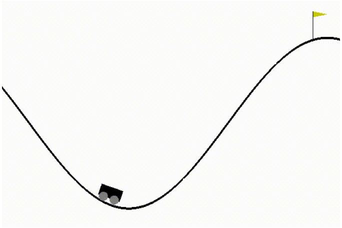
Policy Gradient in practice

Don't become an alchemist :)

Olivier Sigaud

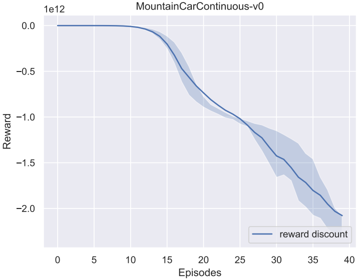
Sorbonne Université
http://people.isir.upmc.fr/sigaud

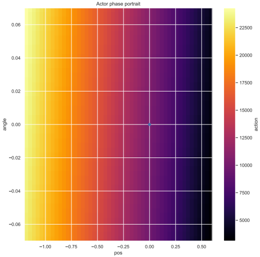
Continuous Mountain Car: Setup



- Bring the car to the flag by pushing
- Reward +100 for reaching the flag, small penalty for pushing force
- The slope is too strong for the engine
- Need to move left before going right
- A Bernoulli policy cannot find weak actions
- Deceptive gradient effect: without successful exploration, should stop moving

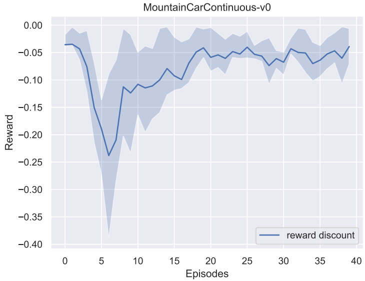
Unbounded actions

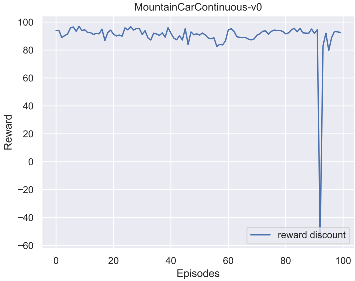




- With Gaussian policy, huge negative reward
- The action is unbounded, and goes far away from 1 (the reward considers the unbounded action)
- A squashed Gaussian policy may avoid this

Reward Normalization, Exploration Issue

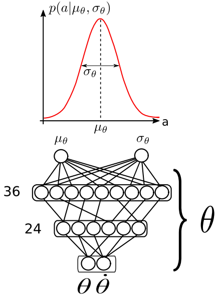




- Adding 0.05 to the reward prevents divergence
- No initial Bernoulli nor Normal policy can reach the flag
- Initialize policy with behavioral cloning: sometimes it works...
- Alternative: use more efficient exploration methods...

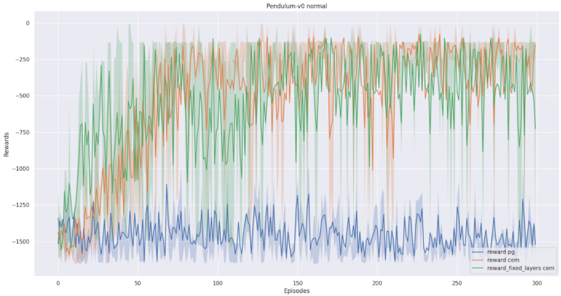
Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer (2018) GEP-PG: Decoupling exploration and exploitation in deep reinforcement learning algorithms. *arXiv preprint arXiv:1802.05054*

The Pendulum-V0 environment



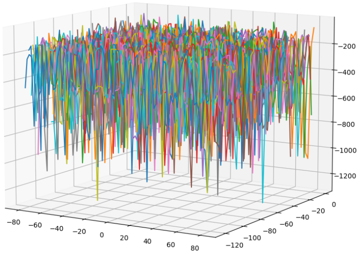
- Two state variables: θ, $\dot{\theta}$
- One continuous action (rotation torque τ)
- Reward function: $r = -\theta^2 + 0.1\dot{\theta}^2 + 0.001\tau^2 \in [0, -16.273604]$
- Studied with a Normal policy

Superiority of CEM



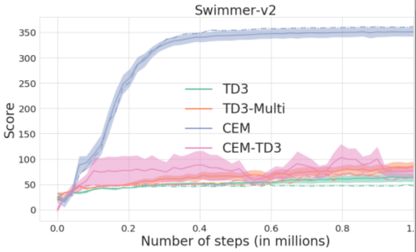
- The Cross-Entropy Method outperforms REINFORCE
- DDPG and SAC perform well too
- Key: strong variance depending on the starting point
- A large minibatch, a replay buffer and entropy help

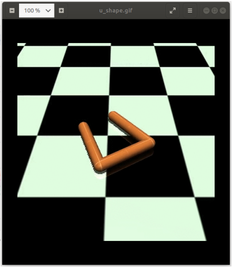
Reward landscape

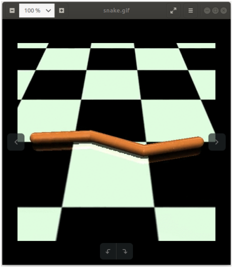


- Slightly changing policy parameters changes a lot the performance
- Not appropriate for gradient techniques

Swimmer behaviors



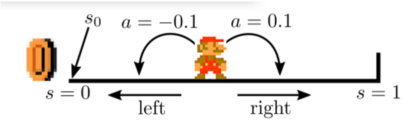




- CEM strongly outperforms all deep RL approaches (perf = 300 vs 40)
- Reward over 400 steps. $0.99^{400} = 0.01795$, $0.9999^{400} = 0.96$,
- Discounting with 0.99 favors reward over the initial time steps

Pourchot, A. & Sigaud, O. (2018) CEM-RL: Combining evolutionary and gradient-based methods for policy search. *arXiv preprint arXiv:1810.01222* (ICLR 2019)

Conclusions



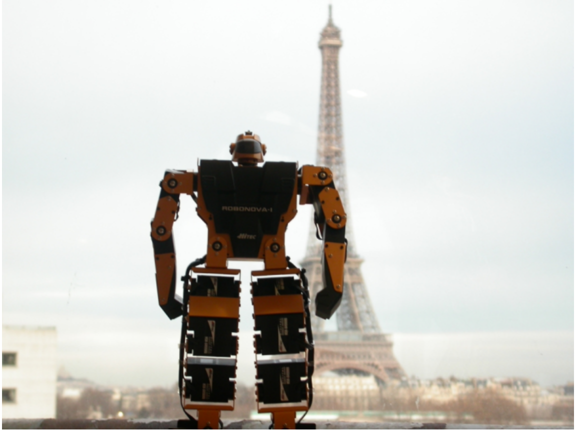
- Each environment comes with its own issues
- CartPole is the easiest gym classic control benchmark
- Basic policy gradient algorithms somewhat work after some tuning
- Making it work requires investigating and understanding phenomena
- SOTA Deep RL algorithms are more powerful, but may still fail on simplistic benchmarks

Guillaume Matheron, Nicolas Perrin, and Olivier Sigaud. (2019) The problem with DDPG: understanding failures in deterministic environments with sparse rewards. *arXiv preprint arXiv:1911.11679*

Take home message

- Science is when it does not work, but we know why
- Engineering is when it works, but we don't know why
- Continuous action RL combines science and engineering:
- It does not work, and we don't know why!

Any question?



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

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer.
GEP-PG: Decoupling exploration and exploitation in deep reinforcement learning algorithms.
arXiv preprint arXiv:1802.05054, 2018.

Guillaume Matheron, Nicolas Perrin, and Olivier Sigaud.
The problem with DDPG: understanding failures in deterministic environments with sparse rewards.
arXiv preprint arXiv:1911.11679, 2019.

Aloïs Pourchot and Olivier Sigaud.
CEM-RL: Combining evolutionary and gradient-based methods for policy search.
arXiv preprint arXiv:1810.01222, 2018.