MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Learning Tasks in a Complex Circular Maze Environment

van Baar, J.; Jha, D.; Romeres, D.; Sullivan, A.; Nikovski, D.N.

TR2018-169 December 29, 2018

Abstract

The purpose of this article is to introduce a circular maze system as a challenging environment to solve, which could be of interest to the robot and reinforcement learning community. Recently, there have been rapid developments in the fields of machine and reinforcement learning, largely due to the success of deep learning approaches. This has also led to increased interest in the area of learning physics based systems. The circular maze environment that we present here is a low-DoF complex system which could be used to investigate many interesting learning problems. We propose some initial results using both model-free and model-based learning approaches to solve the environment with a single and multiple marbles, to demonstrate some of the challenges that this system presents. We hope to opensource the simulation software and hardware design details of the system in the near future.

Modeling the Physical World: Perception, Learning, and Control, NIPS Workshop

Learning Tasks in a Complex Circular Maze Environment

Devesh K. Jha*, **Diego Romeres***, **Jeroen van Baar***, **Alan Sullivan, Daniel Nikovski**
MERL, Cambridge, MA
{jha,romeres,jeroen,sullivan,nikovski}@merl.com

Abstract

The purpose of this article is to introduce a circular maze system as a challenging environment to solve, which could be of interest to the robot and reinforcement learning community. Recently, there have been rapid developments in the fields of machine and reinforcement learning, largely due to the success of deep learning approaches. This has also led to increased interest in the area of learning physics-based systems. The circular maze environment that we present here is a low-DoF complex system which could be used to investigate many interesting learning problems. We propose some initial results using both model-free and model-based learning approaches to solve the environment with a single and multiple marbles, to demonstrate some of the challenges that this system presents. We hope to open-source the simulation software and hardware design details of the system in the near future.

1 Motivation and Introduction

Recent developments in the field of artificial intelligence have been mainly about sophisticated probabilistic algorithms and deep learning representations for pattern recognition problems. Most of the algorithms that are currently studied are being compared to human accuracy in performing the related tasks [1]. However, the area of learning in physical systems has not seen the same degree of advancement. Most of the advanced learning algorithms fail to learn efficiently in complex problems which are very intuitive to human intelligence [2, 3]. In this paper, we present a circular maze environment (CME) (see Figure 1) which is a complex learning domain due to its constrained geometry, nonlinear dynamics and long planning horizon with several discontinuities. We hope that the presented system will be used as a benchmark system for reinforcement and robot learning applications.

For reinforcement learning (RL) algorithms to solve tasks efficiently, the algorithms need to model and predict the physical world well. This problem becomes apparent in the field of robot learning, where the challenge is to use data-efficient algorithms to learn policies to perform tasks in environments with high action-dimension in order to minimize the stress to the physical system. Most policy gradient algorithms can learn efficiently only in the presence of an initial solution [4]. Unfortunately, designing an initial controller for a complex, non-linear system is very challenging, even for low-dimensional systems (e.g., the proposed CME). Model-based RL algorithms have gained a lot of traction in recent literature, because the models can be re-used for learning multiple tasks [5]. However, learning global models is a difficult problem, and thus most of the work in the literature has been focused on learning local models [6]. Moreover, learning local models efficiently is also challenging as it requires perturbing the system in a local manifold (e.g., along some controlled trajectories). In light of these facts, learning in a complex, non-linear environment remains a very challenging task. The

*Authors with equal contribution

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

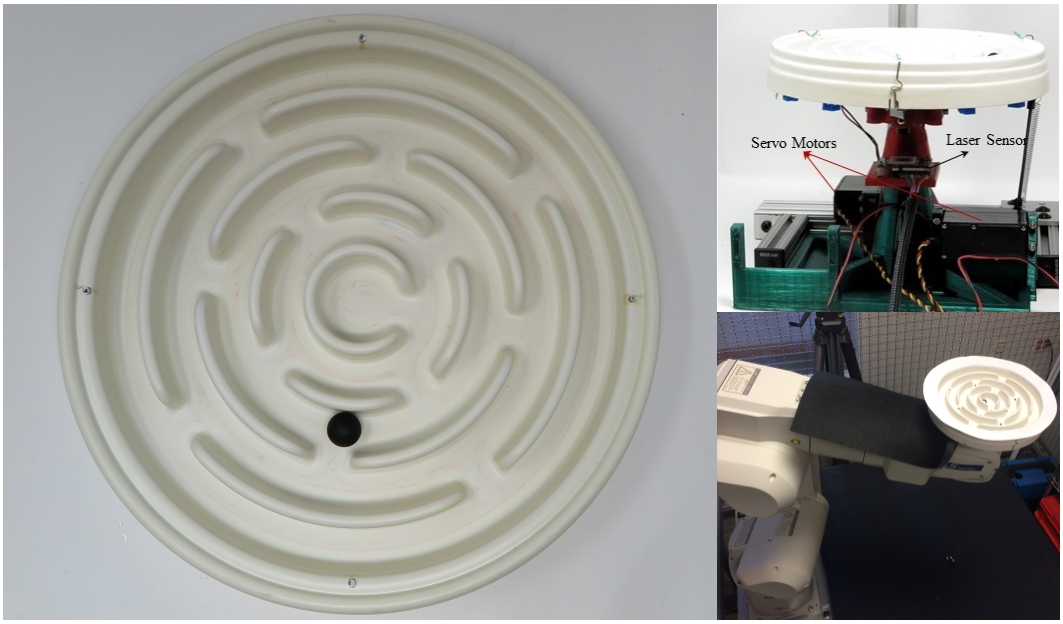

Figure 1: Circular Maze Environment (CME) with a single marble–a toy bought on Amazon. The goal is to maneuver the marble(s) from the outermost ring through a series of gates into the center area. **Right** Experimental platform used for model-based learning, and **Bottom-Right** experimental platform used for transfer learning.

CME we present in this paper highlights a lot of these issues and thus provides a good test-bed to benchmark a lot of different algorithms.

To understand and research different aspects of robot learning problems that come with such a complex system, we have investigated two different approaches for learning in the CME. One is a model-based approach that uses semiparametric Gaussian process (GP) regression [7, 8] to combine physics-based models for the motion dynamics with data-driven models, to accurately estimate the forward dynamics of marble in the environment. This model is then used for trajectory optimization using a stochastic control algorithm. The other approach uses a model-free deep RL algorithm to learn a policy in simulation and then transfer it to a real setup. The main idea of this approach is to learn a policy in simulation and transfer it to the real system with minimum fine-tuning. Both of these approaches are first used to learn models and policies with just one marble in the environment. In order to add further complexity to the problem, we considered more marbles in the environment. This leads to an even more complex planning problem as well as more complex models. However, for humans these tasks are intuitive.

	P [rad]		PI [rad]		NP [rad]		NPd [rad]		SP [rad]	
Ring	n=20	n=40	n=20	n=40	n=20	n=40	n=20	n=40	n=20	n=40
1	1.30	3.86	0.40	1.32	0.11	0.26	1.36	3.01	0.11	0.25
2	1.30	4.05	0.21	0.57	0.15	0.33	0.87	2.09	0.15	0.33
3	1.25	4.03	0.24	0.64	0.19	0.34	0.99	2.26	0.18	0.32
4	1.51	0.86	1.81	1.21	0.46	1.76	0.65	2.45	1.74	1.18

Table 1: Rollout performance for all the rings evaluated as the average absolute error in θ at 20 and 40 steps ahead.

2 Results

For the model-based RL approach, the full problem of navigating the marble to the center of the maze is simplified into sub-goals, which are to position the marble in front of an opening and then pass it through the opening. A model is learned for each individual ring of the maze using a semiparametric Gaussian process regression. For details, we refer the readers to [9]. Table 1 compares the rollout performances of several different models over a horizon of 20 and 40 steps (at a control rate of 30 Hz). In Table 1 mode "P" refers to a physical-based model derived using Lagrangian approach, "PI" refers to a physics-inspired GP model where the basis functions were provided by the physical model, "NP" refers to the nonparametric GP model, "NPd" refers to the discrete nonparametric model and "SP" refers to the semiparametric model (see [9] for more details) . The SP models were used to compute the controllers for the marble using the iterative LQG [10] algorithm. Composing the sequence of controllers we were able to design a controller to navigate the marble to the goal state. However, there are some failure cases due to the contacts of the marble and the walls during the motion. We are currently trying to use a guided policy search (GPS) [6]-type approach to design a full policy where the plan can also be learned along with the control. However, a GPS-like algorithm [11] is not directly applicable to the CME as a local policy may not be robust to disturbances due to contacts, friction, changes in the initial conditions etc.

In the model-free RL approach, we learn robust policies in simulation through randomization of appearance, physics and system parameters during the episodal learning (based on [12]). For details we refer the reader to [13]. We have demonstrated substantial reduction in fine-tuning for transfer to a real setup. Table 2 compares transfer learning for robust and non-robust policies. We report both the number of steps (observations of state) and success rate, i.e. getting the marble from the outermost ring into the center. For a single marble, the fine-tuning with a robust policy achieves 100% success rate, compared to 85% for non-robust policy. In addition, in the case of the non-robust policy the average episode length to solve the puzzle is longer. In the case of a two marble puzzle, fine-tuning a robust policy learning in simulation achieves about 75% success rate, while training on the real setup is still not able to learn a successful policy after more than 1M steps.

	Real	**Simulator**	**Fine-tune**
1 Marble–Robust	$\sim$3.5M (100%)	$\sim$4.0M (100%)	$\sim$55K (100%)
1 Marble–Non-Robust	$\sim$3.5M (100%)	$\sim$4.5M (100%)	$\sim$70K (85%)
2 Marbles–Robust	$\sim$1M (0%)	$\sim$3.0M (100%)	$\sim$225K (75%)

Table 2: Comparing TL for robust and non-robust policies for a single marble maze and robust policy transfer for a two marble maze game. Both the number of steps and success rate are reported.

3 Discussion

We have recently seen a rapid growth in the area of learning for physics-based systems, involving simple physics and relatively short planning horizons. The CME presented in this paper requires the learning of fast dynamics in the presence of friction, contacts, etc. Combined with a long planning horizon, this provides a challenging test-bed which we hope will be used in further research. Here, we list several possible research directions. We are interested in combining the prior physics knowledge with a neural network learning framework [14, 15, 16] and compare it with the proposed semiparametric GP models. It would also be interesting to see how the models learned for a single marble could be leveraged to compose accurate motion models for more than one marble in the environment. The current system is also a good physics-based system to study hierarchical RL [17]. For the end-to-end deep RL setting, we are working towards using a combination of model-based and model-free RL algorithm [18] so that learning could be more efficient and robust. Furthermore, leveraging transfer learning for sim-to-real is another interesting research direction, e.g., learning in both simulation and real environments [19]. Rather than specify a reward function, it would be desirable to use inverse RL techniques to recover the goal and sub-goals directly from trajectory observations. Another interesting direction is to solve this task in a manner similar to how humans would. To handle different instances of maze puzzles, e.g., a square wooden puzzle, the geometry and some material properties may be discovered from observations of interactions with the puzzle. Once the geometry has been recovered, the solution can be formulated as a path planning problem, and perhaps solved as a collection of sub-tasks [20].

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[2] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[3] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[4] Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.

[5] Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):408–423, 2015.

[6] Sergey Levine and Pieter Abbeel. Learning neural network policies with guided policy search under unknown dynamics. In *Advances in Neural Information Processing Systems*, 2014.

[7] Diego Romeres, Mattia Zorzi, R. Camoriano, and Alessandro Chiuso. Online semi-parametric learning for inverse dynamics modeling. In *Conference on Decision and Control*. IEEE, 2016.

[8] D. Romeres, M. Zorzi, R. Camoriano, S. Traversaro, and A. Chiuso. Derivative-free online learning of inverse dynamics models. *ArXiv e-prints*, September 2018.

[9] Diego Romeres, Devesh Jha, Alberto DallaLibera, Bill Yerazunis, and Daniel Nikovski. Learning hybrid models to control a ball in a circular maze. *arXiv preprint*, abs/1809.04993, 2018.

[10] Yuval Tassa, Tom Erez, and Emanuel Todorov. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *Intelligent Robots and Systems (IROS)*. IEEE, 2012.

[11] Sergey Levine, Nolan Wagener, and Pieter Abbeel. Learning contact-rich manipulation skills with guided policy search. In *Intern. Conference on Robotics and Automation*. IEEE, 2005.

[12] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016.

[13] Jeroen van Baar, Alan Sullivan, Radu Corcodel, Devesh Jha, Diego Romeres, and Daniel Nikovski. Sim-to-real transfer learning using robustified controllers in robotic tasks involving complex dynamics. *arXiv preprint*, abs/1809.04720, 2018.

[14] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. *Advances in Neural Information Processing Systems*, 2017.

[15] Sébastien Ehrhardt, Aron Monszpart, Niloy J. Mitra, and Andrea Vedaldi. Unsupervised intuitive physics from visual observations. *arXiv preprint*, abs/1805.05086, 2018.

[16] Anurag Ajay, Jiajun Wu, Nima Fazeli, Maria Bauza, Leslie P Kaelbling, Joshua B Tenenbaum, and Alberto Rodriguez. Augmenting physical simulators with stochastic neural networks: Case study of planar pushing and bouncing. *arXiv preprint arXiv:1808.03246*, 2018.

[17] Ofir Nachum, Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *arXiv preprint arXiv:1805.08296*, 2018.

[18] Anusha Nagabandi, Gregory Kahn, Ronald S. Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. *arXiv preprint*, abs/1708.02596, 2017.

[19] Paul F. Christiano, Zain Shah, Igor Mordatch, Jonas Schneider, Trevor Blackwell, Joshua Tobin, Pieter Abbeel, and Wojciech Zaremba. Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv preprint*, arXiv/1610.03518, 2016.

[20] Marc Toussaint, Kelsey Allen, Kevin Smith, and Josh B Tenenbaum. Differentiable physics and stable modes for tool-use and manipulation planing. In *Proc.of Robotics Science&Sysem*, 2018.