Comparing Action Sets: Mutual Information as a Measure of Control

Sascha Fleer(✉) and Helge Ritter

Neuroinformatics Group, EXC Cognitive Interaction Technology (CITEC),
Bielefeld University, Bielefeld, Germany
{sfleer,helge}@techfak.uni-bielefeld.de
http://www.neuroinformatik.de

Abstract. Finding good principles to choose the actions of artificial agents like robots in the most beneficial way to optimize their control of the environment is very much in the focus of current research in the field of intelligent systems. Especially in reinforcement learning, where the agent learns through the direct interaction with the environment, a good choice of actions is essential. We propose a new approach that allows a predictive ranking of different action sets with regard to their influence on the learning performance of an artificial agent. Our approach is based on a measure of control that utilizes the concept of mutual information. To evaluate this approach, we investigate its prediction of the effectiveness of different sets of actions in "mediated interaction" scenarios. Our results indicate that the mutual information-based measure can yield useful predictions on the aptitude of action sets for the learning process.

Keywords: Reinforcement learning · Environment control · Q-learning · Mutual information · Mediated interaction learning · Physics-based simulation

1 Introduction

One of the bigger visions in the field of intelligent systems is to endow an artificial agent with the ability to solve human-level problems. To deal with such complicated tasks, the agent has to explore the learning domain autonomously by performing actions that affect the environment directly and learn from these effects. This important aspect is a distinct focus of reinforcement learning where the agent learns through the direct interaction with the environment [11]. To explore the terrain, the agent executes actions which alter its surroundings and lead to a feedback how much the chosen action benefits the agent in its current situation with respect to the primary learning goal. Therefore, actions play a crucial role in reinforcement learning as they determine how much control the agent has over the environment and by thus influence the effectiveness of exploration and learning.

This motivates the following question: are there general features that distinguish action sets that facilitate exploration, learning and control ("good" action

A. Lintas et al. (Eds.): ICANN 2017, Part I, LNCS 10613, pp. 68–75, 2017.
https://doi.org/10.1007/978-3-319-68600-4_9

sets) from action sets for which exploration, learning and control is more difficult? Obviously, criteria to recognize such action sets would be of interest for designing interactive learning algorithms that are fast and efficient.

In the present paper, we consider this question for choosing a good action set for a reinforcement learning agent that has to learn a challenging "mediated interaction" task that can only be solved when the agent recognizes to use a "mediator object" as a tool to reach its goal. Using a simulation study with simulated physics, we present results that indicate that a simple, entropy-based measure can rank different possible action sets in a way that correlates well with the learning performance in the reinforcement learning task.

By defining actions relative to a coordinate system, we connect the choice of an action set with the choice of a coordinate system. In this way, we can use our approach also to rank different options for choosing a coordinate system that is "favorable" for the learning task at hand. Our findings are consistent with the expectation that "good" coordinate systems should be those that make uncertainty-reducing actions easy to express. For the task at hand, this turns out to be better achieved with "relational" instead of "absolute" coordinate choices.

In the next section we briefly anchor our notation to define action sets for a reinforcement learning agent and then describe our measure. Section 3 presents the learning domain, Sect. 4 reports the experiments and results and Sect. 5 provides the conclusion.

2 Comparing Action Sets: Mutual Information as a Measure of Control

The concept to maximize the information over the environment to gain more control is studied in various fields [5,12]. In our approach, the mutual information is employed to compare different action sets $\mathcal{A}$, defined by different sensorimotor coordinate systems that determine the agents motions. The ranking order is then used as a criteria for predicting the agents learning performance while using the respective action set.

Reinforcement learning is a class of machine learning algorithms for solving sequential decision making problems through maximization of a cumulative scalar reward signal [11]. It can be defined by the standard formulation of a Markov decision process $(\mathcal{S}, \mathcal{A}, P^{\mathcal{A}}, \mathcal{R}, \mathcal{S}_0)$, where $\mathcal{S}$ denotes the set of states and $\mathcal{A}$ the set of admissible actions. $P^{\mathcal{A}}$ is the set of transition matrices, one for each action $a \in \mathcal{A}$ with matrix elements $\mathcal{P}^a_{ss'} : \mathcal{S} \times \mathcal{A} \longrightarrow \mathcal{S}'$ specifying the probability to end up in state s' after taking action a when in state s. The probability to execute action a in state s can be defined as $\mathcal{P}^a_s$. Finally, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$ is a scalar valued reward function and $\mathcal{S}_0 \subseteq \mathcal{S}$ is the set of starting states.

If the agent induces a state transition from state $s \in \mathcal{S}$, the final state s' is within a subset of possible states $\mathcal{S}' \subseteq \mathcal{S}$. This can be described by the uncontrolled probability $\mathcal{P}_{ss'}$, which fulfills $\mathcal{P}_{ss'} = \sum_a \mathcal{P}^a_s \mathcal{P}^a_{ss'}$. To measure the uncertainty about the next state, the *entropy* [6,9] $\mathcal{H}_s(\mathcal{S}')$ of the current state s can be computed.

$$\mathcal{H}_s(\mathcal{S}') = -\sum_{s' \in \mathcal{S}'} \mathcal{P}_{ss'} \ln\left(\mathcal{P}_{ss'}\right) \tag{1}$$

By introducing surprise [2] (or self-information) which is defined as the negative logarithm of the probability, i.e. $-\ln(\mathcal{P}_{ss'})$, the entropy can be interpreted as the average surprise to end up in one of the possible states s' that can be reached within one transition step. Thus, a small entropy indicates a better prediction of s' while a large entropy implies a high uncertainty of the next state.

Additionally the *conditional entropy* [6] $\mathcal{H}_s(\mathcal{S}'|\mathcal{A})$ of state s can be computed. It measures the average surprise of state s to end up in a state s', conditioned on the actions $a \in \mathcal{A}$, resulting in

$$\mathcal{H}_s(\mathcal{S}'|\mathcal{A}) = -\sum_{a \in \mathcal{A}} \mathcal{P}^a_s \left[\sum_{s' \in \mathcal{S}'} \mathcal{P}^a_{ss'} \ln\left(\mathcal{P}^a_{ss'}\right)\right]. \tag{2}$$

The rate of influence enforced by the set of actions $\mathcal{A}$ on the uncontrolled transitions $\mathcal{P}_{ss'}$ of a state s is thus given by the difference of the state's entropy (1) and the conditional entropy (2) leading to

$$\mathcal{M}_s(\mathcal{S}', \mathcal{A}) = \mathcal{H}_s(\mathcal{S}') - \mathcal{H}_s(\mathcal{S}'|\mathcal{A}). \tag{3}$$

Equation (3) is known as the *mutual information* [6,9]. It measures the reduction of uncertainty of the final states $s' \in \mathcal{S}'$ due to the control of action set $\mathcal{A}$. The ranking order of the expected mutual information, which is (3) averaged over all available states $s \in \mathcal{S}$,

$$\mathcal{M}(\mathcal{S}', \mathcal{A}) = \mathbb{E}_{s \in \mathcal{S}}\left[\mathcal{M}_s(\mathcal{S}', \mathcal{A})\right] \tag{4}$$

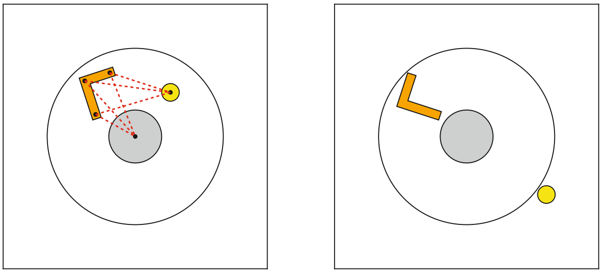
turns out to be highly correlated with the learning performance of the reinforcement learning agent. Therefore, we propose to use the action set which leads to the lowest uncertainty of events within the domain and to select the coordinate system according to $\mathcal{A}^* = \text{argmax}_{\mathcal{A}} \mathcal{M}(\mathcal{S}', \mathcal{A})$.

This choice has the interpretation that the best coordinate system maximizes the expected mutual information (4).

3 Learning Domain

To compare different action sets using the entropy-based measure of control, we employ a 2D simulation world in which an agent has to solve a mediated interaction task. The world is illustrated in Fig. 1 and consists of an agent, a disc-shaped "target-object" and an L-shaped "mediator-object" ("tool"). The simulated learning domain further utilizes the open source `Box2D` physics engine [1] for interaction and collision handling.

The task of the agent is to bring the target-object into the shaded circle in the center ("goal area"). To this end, the agent can at each time step "pick" the target-object or the mediator-object and exert a (discretized) force/torque



(a) Illustration of the used sensory information

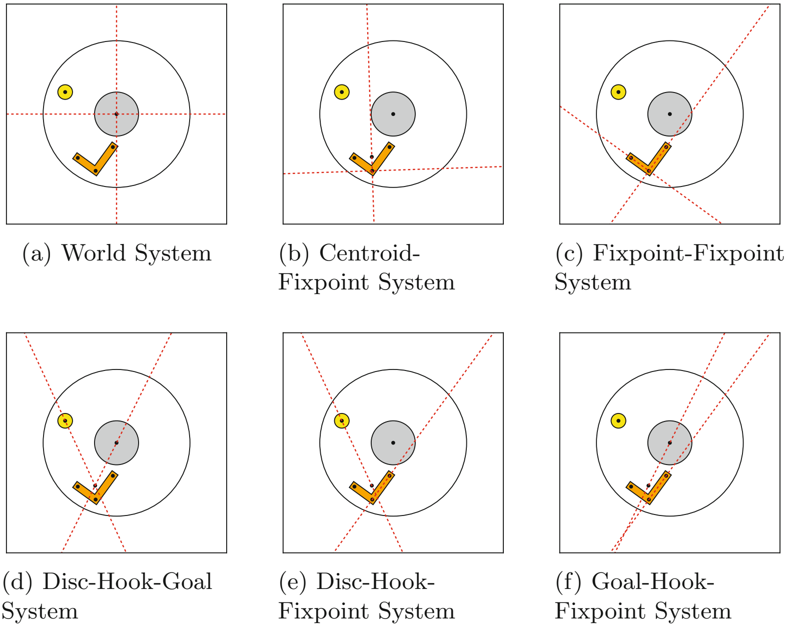
(b) An "extension-of-reach" tool-using task

Fig. 1. The simulation world

at the chosen picking location. Picking locations (indicated by black dots in Fig. 1a) are discretized and fixed at the objects: the target-object offers a single picking location at its center, the mediator-object offers three picking locations, two at its ends and one in the middle. Furthermore, there is an additional picking location in the center of the domain which deals as an unbiased starting location for the agent and is further integrated to be an absorbing state that increases the stability of applied learning algorithms. However, the agent can only reach picking locations that lie inside the circular area. Therefore, when the target-object is outside the circle, the agent must first "discover" that the mediator-object can be used to extend the agent's reach beyond the circle boundary. We assume that the agent has a simple relational perception of the world state consisting of the six scalar distances between the three picking locations on the mediator-object and the center of the target-object, and the three distances of the picking locations and the domain's origin. They are visualized by the dotted lines in Fig. 1a. Additionally, the sensory representation encodes the agent's current picking choice. Learning occurs in discrete episodes, each episode being limited to 100 interaction-steps. If the agent is able to navigate the target object in the goal area, it receives a fixed reward of $R = 10$. The learning is handled by an ϵ-greedy Q-Learning algorithm with eligibility traces and linear function approximation [13]. To make the learned algorithm more stable, artificial noise is integrated into the system, that makes the agent execute a random action with a probability of 0.1. For performing and evaluating the learning process, the `RLPy` learning framework [3] is used. To adapt it to the specific needs of this work, it is extended by the presented learning domain and some additional functions.

4 Experiments

Six different coordinate systems (Fig. 2) were designed that exploit the different salient points within the learning domain (Fig. 1a). They are utilized by the agent as action sets $\mathcal{A}_i$ that define in which way the objects can be moved through the



(a) World System

(b) Centroid-Fixpoint System

(c) Fixpoint-Fixpoint System

(d) Disc-Hook-Goal System

(e) Disc-Hook-Fixpoint System

(f) Goal-Hook-Fixpoint System

Fig. 2. Illustration of the different coordinate systems, representing the different action sets $\mathcal{A}_i$

environment. All coordinate systems, except Fig. 2a, are not fixed in the world but alter according to the objects positions.

The influence of each of these coordinate systems $\mathcal{A}_i$ – more precisely, its associated action set, as described in Sect. 2 – on the agent's learning performance is now compared with their ranking according to the mutual information measure $\mathcal{M}(\mathcal{S}, \mathcal{A}_i)$, introduced in (4). Estimating entropies from finite samples of probability densities can be a challenging problem and has been discussed in many works [8,10]. We here adopt the most basic approach to estimate the entropy [10]. The probability densities $\mathcal{P}^a_{ss'}$ and $\mathcal{P}^a_s$ used to compute $\mathcal{M}(\mathcal{S}, \mathcal{A}_i)$ are approximated by tessellating the agent's state space into $N = 10 \cdot 10^3$ Voronoi cells. In this space $200 \cdot 10^3$ tuples (s, a, s') were counted while the agent was performing a random walk. (1), (2) are then used to estimate the mutual information. For each coordinate system the results are averaged over 20 runs of the experiment while the standard deviation of the mean is used as the estimate of the error.

To this end, we consider three exemplary learning scenarios:

Single-Object Interaction Scenario. At the beginning of a learning episode, the positions of the mediator-object and the target-object are sampled from the uniform distribution over the simulation world inside the agent's interaction range (e.g. Fig. 1a). The agent now has to learn how to move the target-object into the goal area. After successfully solving the task instance or exceeding the limit of possible interaction-steps per episode, the task starts anew with different initial object positions that are again within the agents interaction range.

Mediated Interaction Scenario. This "extension-of-reach" scenario is structured like the first one, but the target-object is distributed *outside* the border of the agent's interaction range (e.g. Figure 1b). Now it is only possible for the agent to solve this task by learning to exploit the mediator-object as a tool to pull the target-object inside the agent's interaction range.

Mixed Interaction Scenario. The last scenario is a mixture of the **Single-Object-Interaction** and **Mediated-Interaction** task, where one of the mentioned scenarios is chosen at the beginning of each episode with equal probability.

To get a preferable general measure of the learning performance for every $\mathcal{A}_i$, each of them is utilized to learn the three presented learning scenarios. All scenarios were learned over $500 \cdot 10^3$ steps. We evaluate the learning performance for two kinds of linear function approximators, representing the sensor vector of the learning domain. The first, more efficient real valued representation is based on Gaussian radial basis functions (RBF) [4,7]. The second is a simpler binary fixed sparse representation (FSR) [4].

For evaluating the efficiency of the learning processes, we depict the number of learning steps as a function of the average reward per episode $\langle R \rangle$ that is received by the agent. To compute $\langle R \rangle$, the learning performance under the current policy was evaluated over 100 episodes for each of the 25 evaluated data points. The results were then averaged over 20 distinct learning runs, where the standard deviation of the mean is used as the error.

As an example, Fig. 3 shows the agent's learning performance of the "Mediated Interaction Scenario" for the two used state representations. An optimal performance for solving the task is reached at $\langle R \rangle = 10$. In both plots, the learning performance is highly varying for each used coordinate system. Half of the coordinate systems achieve completely different results within the learning process for the two used state representations. Nevertheless there are 3 coordinate systems (see Fig. 2a, c and d) that lead to similar results. These coordinate systems include the two best ones and one with poor performance.

To evaluate the performance that takes all learned scenarios and state representations under consideration, the global reward $\mathcal{R}_{\text{global}}$ is defined as the sum over all $\langle R \rangle$, evaluated for the coordinate system $\mathcal{A}_i$. Table 1 now ranks the coordinate systems according to $\mathcal{R}_{\text{global}}$ and additionally lists the respective expected mutual information of all available states $\mathcal{M}(\mathcal{S}, \mathcal{A}_i)$. Although the two rankings are not exactly aligning, the coordinate system with the best and worst global reward can be clearly identified by using the mutual information. This two systems are the ones which performs best and worst in "all" evaluated scenarios. Three of the four not aligning frames are exactly the ones that behave diverse for the different kinds of scenarios and representations, as e.g. illustrated in Fig. 3. A further mentionable point is that the mutual information of similar constructed coordinate systems like Fig. 2b and c is also similar. The ranking is reliable within the facility of predicting the best action set. An efficient ranking

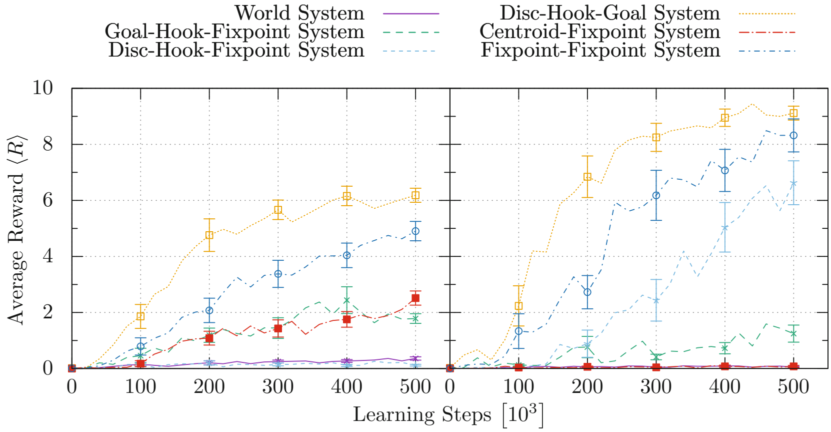


Fig. 3. Time course of the agent's average reward $\langle R \rangle$ of solving the "Mediated Interaction Scenario" using the different coordinate systems shown in Fig. 2. In the left plot, the sensor vector is represented by binary features while in the right plot the real-valued representation, based on RBFs, is used during learning.

Table 1. The ranking of the different movement frames according to R_{global}. The movement frames where the ranking of $\mathcal{M}(\mathcal{S}, \mathcal{A}_i)$ differs from the ranking of $\mathcal{R}_{\text{global}}$ are the ones between the "Disc-Hook-Goal System" and the "World System".

$\mathcal{R}_{\text{global}}$	$\mathcal{M}(\mathcal{S}, \mathcal{A}_i)$	Coord. Systems $\mathcal{A}_i$ - Figure
1183.25 ± 36.3358	1.8147 ± 0.0020	Disc-Hook-Goal System - 2d
650.54 ± 38.2204	1.7786 ± 0.0028	Disc-Hook-Fixpoint System - 2e
547.015 ± 50.0568	1.7581 ± 0.0029	Fixpoint-Fixpoint System - 2c
471.995 ± 41.7003	1.8038 ± 0.0015	Goal-Hook-Fixpoint System - 2f
235.925 ± 23.3319	1.7645 ± 0.0032	Centroid-Fixpoint System - 2b
135.96 ± 9.23605	1.7364 ± 0.0028	World System - 2a

of the non-optimal sets might be hindered by the much larger space of solutions. It is further undermined by the strong dependence of the learning performance on the chosen set of parameters used by the learning algorithm. However, the ranking still provides some orientation for the non-optimal candidates.

5 Conclusion

In this work, we investigate the impact of action sets arising from different sensorimotor coordinate frames on the efficiency of learning mediated-interaction scenarios. Therefore the learning performance of different action sets were evaluated on solving different single-object and multi-object interaction tasks while using reinforcement learning methods. Additionally, the mutual information was computed for each action set which measures the reduction of uncertainty of the agent's next state due to the control of the used action set. After the empirical demonstration that different action sets lead to different learning performances, their performance ranking is compared with the ranking of their mutual information within the environment.

We find that the concept of mutual information, conditioned on the chosen action set, is well suited to predict the ranking of the general learning performance. Although these two rankings are not exactly aligning to each other, there are lots of similarities. In addition, the worst and the best action set can be clearly identified. Based on these findings, further investigations in this matter may lead to a better understanding of the relationship between the mutual information and the agent-environment interaction which can be used to guide the choice of actions within difficult learning scenarios.

Acknowledgments. This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

1. Catto, E.: Box2d (2010). http://www.box2d.org
2. Friston, K.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. **11**(2), 127–138 (2010)
3. Geramifard, A., Dann, C., Klein, R.H., Dabney, W., How, J.P.: Rlpy: a value-function-based reinforcement learning framework for education and research. J. Mach. Learn. Res. **16**, 1573–1578 (2015)
4. Geramifard, A., Walsh, T.J., Tellex, S., Chowdhary, G., Roy, N., How, J.P., et al.: A tutorial on linear function approximators for dynamic programming and reinforcement learning. Foundations and Trends®. Mach. Learn. **6**(4), 375–451 (2013)
5. Gottlieb, J., Oudeyer, P.Y., Lopes, M., Baranes, A.: Information-seeking, curiosity, and attention: computational and neural mechanisms **17**(11), 585–593 (2013)
6. Jones, D.S.: Elementary Information Theory. Oxford University Press, New York (1979)
7. Moody, J., Darken, C.J.: Fast learning in networks of locally-tuned processing units. Neural Comput. **1**(2), 281–294 (1989)
8. Paninski, L.: Estimation of entropy and mutual information. Neural Comput. **15**(6), 1191–1253 (2003)
9. Shannon, C.E., Weaver, W.: The mathematical theory of communication. Mathe. Gaz. **34**(310), 312 (1950)
10. Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R.: Entropy and information in neural spike trains. Phys. Rev. Lett. **80**(1), 197–200 (1998)
11. Sutton, R.S., Barto, A.G.: Introduction to Reinforcement Learning, vol. 135. MIT Press Cambridge, Cambridge (1998)
12. Tishby, N., Polani, D.: Information theory of decisions and actions. In: Cutsuridis, V., Hussain, A., Taylor, J. (eds.) Perception-Action Cycle. Cognitive and Neural Systems. Springer, New York (2010)
13. Watkins, C.J., Dayan, P.: Q-learning. Mach. Learn. **8**(3–4), 279–292 (1992)