Reflexive Reinforcement Learning: Methods for Self-Referential Autonomous Learning

B. I. Lyons and J. Michael Herrmann
Institute of Perception, Action and Behaviour, University of Edinburgh,
10 Crichton Street, Edinburgh, EH8 9AB, U.K.

Keywords: Reinforcement Learning, Exploration-Exploitation Dilemma, Intrinsic Motivation, Self-Referential Learning, Empowerment, Autonomous Agents.

Abstract: Reinforcement learning aims at maximising an external evaluative signal over a certain time horizon. If no reward is available within the time horizon, the agent faces an autonomous learning task which can be used to explore, to gather information, and to bootstrap particular learning behaviours. We discuss here how the agent can use a current representation of the value, of its state and of the environment, in order to produce autonomous learning behaviour in the absence of a meaningful rewards. The family of methods that is introduced here is open to further development and research in the field of reflexive reinforcement learning.

1 INTRODUCTION

Experimentation can be defined as "the process of trying methods, activities etc. to discover what effect they have" (Walter, 2008), and is an essential step in the acquisition of knowledge. This learning criterion implies that for effective learning of models, there is a dichotomy present, in which an agent must act to reduce the prediction error of the model it is generating, whilst also maximising the information gain it can obtain from the environment, that is to say, a learning agent must both learn to accurately perceive the effects of its actions on the environment, but must also aim to select actions with the intent of generating the most information.

This dichotomy is another information centric representation of the *exploration-exploitation dilemma*, and has attracted researchers from multiple disciplines following the publication of J.G. March in 1991 (March, 1991), with greatly cited publications in neuroscience (Laureiro-Martínez et al., 2010), marketing (Prange and Schlegelmilch, 2009) and our own, computer science (Sutton and Barto, 1999).

In many machine learning applications we consider a given function to be optimised. In autonomous learning this is not necessarily the case. We here are interested in an agent that has no specific goal, or indeed, is unable to obtain any information related to the goal temporarily. In such an instance that a robot is lost or unable to determine the goal, it should default to an intrinsic motivation to move to an area where it can learn or be suitably located to perform a task in the future, whilst still being adaptive to its environment.

As such it is important to consider the value of a state in a way that is not tied directly to some desired goal, but instead considers the most *interesting* state as one which is most valuable. Such approaches have been discussed as alternative to reinforcement leaning, for example, in the theories of empowerment (Klyubin et al., 2005b) (see also Sect. 2.2), in homeokinesis (Der and Martius, 2012) or in Friston's programme (Friston et al., 2006) to employ the free energy to the same goal, as well as some studies that have considered the self-referential aspect alongside traditional reinforcement learning methods (Pathak et al., 2017). All of these approaches have their own weaknesses which include, immense computational cost, limitation to low-level behaviours, or conceptual relation to the explanandum, at least some of which may be a necessary cost for the gain of autonomy of the agent that relies on the resp. principle. We will follow here a slightly different path.

The methods discussed in this paper are what we here call *Reflexive Reinforcement Learning* (RRL) and each have in common that the reward in a reinforcement learning task refers reflexively to the values that are built by the algorithm based on the reward. This is problematic but interesting for autonomous learning, as the agent can use its own learn-

Lyons, B. and Herrmann, J.
Reflexive Reinforcement Learning: Methods for Self-Referential Autonomous Learning.
DOI: 10.5220/0009997503810388
In *Proceedings of the 12th International Joint Conference on Computational Intelligence (IJCCI 2020)*, pages 381-388
ISBN: 978-989-758-475-6

ing progress as a source of information. There are three conditions that need to be observed:

- The information accumulated should be meaningful, i.e., the agent used the time when no goal or target behaviour is to be followed for the acquisition of information that is likely to be useful later. This can include the prediction of state transitions, the discovery of critical states in the environment (such as a doorway), or the improvement of the consistency of the representation;
- The learning progress needs to be stable. Information that is fed-back into the system can in principle lead to divergences of the value, which needs to be avoided;
- The representation needs to remain sensitive to the introduction of any goal-related information. E.g., if the agent that got "lost", receives goal related information, the autonomous learning phase should blend in smoothly and beneficially to the standard learning task.

The rest of this paper is organised as follows: After discussion of prior relevant work in Sect. 2, we specify the reinforcement learning problem that we are going to study, the theory underpinning the robotic implementation, the algorithm that we used, as well as a description of the experimental setup that was utilised in Sect. 3. The results of the experiments are provided and analysed also in Sect. 4, and the conclusions of the work and future work are given in Sect. 5.

2 BACKGROUND

2.1 Reinforcement Learning

For the current purpose it suffices to consider the basic reinforcement learning (RL) algorithm (Sutton and Barto, 2018). Given a finite Markov decision problem (MDP), represented as a tuple, (X, x_0, A, R, P), where X is a finite set of states with start state x_0, A a finite set of actions, and R is a function that assigns to each state-action pair[1] a number r (or a random variable with mean r) which provides a direct or delayed (stochastic) evaluation of this pair.

The task is usually to maximise the expected, accumulative, discounted amount of this number which means to choose the actions such that either now, or soon, high values of r are incurred. This task can be easily achieved if a function Q is known that contains the information about the expected reward, so that a large part of RL research is related to function approximation techniques.

In this paper we use a traditional function approximation approach for state-action pairs

$$Q(x,a) = E\left[\sum_{t=0}^{\infty}\gamma^t r_t | x_0 = x, a_0 = a\right] \quad (1)$$

$$V(x) = E\left[\sum_{t=0}^{\infty}\gamma^t r_t | x_0 = x\right] \quad (2)$$

where $Q(x,a)$ is updated by (Sutton and Barto, 1999) with discount γ

$$\Delta Q(x,a) = \alpha\left(r_t + \gamma V(x_{t+1}) - Q(x,a)\right) \quad (3)$$

with learning rate α and the value is given by

$$V(x_{t+1}) = \max_a Q(x_{t+1}, a). \quad (4)$$

Because we essentially modify only the reward signal r, our approach can also be used with many other RL algorithms.

We are interested here in extracting knowledge from these values, which has been done related to temporal fluctuations, but it becomes more interesting if spatial variations are considered because in this way information directly related to the environment can be extracted, whereas temporal fluctuations mainly provide information about the learning process (Herrmann and Der, 1995).

2.2 Empowerment

The concept of *empowerment* (Klyubin et al., 2005b) is a way to express the value of a state without consideration of any goal-oriented behaviour. It can be understood as a quantification of an agent's control over its environment (Salge et al., 2014) and is a quantification of the freedom of choice of actions in combination with the level of reproducibility of an sequence of actions. The aim is not to motivate exploration from an agent, but instead to identify *preferred* states in an environment that is already known. If the agent is within the state s_t the n-step empowerment is defined based on mutual information

$$\mathfrak{E}_n(s_t) = \max_{\pi: s\to a} I\left(s_{t+n}; a_{t+n-1}, \ldots, a_t\right) \quad (5)$$

so that the task is to find a policy for which the mutual information between the next actions and the set of states is maximal.

Empowerment usually requires full prior information as well as the evaluation of all possible time series and states to determine which state or states are best for the agent to occupy over a given n-step time horizon. In this sense it is quite similar to POMDPs.

[1] or possibly to triplets: state, action and following state

In the context of RL, this level of computational complexity appears unnecessary because the precise value of $\mathfrak{E}_n$ (5) is largely irrelevant, alternatively an approximation of the empowerment can be produced iteratively by considering the entropy gain per step $\mathfrak{E}_1(s_t)$ and then summing over the time horizon specified by the RL discount factor γ.

$$\mathfrak{E}_\gamma(s_t) = \sum_{t=t_0}^{\infty} \gamma^{t-t_0} \mathfrak{E}_1(s_t) \tag{6}$$

The concept of γ-empowerment introduced in Eq. 6 is similar in information provided, but not equivalent to the original concept (5), because it does not have a crisp time horizon, because it depends in the non-asymptotic case on the behaviour of the agent, and also because it allows for different measures of the local empowerment $\mathfrak{E}_1(s_t)$ as long as the tendency of the agent to roam without restrictions is captured in an appropriate way, see Sect. 3.1.

3 METHODS

3.1 Actions and Policy

Each of the tested agents can move in any of the four cardinal directions. The agent is unable to remain in the same state, with the exception that if it attempts to move into an obstacle or wall, its position will remain unchanged. We also maintained a very high exploration rate, in the sense of an ε-greedy policy with $\varepsilon = 0.75$ such that the agent will learn to plan ahead as early as possible, because at this high level of randomness errors can often not be corrected in the next or following steps.

As our aim here is mainly that of illustration if the principle of Reflexive RL (RRL), we opted to use a box function over the entire state action space rather than a reduced number of basis functions, with a traditional ε-greedy policy; however, the approach will work in such a space.

3.2 Rewards

Here we discuss the reward functions for the agent across each of the tested approaches in order of appearance in the results section.

When prioritising the maximisation of entropy the agent was rewarded

$$R(x,a) = \begin{cases} \mathcal{H}(x,a) - 1, & \text{if collision} \\ \mathcal{H}(x,a), & \text{else} \end{cases} \tag{7}$$

where

$$\mathcal{H}(x,a) = -\sum_{x'} p(x'|x,a) \log p(x'|x,a). \tag{8}$$

When prioritising the maximisation of γ-empowerment, the agent was rewarded

$$R(x,a) = \begin{cases} -1, & \text{if } x_{t+1} = x_t \\ 0, & \text{else} \end{cases} \tag{9}$$

When prioritising the reduction of prediction error, the agent was rewarded

$$R(x,a) = \begin{cases} 1, & \text{if } x_{t+1} = x_t \\ 0, & \text{else} \end{cases} \tag{10}$$

When prioritising the visitation of corners, the agent was rewarded

$$R(x,a) = \begin{cases} 1, & \text{if } s(x_{t+1}) > 1 \\ 0, & \text{else} \end{cases} \tag{11}$$

where $s(x) \in [0,1,2]$ is the number of occupied adjacent squares in the 4-neighbourhood of the agent.

3.3 Environments

Each of the environments were selected based on the features they present. The empty arena was chosen as the base case. The second environment **(b)** presents a winding corridor which ends in a dead end, chosen to observe the effects over corridors of varying size and the effects of the surrounded end. Environment **(c)** observes the effect of large obstacles in the state space and the effects of irregular shapes on the algorithm.

The final environment, environment **(d)** consists of a smaller room and a large room, chosen to observe the effects of differently sized regions of the state space, and observe the value the agent places on these objects. As the goal of the paper was to approach similarity with the concept of empowerment, this was a useful environment for observing the agents preference for different spaces in similarity to n-step empowerment with small n.

3.4 Reflexive Reinforcement Learning

In the case of reflexive reinforcement learning, the *reflexive component* informs state valuation through standard reinforcement learning as seen in Eq. 3. This additional component allows that we are able to switch between a variety of different components for different needs, as we will show in Sect. 4.

As seen in Fig. 1, the reflexive component will receive external rewards from the environment through the observation of the state, and this adjusted reward

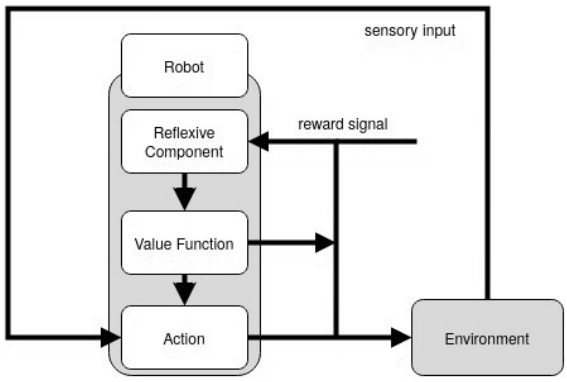

Figure 1: This diagram represents the function of the RRL algorithm as utilised here.

Algorithm 1: Reflexive Reinforcement Learning.

Require: Reflexive policy $\pi_{\theta_{\hat{r}}}(a|x) = p(a|x,\theta_{\hat{r}})$, with initial parameters $\theta_{\hat{r}} = \theta_{\hat{r}_0}$.
Require: Task related policy $\pi_{\theta_r}(a|x) = p(a|x,\theta_r)$, with initial parameters $\theta_r = \theta_{r_0}$.
  while $e < E$ **do**
    Draw starting state $x_0 \sim p(x)$
    while $t < T$ **do**
      if $r \equiv 0$ {no task specific rewards} **then**
        Draw action $a_t \sim \pi_{\theta_{\hat{r}}}(a|x)$ (RRL)
      else
        Draw action $a_t \sim \pi_{\theta_r}(a|x)$ (RL)
      end if
      Observe next state $x_{t+1} \sim p(x_{t+1}|x_t, a_t)$
      Reflect: Observe rewards $\hat{r}_t$ and r_t
      $\theta_{\hat{r}} \mathrel{+}= \alpha \nabla_{\theta_{\hat{r}}} \log \pi(a|x)$
      $\theta_r \mathrel{+}= \alpha \nabla_{\theta_r} \log \pi(a|x)$
      $t \leftarrow t+1$ {time steps within episode}
    end while
    $e \leftarrow e+1$ {continue to next episode}
  end while

is used to inform the state valuation or valuations. In this manner it is possible for the agent to continue to receive information pertinent to potential tasks as it maintains its motivation to explore the environment through the different valuations during periods of no task or when in a state where it is lost.

4 EXPERIMENTS

To observe the effects of the various maxims, each agent was run for 10^7 episodes, each being 42 time steps long. The reflexive reinforcement learning maxims we are here comparing the direct computation of entropy, and a simplified alternative for entropy we here call γ-*empowerment*. In addition to this, we also considered an approach we here refer to as *prediction error reduction*, as complement of entropy and γ-entropy, where the agent favours regions where features are visible.

4.1 Entropy

For the maximisation of entropy, we computed entropy directly and supplied this as a reward bonus to the agent as seen in Eq. 7.

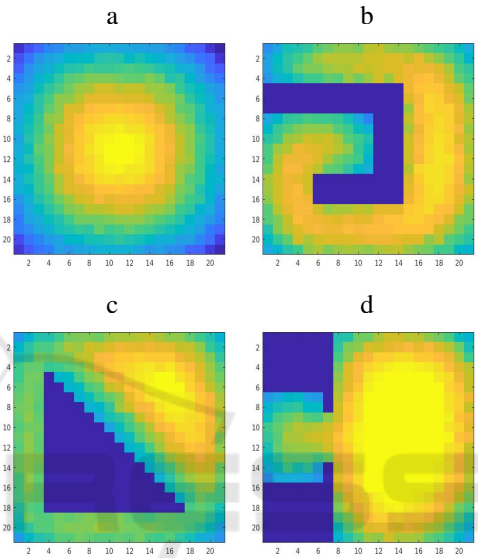

Figure 2: These colour maps represent the value of the various arenas the agent was placed in when aiming to purely maximise entropy, where there are 10^7 episodes, each being 42 time steps long, $\varepsilon = 0.75, \gamma = 0.9, \alpha = 0.1$. **(a)** is an empty arena. **(b)** is a snaking obstacle. **(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 we see that purely maximising entropy leads to increased valuations of regions far from walls and relative obstacles. In Fig. 2(**b**) we see a clear increase in valuation as the agent moves away from either of the "dead-end" regions, with the greatest value being seen in the space on the right hand side, which allows for greatest *n*-step access to the remainder of the environment.

Similarly in Fig. 2(**d**), the environment containing two different sized rooms, we observe that the greater values in the respective rooms are toward the centre, giving greater access to the remainder of the environment; however, as can be noted, the restricted region of the path between the two rooms also sees a greater valuation than other restricted regions, as this is the area that must be traversed to receive increased entropy. This is consistent with what is observed in empowerment (Klyubin et al., 2005b), par-

ticularly the cases of the mazes where an agent is in the state of greatest empowerment when it is not enclosed in walls, and over a defined n-step time horizon can actualise the greatest number of future states from the current state.

4.2 γ-Empowerment

When using γ-Empowerment as the reflexive components, we calculated the reward to be provided to the agent as in Eq. 9.

As remaining in the same state between time steps is only possible in the case of colliding with an obstacle, and the high level of epsilon makes this much more likely near corners or walls, we felt this was an appropriate quantity to consider in consort with entropy and empowerment, since under this scheme an agent should more highly value regions that provide future freedom of movement, and as opposed to the entropy case above, requires no calculation, and is easy to work with on-policy.

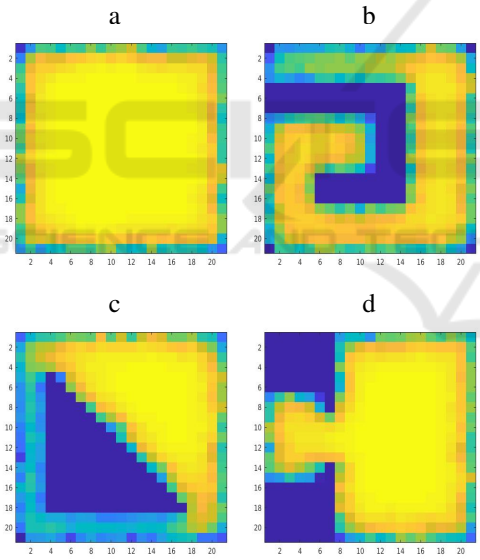

Figure 3: These colour maps represent the value of the various arenas the agent was placed in when aiming to purely maximise γ-empowerment (6), with the obstacles, episodes, episode length and parameters as in Fig. 2.

The resulting graphs can be seen in Fig. 3. Here we see similar regions of high valuation, with significantly increased value in the surrounding regions. This is consistent with what we would expect in an empowerment case, though with a substantially increased value for n in the traditional case.

We would not expect such a high value directly up to the wall regions, where in the maze variants seen in Ref. (Klyubin et al., 2005b) paper there is a smoother gradient between values, with much more distinct regions of increased empowerment, closer to what we see in the entropy case.

4.3 Prediction Error Reduction

Another approach we considered is that of an agent attempting to reduce prediction error. When an agent is attempting to find a task, it will be essential for the agent to be adequately localised to better enable the finding of a goal or regions deemed to be interesting, as such it is common in SLAM approaches to use walls and other fixed features of the environment to localise, in this approach we consider an agent's need to find such features of the environment.

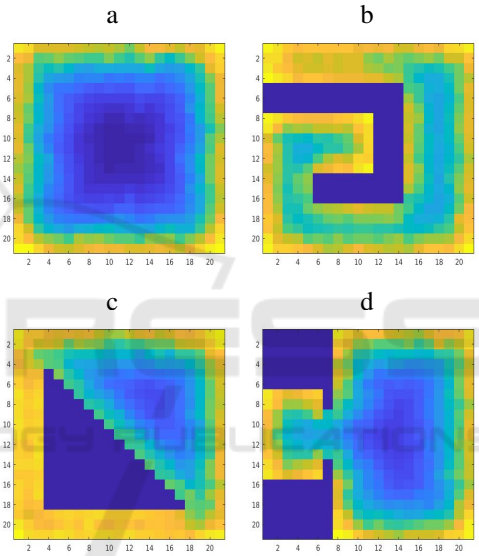

Figure 4: These colour maps represent the value of the various arenas the agent was placed in when being rewarded for obtaining sensory values of obstacles or walls in the environment, with all parameters as in Fig. 2.

By simply rewarding the agent for remaining in the same state in the following time steps, we are able to consider these external, easily referable regions of the environment with a significant increase in value. This approach has the benefit that it can be implemented alongside traditional SLAM architecture (Thrun, 2002) so as to reward moving to easily identifiable regions of the state space, and in future work we intend to implement this, with the additional rewards for correctly predicting the agents state.

4.4 Corner Favouring

Our final approach considered is that of corner favouring. When an agent is lost and attempting to find

a task in noisy environments, corners offer regions of significant information about location, and in dynamic environments these regions can typically be considered to be "out of the way". When considering the prediction error reduction case, we focused on rewarding the agent simply by sensing anything in any of the sensor directions, whereas here the focus is on rewarding multiple sensor inputs, and as such, can be considered to be moving to regions of significant information.

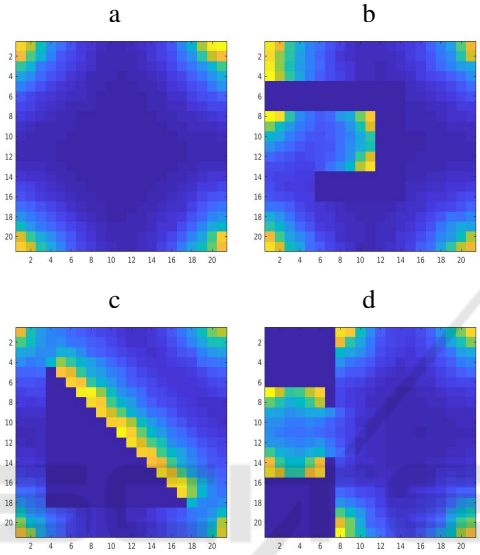

Figure 5: These colour maps represent the value of the various arenas the agent was placed in when being rewarded for obtaining multiple sensory values of obstacles or walls in the environment, with all parameters as in Fig. 2.

By rewarding the agent with a small scalar reward where multiple sensor inputs are engaged, as seen in Eq.11, the agent highly values corner regions in the state space in Fig. 5(**a**), (**b**), and (**d**). This also occurs in Fig. 5(**c**); however, here we also see a strict favouring of the wall along the diagonal region of the object, which is a feature of the discretisation of the state space.

4.5 Remarks

Rewarding an agent based on entropy leads to valuation of states which is commensurate with what one would expect in an empowerment approach, with the benefit that it can be computed on policy, without the necessity to exhaustively compute over all policies and time series.

Similarly, considering the easier to compute γ-empowerment, we are able to obtain a valuation similar to what we would expect from an empowerment approach; however, we observe that the valuation remains very similar over the open regions with no clear peak value in the environment, which may not prove as useful to the goal of intrinsic motivation over potentially dynamic environments as it shows a tendency to prefer vast regions in the state space, which will make isolating the most interesting or free regions over a complex or dynamic space much more unlikely.

Where reduction in prediction error and relocalising are key priorities for a lost agent, we can also consider employing our variants in Sect. 4.3 and Sect. 4.4 to move to regions at the edge of the state space, where there may be less task dependent information, yet the information is consistent and stable to relocalise and return to searching for task relevant information with a better understanding of where such information seen in the other variants can be found.

All of these variants here serve as a compliment to traditional reinforcement learning approaches through the use of the reflexive component, and indeed, can also be considered in tandem with one another. An intrinsically motivated, agent should seek out regions which are interesting or surprising as in sections 4.1 and 4.2, where no task relevant information is available in these identified regions of interest, the agent should return to regions where prediction error can be minimised, and relocalisation is possible, and the cycle should repeat, in a control system, perhaps after sufficient searching of the state space, the agent should move to regions which have minimal impact on a potentially dynamic environment, such as a corner, and wait to search again later.

Alternatively, we consider that if there is no task dependent information available, a continually learning agent should seek out these surprising regions, with the aim of learning more about the environment and correcting the model, by more accurately learning state transition probabilities, or learning about features present in these high interest subspaces, to better perform tasks in the future when this information becomes available.

5 DISCUSSION

5.1 Exploration Vs. Exploitation

The exploration-exploitation dilemma is not a solved problem in any of the various domains that it has been researched in. We presented here a use case for utilising entropy as a reward on its own or in conjunction with other rewards to highlight the best regions available to an agent in an environment where there is

no clear goal. In doing so we have found that these regions which are considered highly valued are similar to those found in empowerment, where the agent more highly values regions from which it is able to access a larger subset of the state space over any given discrete time frame.

As opposed to having to create sophisticated models for task location, the use of entropy maximisation may enable agents to *find* a task or location when lost in a changing environment. In future work we intend to consider the problem using actor-critic algorithms, where the actor and critic use different state-action value functions, and to employ this in a dynamic environment as well as in combination with a hierarchical model (Smith et al., 2020) alongside other functions or goals to study robotic self-motivation.

5.2 Bayesian Theory

The methods for self-supervised learning discussed here in the context of RRL have a general form as rewards in a reinforcement learning algorithm. This view has been discussed a decade ago in contrast to active inference (Friston et al., 2009), but recently it was shown that the active inference can in integrated with RL (Tschantz et al., 2020), such that it is an interesting question whether our reward-based approach can also be integrated with the Bayesian view and act as a prior for exploration in a general learning task.

Our approach can contribute also in the Bayesian context an efficient estimation of information-theoretical quantities such as shown here for entropy or empowerment. The advantage would be the state augmentation by a continuous entropy value is more naturally realisable in a Bayesian approach, whereas here it requires in principle a continuous state space in order to guarantee Markovianity and thus convergence of the RL algorithm. In continuous problems, where Markovianity cannot be exploited in the same was as in grid world, this may be a less critical issue.

5.3 Intrinsic Motivation

The search for an intrinsic motivation for an agent to perform any given task, develop new behaviours, or learn its own embodiment is and take advantage of that is a key task in the development of continually learning, adaptable agents which are capable of working in highly dynamic environments. It is essential that an agent is able to identify important or interesting regions in the sensorimotor space, both to learn the model, or in fact learn a goal where no clear goal is immediately visible.

Empowerment seeks to do this (Salge and Polani, 2017) by defining an empirical measure which can be performed over the state-action space to definitively state the best possible states for an agent to be in to have sufficient future degrees of freedom. This valuable concept is unfortunately subject to the curse of dimensionality, and as such, other approaches to estimating empowerment have been sought (Zhao et al., 2019). We believe that we have shown that entropy maximisation allows for an agent to approximate such a position utilising an on-policy approach over varying environments, by instead considering more interesting or surprising regions of the state space to be the most valuable. This can more concisely be thought of as a form of information empowerment, where, as opposed to the mantra “All else being equal, be empowered” (Klyubin et al., 2005a), we consider that perhaps the notion in a adaptive learning agent should be “all else being equal, be interesting”.

5.4 Applications

We believe that RRL has a variety of potential applications in terms of control architecture for an autonomous agents, as well as less lofty pursuits. As we see in Fig. 2 and Fig. 3, the agent shows highlighted regions of preference around the obstacles, preferring to avoid walls and obstructions. We believe there is potential here to consider the notion of “curiosity path planning”, where an agent plans the route on the basis of interesting regions within a known environment to better learn about, or be available for future tasks.

The full potential of RRL will become available only if the methods discussed here are incorporated into a more general framework that includes continuous state and action spaces as well as higher-order reflexion. A very promising option to enable RRL in a wider range of contexts appears to be a combination with inverse reinforcement learning (Ng and Russell, 2000), where the agent derives a reflexive reward signal from the performance of a parallel inverse learner. A detailed discussion of this approach is beyond the scope of this paper and is subject of current research.

6 CONCLUSION

Reflexive reinforcement learning (RRL) is a new direction in machine learning. It is based on the observation that in learning problem where no direct gradient can be used in order to adapt to a particular, the representation of information from the environment (such as state information or evaluative information) requires not only guiding principles (such as smooth-

ness, consistency and locality), but also provides information that can be used to decide about the actions of an agent.

The advantage of reflexive reinforcement learning is that an agent can learn even in the absence of an evaluative signal (reward and punishment), it can bootstrap elementary actions (as in homeokinesis (Der and Martius, 2012)) or can learn about options in the environment (as in empowerment (Klyubin et al., 2005b)), and obtain more meaningful and generalisable representations (see (Smith and Herrmann, 2019)).

The unavoidable difficulty in reflexive reinforcement learning consists in the fact that the use of quantities that are eventually based on the reward as a reward, introduces a feedback loop which can lead to instabilities or divergences. This is not unknown in RL, where e.g., an often visited source of low reward can dominate a better source of reward that is rarely found, or in cases where correlations among basis functions lead to divergences as notice already in Ref. (Baird, 1995).

In RRL such feedback is even more typical, but can also be used to introduce structure the state space by self-organised pattern formation or to identify hierarchical relationships as will be studied in future. In order to keep the effects of self-referentiality under control and to make use of their potential a dynamical systems theory of reinforcement learning is required that does not only consider the agent as a dynamical system, but the full interactive system formed by the agent, its environment and its internal representations.

ACKNOWLEDGEMENTS

This research was funded by EPSRC through the CDT RAS at Edinburgh Centre for Robotics. Discussions with Calum Imrie and Simon Smith are gratefully acknowledged.

REFERENCES

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier.

Der, R. and Martius, G. (2012). *The playful machine: Theoretical foundation and practical realization of self-organizing robots*, volume 15. Springer Science & Business Media.

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87.

Friston, K. J., Daunizeau, J., and Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS one*, 4(7):e6421.

Herrmann, M. and Der, R. (1995). Efficient q-learning by division of labour. In *Proceedings ICANN*, volume 95, pages 129–134.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005a). All else being equal be empowered. In *European Conference on Artificial Life*, pages 744–753. Springer.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005b). Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135. IEEE.

Laureiro-Martínez, D., Brusoni, S., and Zollo, M. (2010). The neuroscientific foundations of the exploration-exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics*, 3(2):95.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization science*, 2(1):71–87.

Ng, A. Y. and Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *IMCL*, pages 663–670.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17.

Prange, C. and Schlegelmilch, B. B. (2009). The role of ambidexterity in marketing strategy implementation: Resolving the exploration-exploitation dilemma. *Business Research*, 2(2):215–240.

Salge, C., Glackin, C., and Polani, D. (2014). Empowerment – an introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer.

Salge, C. and Polani, D. (2017). Empowerment as replacement for the three laws of robotics. *Frontiers in Robotics and AI*, 4:25.

Smith, S. C., Dharmadi, R., Imrie, C., Si, B., and Herrmann, J. M. (2020). The DIAMOnD model: Deep recurrent neural networks for self-organising robot control. *Frontiers in Neurorobotics*, 14:62.

Smith, S. C. and Herrmann, J. M. (2019). Evaluation of internal models in autonomous learning. *IEEE Transactions on Cognitive and Developmental Systems*, 11(4):463–472.

Sutton, R. S. and Barto, A. G. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An Introduction*. MIT Press.

Thrun, S. (2002). Probabilistic robotics. *Communications of the ACM*, 45(3):52–57.

Tschantz, A., Millidge, B., Seth, A. K., and Buckley, C. L. (2020). Reinforcement learning through active inference. *arXiv preprint arXiv:2002.12636*.

Walter, E. (2008). *Cambridge advanced learner's dictionary*. Cambridge University Press.

Zhao, R., Tiomkin, S., and Abbeel, P. (2019). Learning efficient representation for intrinsic motivation. *arXiv preprint arXiv:1912.02624*.