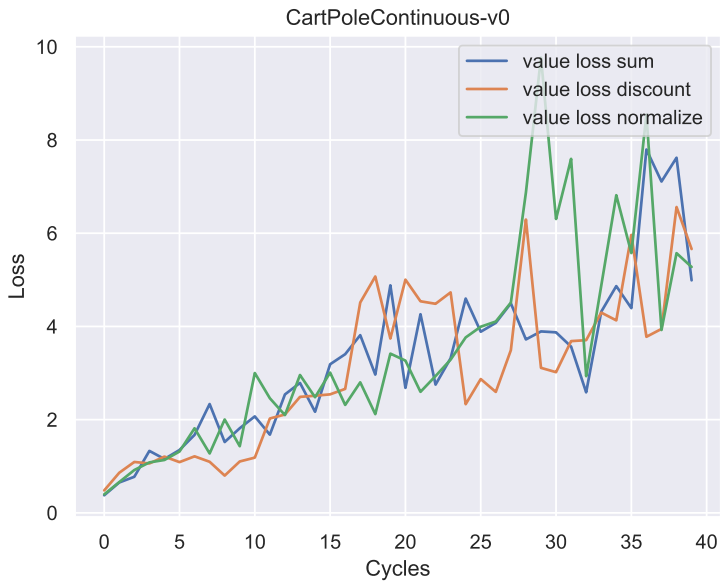
TD: $\delta_t = r + \gamma \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\phi_j}(\mathbf{s}', \pi_{\boldsymbol{\theta}}(\mathbf{s}')) - \hat{Q}^{\pi_{\boldsymbol{\theta}}}_{\phi_j}(\mathbf{s}, \mathbf{a})$

- The targets keep the same: this is a regression problem
- No need to recompute the target from the batch when the critic changes

- In the beginning, critic values are all 0
- Thus the loss are all low
- The TD error ↑, then should ↓ to 0
- Need to recompute the target at each iteration
- (or target critic)

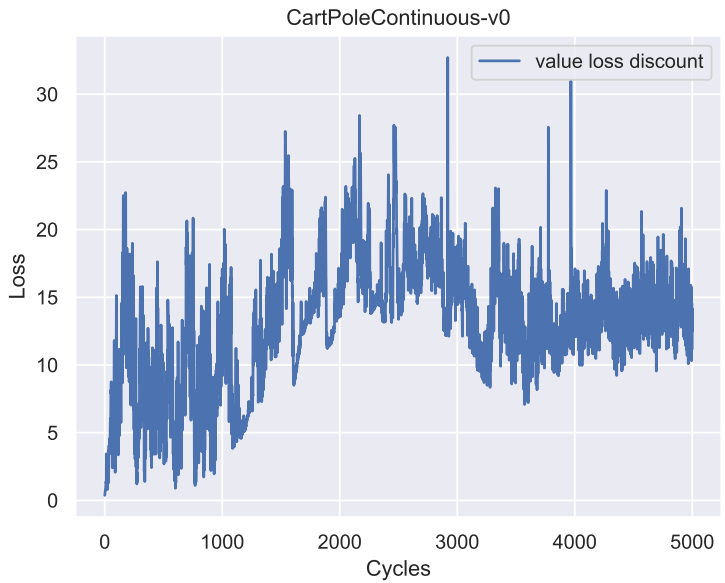
Losses of the critics





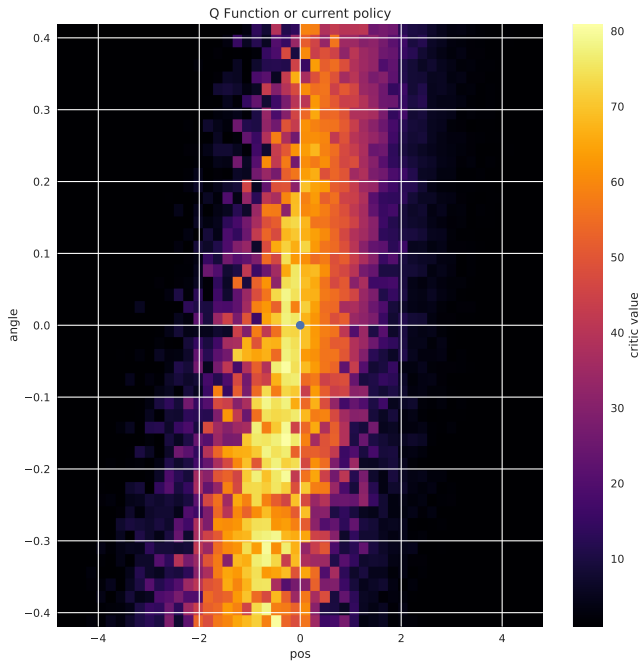
- Bernoulli (left) and Normal (right) policies
- The critic loss does not go to 0

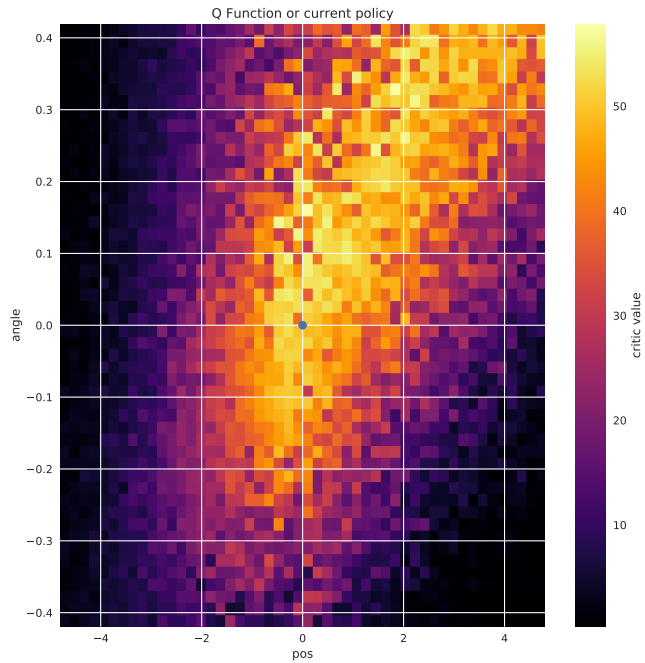
Losses of Bernoulli, longer run



- In Bernoulli policies, randomness does not go down to 0
- In Normal policies, fixed Gaussian variance
- Squashed Gaussian policy: tunable variance, but same story
- If the loss goes to 0, the policy degenerates

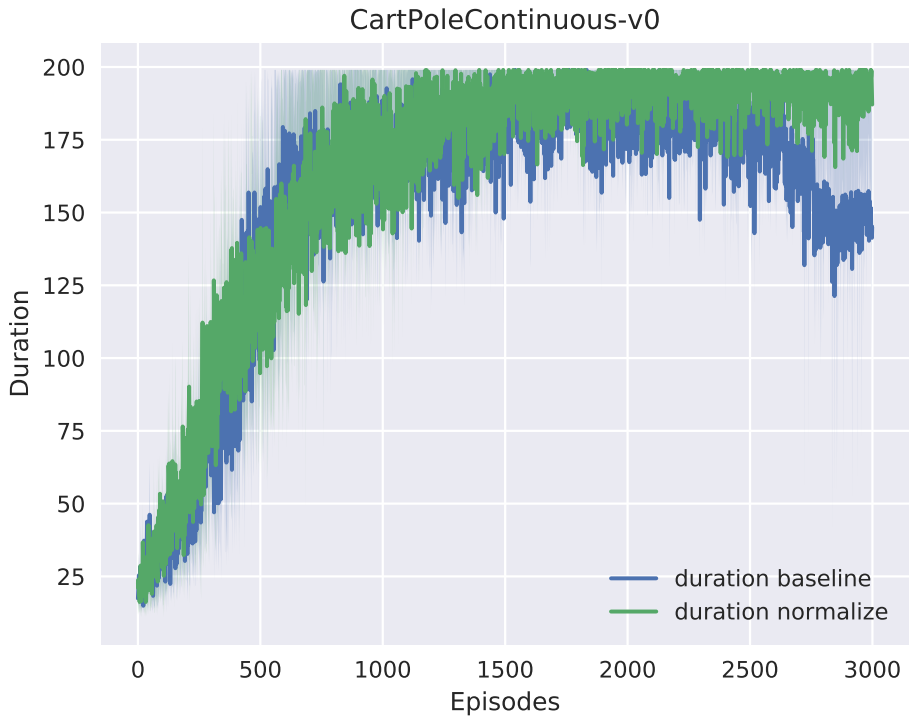
Monte Carlo critic from optimal policy





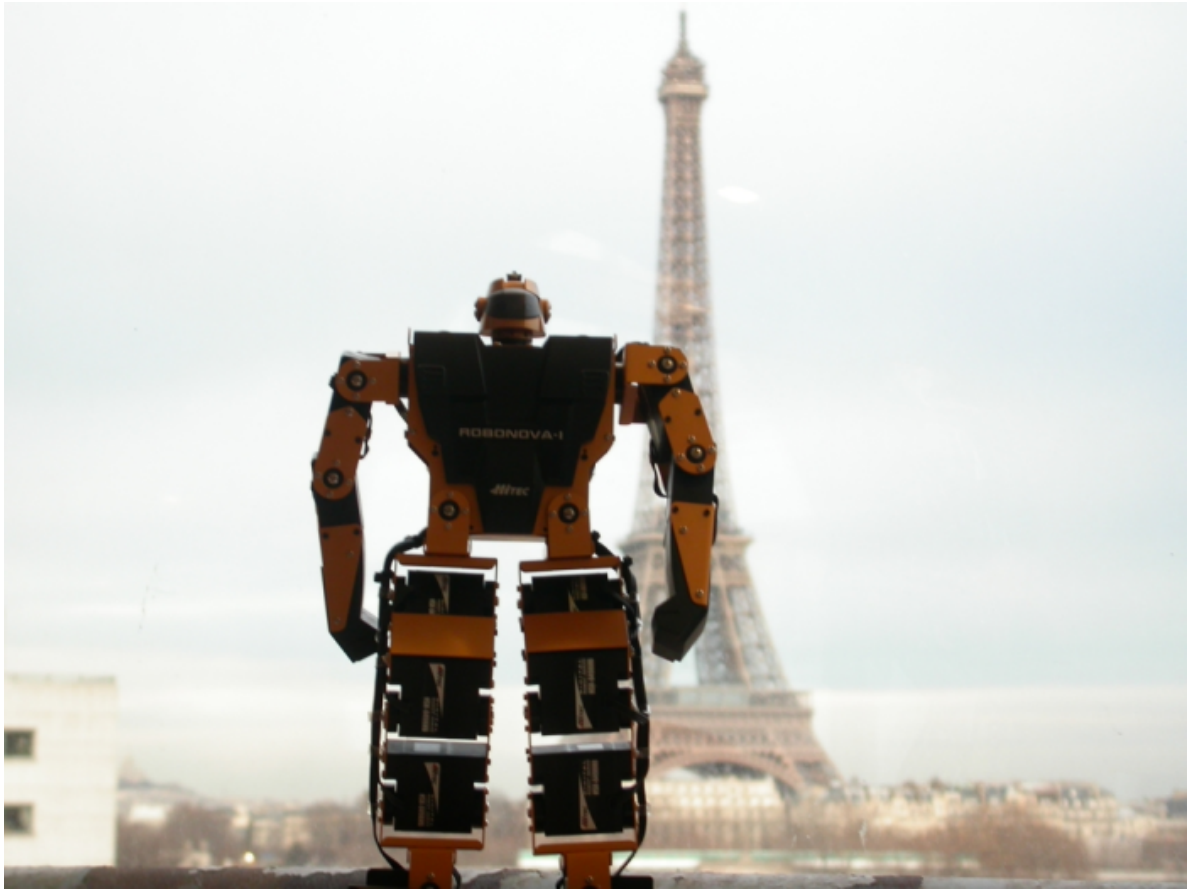
- Trained MC critic from random policy versus from top policy
- From a top policy, it does not work anymore
- Data along the same optimal trajectory: not enough exploration

Policy Gradient with critic baseline



- Learning the baseline (here, a Q-function) works well
- Until the lack of exploration results in critic degeneracy
- Sometimes, degeneracy is much more abrupt

Any question?



Send mail to: Olivier.Sigaud@upmc.fr