State-only Demonstration Driven Exploration for Learning Unstable Dexterous Manipulation and Locomotion

Siqi Shang*, Gagan Khandate*, Matei Ciocarlie

I. Introduction

Model-free RL has been effective for learning numerous diverse sensorimotor learning tasks such as dexterous manipulation [1–5], locomotion [6–10] etc,. With techniques such as domain randomization and policy adaptation, it has also been effective at learning generalist policies that demonstrate these sensorimotor skills on the real hardware.

However, the sample inefficiency of model-free RL remains a major bottleneck that is limiting its widespread adoption. This is especially true when one is interested in learning robust and generalist policies that transfer to real hardware. A number of different sources contribute to the sample inefficiency of model-free RL methods but the effect of exploration, i.e. the lack of ability to effectively explore the state, is most pronounced. Sparse reward structure and long horizon are some of the common reasons for the difficulty in exploration and a number of important ideas have been proposed to address this. Some methods propose to improve exploration through exploration bonuses [11, 12] while other methods use data-augmentation in the form of expert demonstrations [13] or hindsight experience replay [14]. We can also engineer the reset distribution [15] to improve exploration.

In this work, we consider a large subset of sensorimotor learning tasks that are hard to explore due to their intrinsic dynamics. We refer to these as unstable motor control tasks, where the intrinsic dynamics critically limit the exploration of relevant state-space that can be achieved by a sequence of random actions. The state-of-the-art RL methods either fail to learn or exhibit prohibitively large sample complexity. The methods using exploration bonuses also fail as they still rely on random actions for exploration. Obtaining state and action demonstrations is challenging in such tasks with unstable dynamics which prohibits use of such data augumentation methods.

We too follow the approach of engineering the reset distribution but significantly alleviate the issues in engineering this distribution. The goal of this work is to propose a method to generate reset states starting from a single state-only sub-optimal demonstration provided by an external expert.

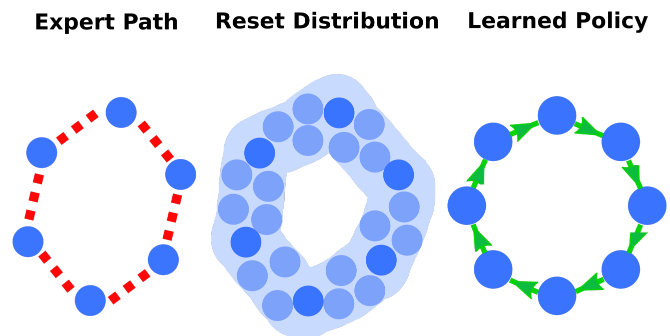

Fig. 1: Engineering the reset distribution from state only demonstration. The solid blobs in blue represent the states of the state-only expert demonstration which are used to sample reset states in their vicinity forming the reset distribution. Training with this reset distribution results in the policy learning to traverse betweeen these states accomplishing the motor control task.

Our initial results show that engineering the reset distribution is critically important for learning complex motor skills with unstable dynamics. We evaluate our methods in simulation and demonstrate that they enable learning for complex motor control problems, in-hand manipulation of objects that require difficult finger gaiting, and climbing with sparse footholds. Preliminary experiments show promise that our method can outperform state-of-the-art RL methods on these tasks.

The main contributions of our work are as follows:

1) We show that reset distribution engineering is critical for learning complex motor control problems characterized by unstable dynamics in large parts of the state space.
2) We propose a simple method to find these reset states starting from a state-only sub-optimal demonstration provided by an external expert. We then show that the resulting states can be used as an effective reset distribution to enable model-free learning without reward engineering.
3) We test our methods in simulation and show that they enable learning for motor control problems such as in-hand manipulation of objects that require difficult finger gaiting, and climbing with sparse footholds. It is, to the best of our knowledge, the first time that successful learning of these tasks has been demonstrated.

* indicates equal contribution

All authors are with Columbia University, New York, NY 10027, USA. Siqi Shang and Gagan Khandate: Department of Computer Science; Matei Ciocarlie: Department of Mechanical Engineering. Emails: siqi.shang, gagan.khandate, matei.ciocarlie @columbia.edu

II. Learning robot control in unstable environments

Our method for sampling reset states consists involves two components: an expert trajectory from demonstration way-points, and a criterion to determine the admissibilty of the sampled state. The expert trajectory is simply the linear interpolant with $u \in [0, 1]$ satisfying the way-points of a state only demonstration i.e ζ_k denotes the expert trajectory for each demonstration where $k = 1, \ldots, K$. The admissibilty criterion determines the boundary of the useful state-space. $\eta < 0$ represents regions of state-space from which recovery into useful state space is deemed impossible. We use this criterion both for sampling a state from which to start a roll-out and for terminating the episode.

The procedure for sampling reset states is shown in Alg 1. Note that it involves perturbation of state obtained via interpolation of the expert trajectory which we found to be critical for learning, especially when using a single demonstration.

Algorithm 1 Sampling reset states using expert state-only demonstration

Input: Expert trajectories $\zeta_k(u)$, Admissibilty criterion η

Initialize reset state buffer, D
for i = 1, …, N **do**
  repeat
    Select expert trajectory ζ_k where $k \sim \{1, \ldots, K\}$
    Sample state $\boldsymbol{s} = \zeta_k(u)$ where $u \sim \mathcal{U}[0, 1]$
    Add perturbation $\boldsymbol{s} \leftarrow \boldsymbol{s} + \boldsymbol{w}$ where $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma)$
    Set simulator state to $\boldsymbol{s}$
    Step the simulation forward by t_s seconds
  until Admissibile state $\eta_t \geq 0$
  Add simulator state $\boldsymbol{s}$ to buffer, $D = D \cup \boldsymbol{s}$
end for

III. Experiments

We test our method in learning in-hand manipulation of objects that require difficult finger-gaiting (ex. an elongated object) and climbing with sparse footholds. In each of the tasks, we assume only proprioceptive sensing, tactile sensing for feedback, and position control for actuation. We use the demonstration gaits shown in Fig 2 for sampling reset states using Alg 1 and collect rollout trajectories by starting exploration from these states during training. For both tasks the admissibility constraint is based on the number of contacts after settling. We require at least four fingertip contacts with the object for in-hand manipulation and at least three contacts of the feet with wall for vertical climbing. To perturb the state we set $\Sigma = \mathrm{Diag}(0.2)$.

Altough our framework is independent of the reinforcement learning method, we use PPO [16] for the results presented here. We also verified that our approach is effective with SAC [11]

In order to evaluate the effectiveness of our proposed method, we compare it with other methods leveraging reset distributions for exploration. The baseline we compare against employs a replay buffer of explored states starting from a fixed initial state as a reset the distribution [17]. We can use episode rewards to priortize sampling state from the replay or simply sample uniformly. We compare against both. We also investigate the effect of perturbation in sampling reset states for both the tasks.

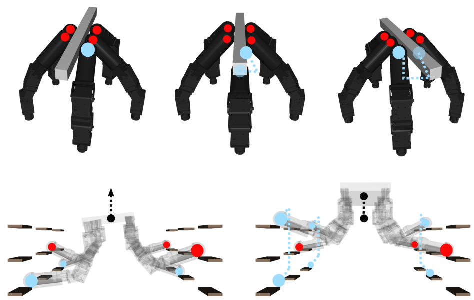

Fig. 2: Expert demonstration for finger-gaiting (top) and multi-legged vertical climbing (bottom)

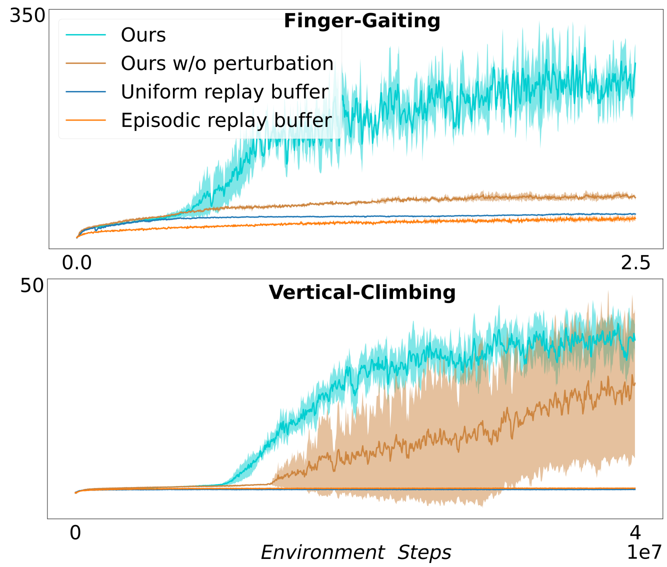

Fig. 3: Training with reset states sampled from expert demonstration. Our method outperforms other methods of reset distribution for hard-to-explore motor control tasks.

As shown by the training curves in Fig 3, our method of using an expert state-only demonstration enables learning these hard-to-explore tasks while the state-of-the-art fail. We also found the perturbation to state is critical, epsecially in learning the finger-gaiting task. While further investigation is required, a likely explantion is that pertubing the state helps in maintaining the diversity of states of the the reset distribution.

In conclusion, this preliminary work shows that engineering the initial state distributions is a promising approach towards learning motor control tasks with uncertainty, including unstable dynamics and a lack of knowledge of the full environment due to intrinsic sensing. In addition, the challenge of designing such distribution can be alleviated using demonstrations and potentially other sources of expert states.

REFERENCES

[1] Wenlong Huang, Igor Mordatch, Pieter Abbeel, and Deepak Pathak. "Generalization in Dexterous Manipulation via Geometry-Aware Multi-Task Learning". In: ().

[2] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. "Learning Dexterous In-Hand Manipulation". In: (Aug. 2018). arXiv: 1808.00177 [cs.LG].

[3] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. "Solving Rubik's Cube with a Robot Hand". In: (Oct. 2019). arXiv: 1910.07113 [cs.LG].

[4] Ilija Radosavovic, Xiaolong Wang, Lerrel Pinto, and Jitendra Malik. "State-Only Imitation Learning for Dexterous Manipulation". In: (Apr. 2020). arXiv: 2004.04650 [cs.RO].

[5] Aravind Rajeswaran, Kendall Lowrey, Emanuel Todorov, and Sham Kakade. "Towards Generalization and Simplicity in Continuous Control". In: (Mar. 2018). arXiv: 1703.02660 [cs.LG].

[6] Joonho Lee, Marko Bjelonic, and Marco Hutter. "Control of Wheeled-Legged Quadrupeds Using Deep Reinforcement Learning". In: *Robotics in Natural Settings*. Springer International Publishing, 2023, pp. 119–127.

[7] Yuxiang Yang, Ken Caluwaerts, Atil Iscen, Tingnan Zhang, Jie Tan, and Vikas Sindhwani. "Data Efficient Reinforcement Learning for Legged Robots". In: *Proceedings of the Conference on Robot Learning*. Ed. by Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura. Vol. 100. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1–10.

[8] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. "Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning". In: (Sept. 2021). arXiv: 2109.11978 [cs.RO].

[9] Jinghong Yue. "Learning Locomotion For Legged Robots Based on Reinforcement Learning: A Survey". In: *2020 International Conference on Electrical Engineering and Control Technologies (CEECT)*. Dec. 2020, pp. 1–7.

[10] Tuomas Haarnoja, Sehoon Ha, Aurick Zhou, Jie Tan, George Tucker, and Sergey Levine. "Learning to Walk via Deep Reinforcement Learning". In: (Dec. 2018). arXiv: 1812.11103 [cs.LG].

[11] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: (Jan. 2018). arXiv: 1801.01290 [cs.LG].

[12] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. "Curiosity-driven Exploration by Self-supervised Prediction". In: (May 2017). arXiv: 1705.05363 [cs.LG].

[13] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations". In: (Sept. 2017). arXiv: 1709.10087 [cs.LG].

[14] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. "Hindsight Experience Replay". In: (July 2017). arXiv: 1707.01495 [cs.LG].

[15] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. "Overcoming Exploration in Reinforcement Learning with Demonstrations". In: (Sept. 2017). arXiv: 1709.10089 [cs.LG].

[16] John Schulman, Xi Chen, and Pieter Abbeel. "Equivalence Between Policy Gradients and Soft Q-Learning". In: (Apr. 2017). arXiv: 1704.06440 [cs.LG].

[17] Arash Tavakoli, Vitaly Levdik, Riashat Islam, Christopher M Smith, and Petar Kormushev. "Exploring Restart Distributions". In: (Nov. 2018).