The BRIO® Labyrinth Game - A Testbed for Reinforcement Learning and for Studies on Sensorimotor Learning

Jan Hendrik Metzen[⋆], Elsa Kirchner[⋆,†], Larbi Abdenebaoui[⋆], and Frank Kirchner[⋆,†]

[⋆] Researchlab Robotic, German Research Center for Artificial Intelligence (DFKI)

[†] Robotics Group, University Bremen

Motivation Applying Reinforcement Learning (RL) in the context of robot learning is a challenging problem mainly due to the continuous state and action spaces, inherent noise in sensors and actuators, and a lack of full observability due to hidden states. Furthermore, learning in the physical world is expensive because of material deterioration and the requirement of continuous human supervision to avoid that the robot gets damaged during learning. Thus, there is a high demand for robotic benchmark scenarios that allow to test existing and develop new RL algorithms suited for robotic applications.

The BRIO® Testbed In this work we present the BRIO® testbed, a system based on the BRIO® labyrinth game. The ultimate goal of this game is to steer a ball indirectly through a maze from a start to a goal position by changing the orientation of the board. In order to allow an RL agent to control the game fully autonomously, an off-the-shelf BRIO® labyrinth game (see Figure 1) has been equipped with two servo motors (Robotis Dynamixel DX-117) and two potentiometers, which allows the remote control and measurement of the orientation of the game's board. The board is mounted on a platform, and a camera is placed above this platform which allows to estimate the current position of the ball on the board. For this purpose, a vision algorithm has been developed which segments the ball in the camera image and maps the position of the ball in the image onto a position in the labyrinth coordinate system. The mapping is calibrated automatically at the beginning of each session using landmarks. Furthermore, it is possible to detect that a ball has fallen through a hole of the maze by means of a piezo sensor that has been integrated in the base of the game and is able to detect the ball's impact on the ground. If such an impact is detected, a new ball is supplied by a custom-made ball depot that is mounted on the game and thus, it is possible to play more than one session without putting the ball back manually.

Besides the physical labyrinth, there exists also a simulation of the whole system. This simulation is based on the Open Dynamics Engine (ODE) [3], a high-performance library for simulating rigid body dynamics in realtime that is based on a Lagrange multiplier method with first order integrator. The friction calculation is based on an approximation to the Coulomb friction model. The simulation environment offers the possibility to interface with the modeled system, to control the simulated actuators, to receive the actual simulated sensors, as well as to change the physical parameters of the simulation. The maze is modeled using a 3D mesh model, which allows a precise reproduction of the original shapes.

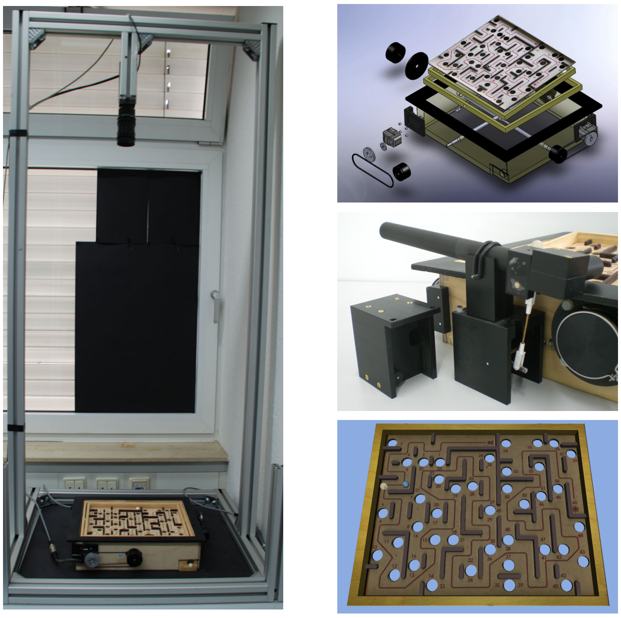

Figure 1: The left image shows the setup of the BRIO® labyrinth testbed with a camera mounted approximately 1 meter above the board and servo motors installed on the game to control the board's orientation. The upper right image shows an exploded view of the BRIO® labyrinth system, the middle right image shows the ball depot, and the lower right image shows the latest version of the simulation.

Learning Policies for the Simulated BRIO® Labyrinth Since the ball can only indirectly be steered through the BRIO® labyrinth by changing the orientation of the board, potential actions for the agent are changes of two rotatable axes. The action space can thus be defined as the crossproduct of the allowed positions for the two axes $[-\phi_x, \phi_x] \times [-\phi_y, \phi_y]$. Relevant information are the current ball position, the current ball velocity vector as well as the current board orientation. Thus, the state space consists of six continuous-valued dimensions. Since it is the goal of the agent to steer the ball as fast as possible from the start state to the goal state, the agent receives a large positive reward when it reaches the goal and a small negative reward for every time step it hasn't reached the goal. In order to simplify the task and to guide the agent into the correct direction, the agent gets an additional positive reward whenever it reaches one of a set of predefined waypoints (subgoals) and a negative reward when it passes a waypoint in the wrong direction. The agent is allowed to execute an action every 100ms. In a first experiment, we used SARSA(λ) learning with a Cerebellar Model Articulation Controller (CMAC) function approximator [4] to learn policies for the BRIO® labyrinth simulation. We used ϵ-greedy exploration with $\epsilon = 0.01$, set the learning rate to $\alpha = 0.1$, the discount factor to $\gamma = 1.0$, and the eligibility trace decay rate to $\lambda = 0.9$. The CMAC function approximator used 10 independent tilings, each consisting of 5 tiles per board orientation and ball velocity dimension and 11 tiles per ball position dimension. This setup was able to learn policies that reached the goal reliably within less than 100000 episodes. A video of a policy that learned to reach the goal after 65000 episodes can be found on `http://robotik.dfki-bremen.de/en/research/projects/cognitive-robotics/brio-labyrinth-2.html`

Learning Policies for the Real-World BRIO® Labyrinth

In this section, we give an outlook of future studies that will be carried out on the BRIO® testbed. These will mainly investigate how policies for the the physical BRIO® labyrinth can be learned. Learning policies for the physical BRIO® labyrinth is difficult since the number of trials that are realizable on the physical system is limited because of the need of human supervision (as in most robotic application). Thus, learning on the real system should require only a small number of trials. One promising approach is to use experience gained in the simulation to boost learning on the physical system, i.e. to apply methods from the area of Transfer Learning (see [5]). This approach is complicated by the fact that the dynamics of simulated and real system remain slightly different even though the simulation has been calibrated to resemble the behaviour of the real BRIO® system as good as possible. It will be studied, which kind of knowledge can most easily be transferred from simulation to physical system: Transferring a learned policy directly is the most straight forward approach. Unfortunately, a specific policy will usually be highly tuned to the specific characteristics of the simulation and does not allow to identify changes of the dynamics between simulated and physical system easily. Transferring a learned model along with a learned policy provides the advantage that situations where the behaviors of simulated and real system deviate significantly can be identified and thus the agent can identify regions of the game where exploration is required to adjust model and policy to the specific dynamics of the physical system. Furthermore, it will also be investigated if an agent can successfully transfer structural knowledge of the problem from simulation to the physical system and whether this knowledge can be utilized to boost learning. This structural knowledge can be e.g. a decomposition of the problem into several independent subproblems (see Abdenebaoui et al. [1]) or an adaptive function approximator (e.g. an adaptive tile coding [6]) whose resolution is tuned to the specific problem such that the resolution is increased for critical parts of the board (close to holes) while resolution is kept low for uncritical parts.

Neurobiological Research on Transfer and Imitation Learning In order to support the aforementioned research in the area of transfer learning, the BRIO® testbed will also serve as platform for investigating sensorimotor learning in humans by means of behavioral, EEG and fMRI studies. These studies might provide new insights into the learning and execution of behavior as well as into the transfer and adaptation of automated behavior that is needed for adaptation to new situations. Transferring behavior is especially interesting since it allows rapid and efficient adaptation, e.g. enabling humans that are able to play the BRIO® labyrinth with two hands to adapt their behaviour within a short period of training so that they are also able to control the board with a joystick by using only one hand.

Furthermore, the BRIO® labyrinth is a good platform for studies in the field of imitation learning (see [2]). Imitation learning takes place in humans while they are watching someone else performing a motor action. Neurons in the premotor and parietal cortex are thought to transform specific sensory information into a motor format and by this allow motor learning. Imitation learning mediates the transformation of a seen action into an ideally identical motor action, and thus it is a very efficient way of learning new behaviours. Studies are planned in which sensor information recorded from a camera and potentiometers while a subject is playing the BRIO® labyrinth is used for accelerating the learning of a control policy by a virtual agent.

References

[1] L. Abdenebaoui, E. A. Kirchner, Y. Kassahun, and F. Kirchner. A connectionist architecture for learning to play a simulated BRIO labyrinth game. In *Proceedings of the 30th Annual German Conference on Artificial Intelligence (KI07)*, pages 427–430, Osnabrück, Germany, September 10-13 2007. Springer-Verlag.

[2] G Buccino, S Vogt, A Ritzl, G Fink, and K Zilles. Neural circuits underlying imitation learning of hand actions: An event-related fMRI study. *Neuron*, Jan 2004.

[3] R. Smith. Open Dynamics Engine (www.ode.org), 2005.

[4] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning).* The MIT Press, March 1998.

[5] M. E. Taylor. *Autonomous Inter-Task Transfer in Reinforcement Learning Domains.* PhD thesis, Department of Computer Sciences, The University of Texas at Austin, August 2008. Available as Technical Report UT-AI-TR-08-5.

[6] S. Whiteson. *Adaptive Representations for Reinforcement Learning.* PhD thesis, Department of Computer Sciences, University of Texas at Austin, May 2007.