Reflexive Reinforcement Learning

——— Supplementary Material ———

A Reflexive Reinforcement Learning (RRL)

A.1 Inverse Reinforcement Learning (IRL)

The self-generated rewards in reflexive reinforcement learning are expected to be self-consistently recoverable by inverse reinforcement learning, i.e. the extracted reward distribution matches the initial rewards up to scale for a known original discount factor γ. If γ is not known, further ambiguities arise as any exponentially smoothed version of the original reward distribution is also a solution, such that a fixed environment-based choice of γ is preferable.

A.2 Numerical stability

The value of γ has an obvious effect on the IRL solubility: For $\gamma \lessapprox 1$, the stability of the inverse equations drop such that the quality of any finite-time estimation of the action-induced transitions will not be sufficient to solve the IRL problem. Likewise, the numerical stability can also suffer from ambiguous reward configuration, such as the smoothening of reward distributions. This may be useful to guide an agent gradually to a particular goal, but this would require precise goal-related prior information which is not assumed in a primarily exploratory problem setting.

A.3 Noise level

The noise level is also a critical parameter in RRL. If the noise is high then a thorough exploration is possible, but less reward-related information is available to the IRL stage. For small noise, the agent will slowly converge such that the reward-related information is not precise. The goal of RRL is to maximise the reward-related information flow.

A.4 Maximum entropy IRL

A single rewarded state would be easiest in the sense of numerical stability, but the success of maximum entropy approaches [6] can serve as an argument for more general reward distributions such that the maximum entropy assumption is a-priorily correct. In adversarial scenarios [1] can provide the guidance necessary to generate the primal exploration schemes.

A.5 Markovianity

Markovianity can be guaranteed in the trivial case of information states, i.e. if all previous state-actions pairs are included in an exponentially large state space. In general, the requirement of Markovianity restricts somewhat the generality of RRL, and obviously if the primal RL does not converge because of the non-Markovianity of the dynamics, then also IRL is bound to fail. This does not mean that a meaningful application of RRL always presupposes Markovianity. Indeed, RRL can be used to reduce the adverse effects of non-Markovianity, which proves advantageous, e.g., if the agent does not have full information about the state. In this case, the reward can be used to encourage state transitions that have observable consequences and that can be mapped to unique internal states which clearly helps to recover the reward signal in IRL.

A.6 Continuous States and Actions

Introduction of a representation of policy and value function in terms of suitable chosen basis functions can improve the convergence rate, which holds both for primal RL and inverse RL. While relations between the representations of policy and value function have been studied [2, 5] the relation between RL and IRL is less understood, although usually the same representation can be used in either case. Thus, if the functional representation is known, then no additional problem arises in RRL in this respect.

In RRL the choice of a system of basis functions is obvious if taken jointly with the decision on the primal reward distribution: A function system is preferable if it represents perfectly the value function arising from the primal rewards. In addition, sparsity and low computational complexity are clearly beneficial.

Using and finding symmetries [3, 4] is one of the goals of RRL. Note that the reduced sensing scenario that we have employed in the simulations already supports the identification of symmetries.

B Pseudocode

This section contains the pseudocode for a basic RRL algorithm. For more general approaches to RRL see the Discussion of the main paper.

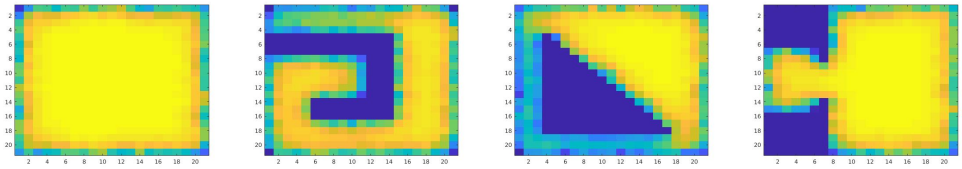
Procedure 1 Reflexive Reinforcement Learning

Input: Reflexive policy $\pi_{\theta_{\hat{r}}}(a|x) = p(a|x, \theta_{\hat{r}})$, with initial parameters $\theta_{\hat{r}} = \theta_{\hat{r}_0}$, its derivative $\nabla_{\theta_{\hat{r}}} log\pi(a|x)$ and basis functions $\phi_{\hat{r}}(x)$ for the value function $V^{\pi_{\hat{r}}}(x)$. Task related policy $\pi_{\theta_r}(a|x) = p(a|x, \theta_r)$, with initial parameters $\theta_r = \theta_{r_0}$, its derivative $\nabla_{\theta_r} log\pi(a|x)$ and basis functions $\phi_r(x)$ for the value function $V^{\pi_r}(x)$.

```
while e < E do
  Draw starting state x_0 ~ p(x)
  while t < T do
    if No task related information is present then
      Draw action a_t ~ π_θr̂(a|x)    (RRL)
    else
      Draw action a_t ~ π_θr(a|x)    (RL)
    end if
    Observe next state x_{t+1} ~ p(x_{t+1}|x_t, a_t)
    Reflect: Observe rewards r̂_t and r_t
    θ_r̂ ← α∇_θr̂ logπ(a|x)
    θ_r ← α∇_θr logπ(a|x)
    t ← t + 1
  end while
  e ← e + 1
end while
```

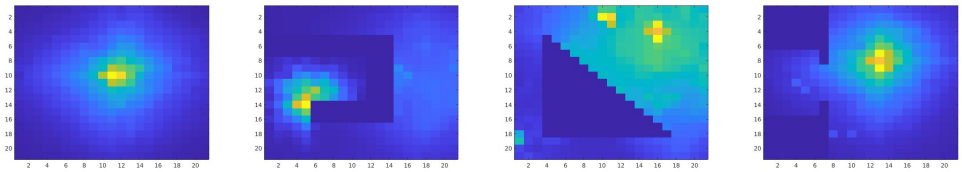
C Additional Figures

In this section you will find the discounted entropy figures left out of the main body of the paper in Fig. 1, and an additional example of the multi agent case in Fig. 2.



(a) (b) (c) (d)

Figure 1: Value function of the various arenas for the agent aiming to purely maximise γ-empowerment. **(a)** is an empty arena. **(b)** is a snaking obstacle. **(c)** has a triangular obstacle with two corridors. **(d)** is an arena consisting of two rooms, where the agent initialises in the smaller room. Here and in the following, agents were trained for 10^7 episodes each of a length of twice the size of the environment with discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.1$. For the single agent experiments the exploration rate was $\varepsilon = 0.75$. Yellow (blue) colour correspond to maximal (minimal) value at an arbitrary scale which is implied by the IRL scenario. All values are non-negative. Inaccessible regions of the environment are assumed to have zero value.



(a) (b) (c) (d)

Figure 2: Value function of the various arenas for the agents aiming to attempting to socialise. Agents are only rewarded for both choosing the action to greet the other within a Manhattan distance of 3. **(a)** is an empty arena. **(b)** is a snaking obstacle. **(c)** has a triangular obstacle with two corridors. **(d)** is an arena consisting of two rooms, where the agent initialises in the smaller room. Here and in the following, agents were trained for 10^7 episodes each of a length of twice the size of the environment with discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.1$, with $\varepsilon = 0.5$. Yellow (blue) colour correspond to maximal (minimal) value at an arbitrary scale which is implied by the IRL scenario. All values are non-negative. Inaccessible regions of the environment are assumed to have zero value.

References

[1] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016.

[2] Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

[3] Anuj Mahajan and Theja Tulabandhula. Symmetry learning for function approximation in reinforcement learning. *arXiv preprint 1706.02999*, 2017.

[4] B. Ravindran and A. G. Barto. Symmetries and model minimization in markov decision processes. Technical report, University of Massachusetts, Computer and Information Science Dept. Graduate Research Center Amherst, MA, USA, 2001.

[5] Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

[6] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proc. 23*rd *AAAI Conference on Artificial Intelligence*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.