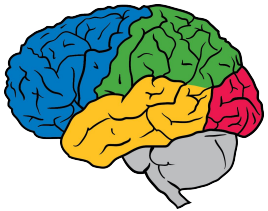
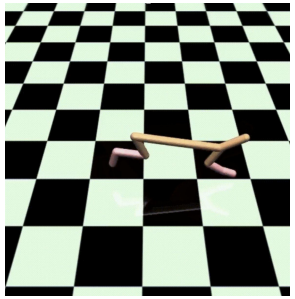
Learning from a Learner

Alexis Jacq (1,2), Matthieu Geist (1), Ana Paiva (2), Olivier Pietquin (1)
1 Google Research, Brain team
2 Instituto Superior Tecnico, University of Lisbon



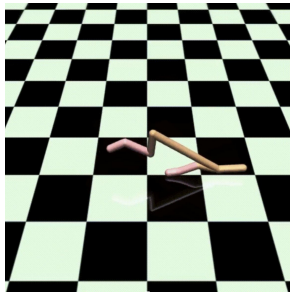
TÉCNICO LISBOA

Goal: You want to learn an optimal behaviour by watching others learning



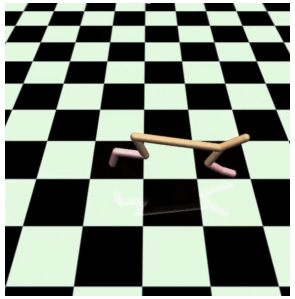
t=20

Learner
improvements



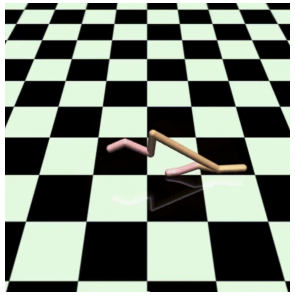
t=0

Goal: You want to learn an optimal behaviour by watching others learning



t=20

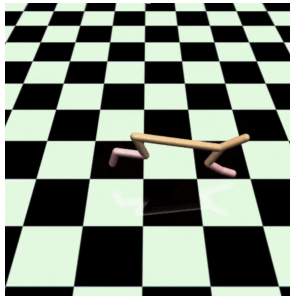
Learner



t=0

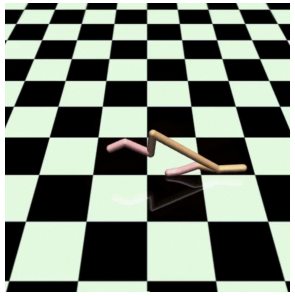
Infer rewards

Goal: You want to learn an optimal behaviour by watching others learning

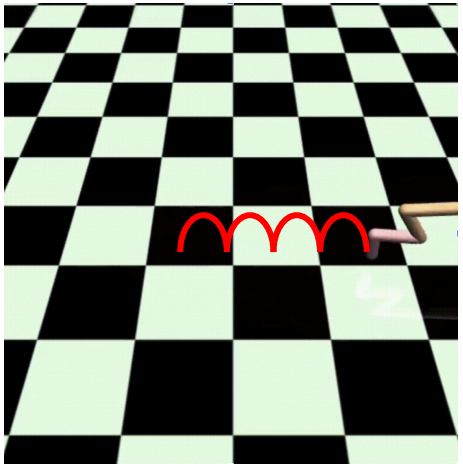


t=20

Learner



t=0



Observer
(after training with inferred reward)

Applications:

- You can observe an agent that **learns** through RL but do not see its reward
- You can observe somebody **training** but have limited access to the environment
- You were able to build **increasingly good** policies for your task but can’t tell why

Assume the learner is optimizing a regularized objective:

$$\mathcal{J}_{\text{soft}}(\pi) = \mathbb{E}_{\pi}\left[\sum_{t \geq 0} \gamma^t \left(r(s_t, a_t) + \alpha \mathcal{H}(\pi(.|s_t))\right)\right]$$

The value of a state-action couple is given by the fixed point of the (regularized) bellman equation:

$$Q^{\pi}_{\text{soft}}(s,a) = r(s,a) + \gamma \mathbb{E}_{s',a'}\left[Q^{\pi}_{\text{soft}}(s',a') - \alpha \ln \pi(a'|s')\right]$$

And one can show that the softmax:

$$\pi_2(a|s) \propto \exp\left\{\frac{Q^{\pi_1}_{\text{soft}}(s,a)}{\alpha}\right\}$$

is an improvement of the policy.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. ICML, 2018.

Given the two consecutive policies, one can recover the reward function:

$$\bar{r}_{1\to 2}(s,a) = \alpha \ln \pi_2(a|s) + \alpha\gamma \mathbb{E}_{s'}\left[\mathrm{KL}(\pi_1(.|s')\|\pi_2(.|s'))\right]$$

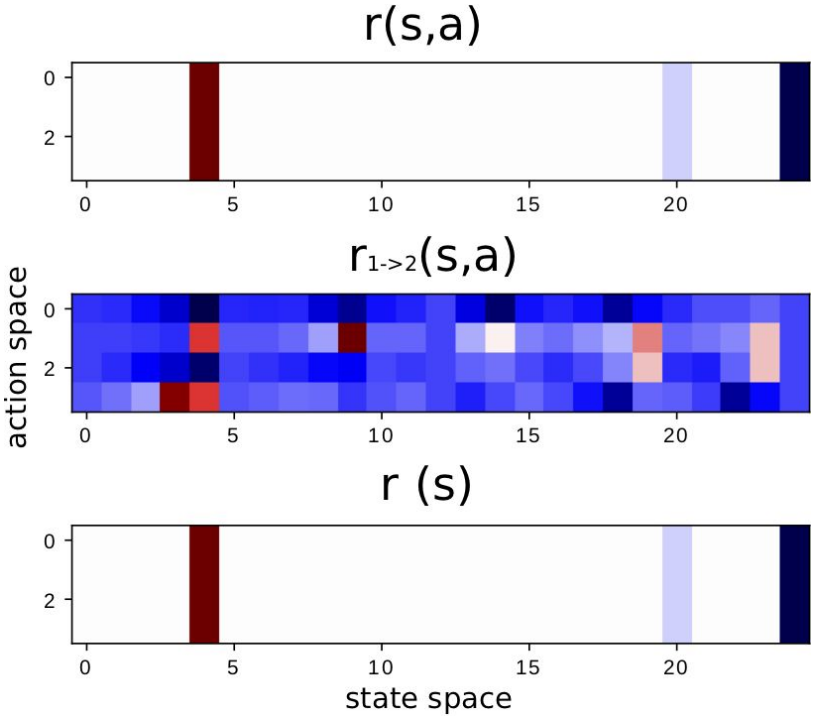
Up to a shaping that does not modify the optimal policy of the regularized Markov Decision Process:

$$\bar{r}_{1\to 2}(s,a) = r(s,a) + f_{1\to 2}(s) - \gamma \mathbb{E}_{s'}\left[f_{1\to 2}(s')\right]$$

Result with exact soft policy improvements in gridworld:

-1 Start	-1	-1	-1	-12
-1	-1	-1	-1	-1
-1	-1	-1 Reset	-1	-1
-1	-1	-1	-1	-1
0	-1	-1	-1	+10 Reset

Result with exact soft policy improvements in gridworld:

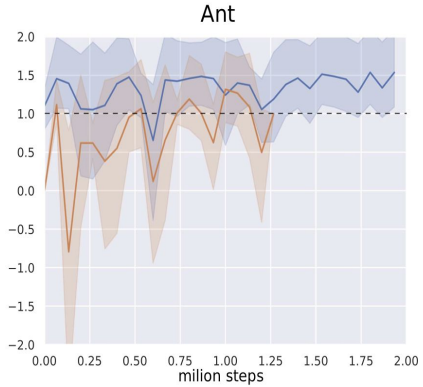


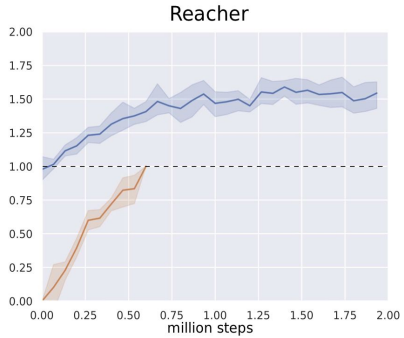
Ground truth reward.

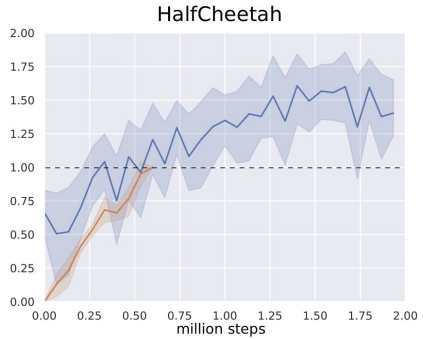
Recovered reward function by inverting soft policy improvement.

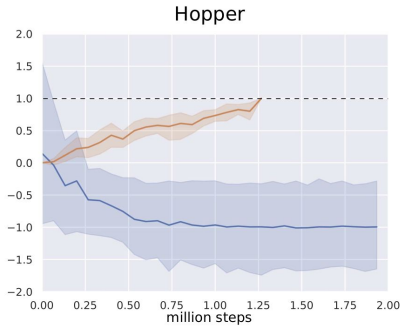
Knowing the reward is state-only.

Result with mujoco and proximal policy iterations:









(Red) Evolution of the *learner*'s score during its observed improvements.

(Blue) Evolution of the *observer*'s score when training on the same environment and using the recovered reward function.

Poster:

06:30 -- 09:00 PM Room Pacific Ballroom