


Reinforcement Learning for Ballbot Navigation in Uneven Terrain

Achkan Salehi[†] 

Abstract—Ballbot (*i.e.* Ball balancing robot) navigation usually relies on methods rooted in control theory (CT), and works that apply Reinforcement learning (RL) to the problem remain rare while generally being limited to specific subtasks (*e.g.* balance recovery). Unlike CT based methods, RL does not require (simplifying) assumptions about environment dynamics (*e.g.* the absence of slippage between the ball and the floor). In addition to this increased accuracy in modeling, RL agents can easily be conditioned on additional observations such as depth-maps without the need for explicit formulations from first principles, leading to increased adaptivity. Despite those advantages, there has been little to no investigation into the capabilities, data-efficiency and limitations of RL based methods for ballbot control and navigation. Furthermore, there is a notable absence of an open-source, RL-friendly simulator for this task. In this paper, we present an open-source ballbot simulation based on MuJoCo, and show that with appropriate conditioning on exteroceptive observations as well as reward shaping, policies learned by classical model-free RL methods are capable of effectively navigating through randomly generated uneven terrain, using a reasonable amount of data (four to five hours on a system operating at 500hz). Our code is available at <https://github.com/salehiac/OpenBallBot-RL>.

I. INTRODUCTION

The term Ballbot (*i.e.* Ball balancing robot) refers to a family of dynamically stable, underactuated robots that are similar in principle to an inverted pendulum mounted on a spherical base [1], [2], [3]. The single spherical base results in omnidirectional mobility and agility/maneuverability, especially in comparison with more conventional, statically stable wheeled robots. However, ballbot control is notoriously challenging due to several factors such as the system’s high non-linearity and sensitivity to perturbations [3], [4], [5], [6]. Almost all approaches to Ballbot control in the literature are based on methods from control theory that require a mathematical model of the ballbot. Those models generally make several simplifying assumptions [7], [8], [6], *e.g.* that the ball moves in a perfectly horizontal plane, that there is no slippage between the floor and the ball and that there is no deformation.

In contrast, Reinforcement Learning (RL) algorithms in general do not require a mathematical model of the robot, and therefore do not require any of the aforementioned simplifications. Furthermore, in an RL context, policies can easily be conditioned on additional observations such inputs from depth cameras, which can result in more adaptive navigation. Despite those advantages, there has been little to no investigation into the capabilities, data-efficiency and limitations of RL based methods for ballbot control and navigation. Furthermore, there is a notable absence of an open-source, RL-friendly simulator for this task.

Our aim in this paper is to demonstrate the potential of Reinforcement Learning for ballbot control/navigation. To this end, we consider the task of navigation in randomly generated uneven terrain, and show that appropriately shaped rewards coupled with conditioning on exteroceptive information allow classical model-free RL methods to succeed in learning an adaptive and robust

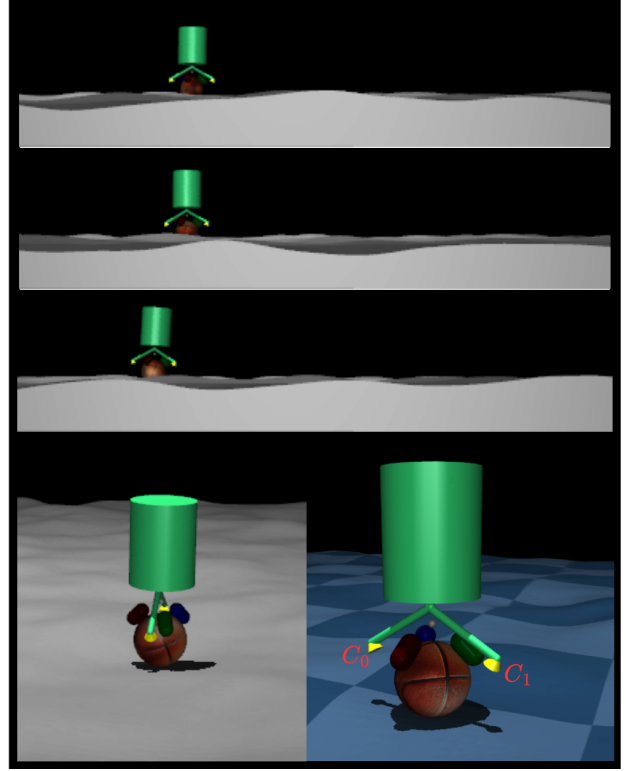


Fig. 1: Screenshots from our open-source simulation, where the learned policy navigates through randomly generated uneven terrain. Note that for simplicity, the three omniwheels controlling the base sphere are modeled as capsules with anisotropic tangential friction. Two low-resolution depth cameras (visible as yellow cones and noted C_0, C_1 in the bottom right image) enable terrain perception. They are both oriented towards the contact point between the ball and the ground.

agent that generalizes to unseen terrain. Figure 1 shows a number of randomly generated terrains as well as our agent navigating those environments. To our knowledge, none of the current control-theory based methods are able to tackle such scenarios.

The simulation environment, which is made available in the public project repository, is based on the high-fidelity MuJoCo [9] physics engine. This choice was motivated by the fact that it offers a good balance between speed and physical fidelity, and that agents trained in MuJoCo have been shown to have high transferability [10].

The paper is structured as follows: the next section is dedicated to preliminaries and problem definition. We present our RL-based approach to ballbot control/navigation in §III, and Section §IV discusses our simulated environment. Experiments are presented in section §V, and are followed by a discussion in §VI. We review the related literature in §VII. Concluding remarks are the subject of section §VIII.

[†]Independent Researcher (formerly with ISIR/CNRS, Sorbonne University), Paris, France. All research in this paper has been done independently. Email: ashsalehi4133@gmail.com.

II. PRELIMINARIES AND PROBLEM FORMULATION

We will first briefly introduce the ballbot and give a high-level overview of the principles underlying most approaches to its control. We then formalize our objectives using Reinforcement Learning terminology.

A. The ballbot

A ballbot is an underactuated, statically unstable but dynamically stable system, similar in principle to an inverted pendulum balancing on a spherical base. While ball drive mechanisms such as inverse mouse-ball drives [1] and spherical induction motors [11] have been used, the most popular drive for ballbots is arguably the omniwheel based setup, an example of which is given in figure 2. An omniwheel applies torque to the sphere along its direction of rotation, while its free-rolling idler rollers allow unconstrained slip along the wheel's axis, enabling omnidirectional motion.

Ballbot control methods in general define $q \triangleq [x, y, \theta_0, \theta_1, \theta_2]^T$ (where $[x, y]^T$ is the planar position and the θ_i specify the body's orientation) and use the Euler-Lagrange Equation

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = \tau_{ext} \quad (1)$$

from which one can derive the equations of motion as

$$M(q)\ddot{q} + C(q, \dot{q}) + g(q) = \tau_{ext} \quad (2)$$

where M, C are respectively the Mass and Coriolis matrices and g is the gravity vector (additional terms such as friction can sometimes be added[7]). These matrices are then estimated depending on the chosen ballbot model (e.g. decoupled motion in xy and xz planes [1], [4] or 3d [12] etc) and then balance/stabilizing/trajectory tracking is done via different combinations of PID and LQR controllers (and related methods) at different levels. Our aim in this paper is to make abstraction of those mathematical models and their underlying assumptions such as the absence of slippage or planar-only motion. Therefore, these control methods will not be discussed any further.

B. Problem formulation

Our aim in this work is to enable the ballbot to navigate in uneven terrains, which we sample from a perlin noise distribution \mathcal{P}_{perlin} [13], [14]. Some example terrains sampled from this distribution can be seen in figures 1 and 8.

Let $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{P}_0, \mathcal{R}, \gamma \rangle$ denote a Markov Decision Process (MDP) with \mathcal{S}, \mathcal{A} respectively the set of states and actions, $\mathcal{P}(s'|s, a)$ the transition probabilities, \mathcal{P}_0 the initial state distribution, $\mathcal{R}(s, a, s')$ the reward function, and $\gamma \in [0, 1)$ the discount factor. Let us also denote $J(\theta) = E_{\tau \sim p_\theta} [\sum_t \gamma^t r_t]$ the discounted cumulative rewards with the expectation taken over trajectories τ sampled from the policy $\pi_\theta \triangleq p_\theta(a|s)$.

Our objective can then be formalized as designing the appropriate \mathcal{S}, \mathcal{R} and policy architecture to enable the ballbot to learn to navigate in terrains sampled from \mathcal{P}_{perlin} , and subsequently solving $\arg \max_\theta J(\theta)$ to obtain the policy.

III. PROPOSED SOLUTION

In this section, we will first discuss the choice of state/observation space. This will be followed by the definition of our reward signal, policy architecture and finally training procedure.

Throughout the rest of the paper, the set of proprioceptive observations at timestep t will be noted $\mathcal{O}_t^{prop} \triangleq (\phi_t, \dot{\phi}_t, v_t, \dot{m}_t, a_{t-1})$. Here, $\phi_t \in \mathbb{R}^3, \dot{\phi}_t \in \mathbb{R}^3, v_t \in \mathbb{R}^3$ respectively denote orientation, angular velocity and velocity in world coordinates. Each component of the $m_t \in \mathbb{R}^3$ vector denotes the angular velocity of the three omniwheels in local coordinates (i.e. relative to their rotation plane). The vector $a_{t-1} \in [0, 1]^3$ is

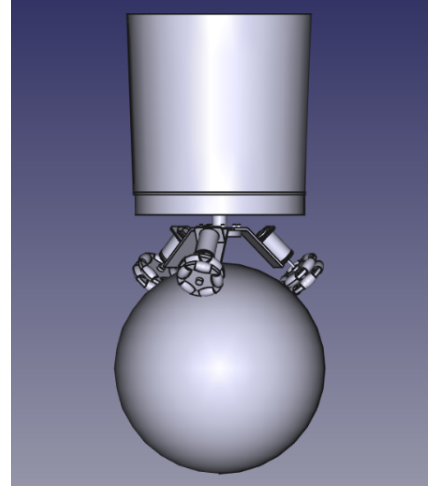


Fig. 2: A CAD model of a ballbot that uses three omniwheels as its ball drive mechanism.

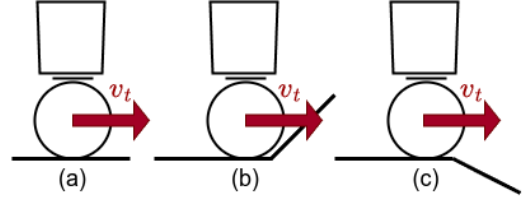


Fig. 3: Using only proprioceptive observations in uneven terrain leads to ambiguities: assuming that the robots in (a), (b), (c) have the same state at time t (in the figure, only the velocity v_t is explicitly shown for clarity) and adopting a model-based view, it is clear that any learned model of the form $s_{t+1} = M(a_t, s_t)$ will be ambiguous, as the state information is not sufficient for predicting whether or not the robot will encounter a flat terrain or an increasing/decreasing slope at $t + 1$. This ambiguity can not be fully resolved by recurrent models or distributional methods based only on proprioceptive data, but can be alleviated by incorporating observations from exteroceptive sensors, such as depth cameras.

the last set of commands sent to each of the three motors.

Observation space. First, note that naively applying RL to a state defined by \mathcal{O}_t^{prop} is sufficient for learning a policy that navigates in a flat plane (the option to train policies in flat terrain, based only on proprioceptive observations, is available in our codebase). However, using only proprioceptive observations leads to ambiguity in uneven terrain (figure 3) and therefore to high uncertainty over actions¹. In order to alleviate that problem, we endow the robot with the ability to observe the terrain via two low-resolution (128×128) depth cameras (figure 5, left), angled towards the contact point between the robot and the floor. In order to increase efficiency, we do not incorporate those images directly into the state. Instead, we pass them to a pretrained encoder and retrieve the low-dimensional embeddings $z^1 \in \mathbb{R}^{20}$ and $z^2 \in \mathbb{R}^{20}$.

A ballbot's control and proprioception usually operate at a high frequency f_{high} that can not be matched by most camera systems. To provide unambiguous observations at the same frequency as

¹While the use of a terrain distribution might be reminiscent of the task distribution considered in meta-learning [15], [16], it should be noted that meta-RL methods are applicable when the physical properties and/or reward functions of a novel task can be estimated with sufficient accuracy from earlier data (e.g. from a short exploration phase). In our case, the problem is one of lack of observability that can not be solved by simply using the history of proprioceptive observations.

f_{high} , we use the most recent depth images $z_{t'}^1, z_{t'}^2$ at each timestep and append the time difference $\Delta_t = (t - t')$ to the observations. The full observation vector is then defined by

$$o_t = (o_t^{prop}, z_{t'}^1, z_{t'}^2, \Delta_t)^T \in \mathbb{R}^{56} \quad (3)$$

where $t' \leq t$.

Policy Network. The policy architecture used in this paper is illustrated in figure 4. Depending on whether the aforementioned depth encoders are trained/fine-tuned or frozen, they can be considered part of the policy. Details can be found in appendix I.

Reward shaping Noting $g \in \mathbb{R}^2$ the desired direction of motion during an episode, we define our reward signal as

$$\mathcal{R}(s_t, a_t) = \alpha_1 v_t^T g + \mathbf{1}_{\{s_t \notin S_F\}} \alpha_2 - \alpha_3 \|a_t\|^2 \quad (4)$$

where $v_t \subset s_t$ is the velocity at time t , the symbol $\mathbf{1}$ denotes the indicator function, $\alpha_i \in \mathbb{R}$ are constants, and S_F is the set of failure states. In this work, a state is considered a failure state if and only if the robot’s tilt from the vertical axis is larger than 20° . Intuitively, the first term in the above rewards the ballbot for moving in the desired direction, the second term gives a positive feedback to the policy for each timestep where it “survives”, and the last term acts as regularization by penalizing large motor commands. Hyperparameters used in this paper are reported in appendix II.

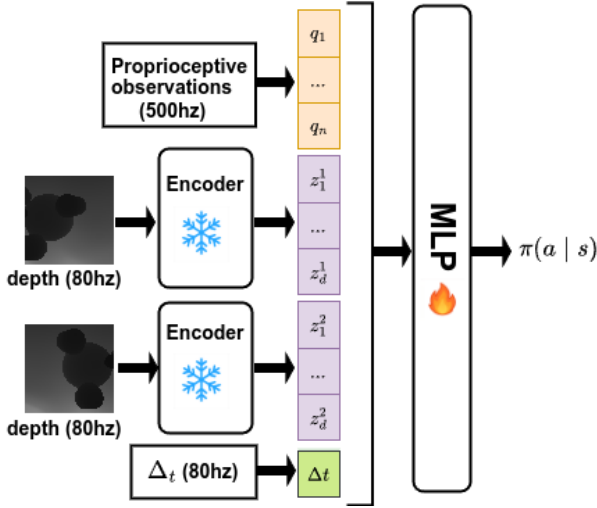


Fig. 4: The policy used to navigate uneven terrain. Two low resolution 128×128 depth images, one from each depth camera, are fed to a pretrained encoder that maps each of them to an embedding in \mathbb{R}^{20} . These embeddings are then concatenated to the proprioceptive observations: orientation, angular velocity, body velocity, angular velocities of each omniwheel (the latter are in local coordinates), as well as the last command vector sent at the previous timestep. To avoid state ambiguities arising from the differences in observation frequencies (500hz for proprioceptive readings vs $\sim 80hz$ for the depth cameras), we also concatenate the time Δt that has elapsed since the last depth image observation was received. This 56 dimensional vector is then fed into a small MLP which is trained to predict torque commands that are sent to the omniwheel motors.

Training. We used the well-known Proximal Policy Optimization (PPO) algorithm [17]. While not as data-efficient as other actor-critic methods such as SAC[18] or model-based approaches such as DreamerV3 [19], it is compute-efficient, has decent exploration capacity (thanks to its entropy term) and its performance is less

sensitive to hyperparameter values, making it ideal for an initial simulation-based proof of concept.

IV. SIMULATED ENVIRONMENT

The simulation environment (available in the public project repository), is based on the high-fidelity MuJoCo [9] physics engine. This choice was motivated by the fact that it offers a good balance between speed and physical fidelity, and that agents trained in MuJoCo have been shown to have high transferability [10].

Figure 5 (left) shows our simulated ballbot, including the two 128×128 depth cameras that are oriented towards the contact point between the ball and the ground. As shown in figure 5 (right), omniwheels are modeled as capsules with anisotropic friction. Note that as of this writing, *The current stable MuJoCo release requires a small modification to allow anisotropic friction to function as intended. This tweak is provided as a patch in our public repository, and was suggested by the developers of MuJoCo*².

Note that robot control is performed at 500hz, and while proprioceptive observations o_t^{prop} are read at that same frequency, the depth cameras operate at $\sim 80hz$. As discussed in section III, unambiguous observations o_t are provided at 500hz by concatenating the relative timestamp Δt to the depth image encodings and proprioceptive readings composing the state.

The distribution \mathcal{P}_{perlin} is based on the `snoise2` function provided by [20].

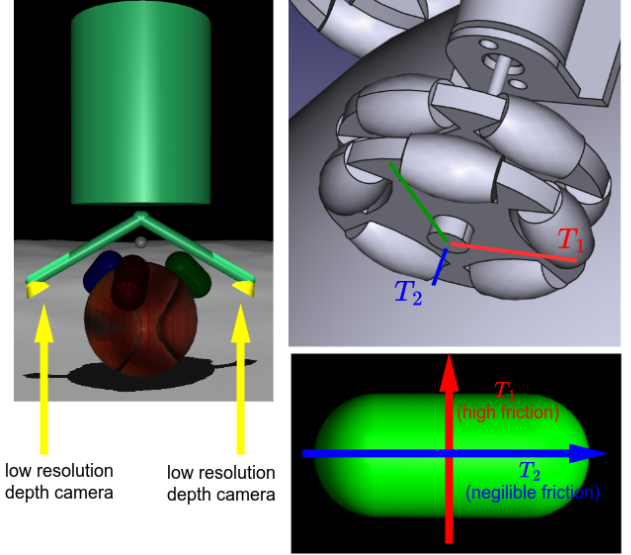


Fig. 5: (left) The ballbot from our open-source simulation. (right) The omniwheels are modeled as capsules with *anisotropic* tangential friction. Friction is high along the T_1 axis, which is the direction along which the wheel applies torque to the sphere. Friction in the T_2 direction, which corresponds to the rotation direction of the omniwheel’s idler rollers, should be negligible.

V. EXPERIMENTS

In this section, we show that our trained policies have effectively learned to navigate in uneven terrain and generalize to previously unseen situations. We also provide some qualitative results, more of which can be found in the project’s public repository. We note that since ballbot navigation in uneven terrain is an extremely under-explored area, we are not aware of any existing method that would serve as a reasonable comparative baseline.

²See the discussion at <https://github.com/google-deepmind/mujoco/discussions/2517>

During training, the maximum length of each episode was fixed to a horizon of $H_{train} \triangleq 4000$ steps, with each episode ending either in case of failure (*i.e.* the robot falling down) or after reaching H_{train} steps. We conducted five training experiments, each with a total training budget of $8e6$ environment steps (one step taking $\sim 2ms$ in simulation time) and a different RNG seed.

The generalization performance of the policy was evaluated at regular intervals, using $N = 10$ randomly sampled, previously unseen terrains. The results are reported in figure 6. In each training experiment, the policy has learned to robustly balance the robot in about $3e6$ timesteps. This is evident from the fact that the average length of evaluation episodes (figure 6, bottom) starts to plateau at ~ 4000 environment steps, which is the maximum that is reachable in our settings (shorter episodes would mean the robot has fallen). At this time, the policy has primarily learned to optimize for the second and third terms of equation 4. During the remainder of each experiment (environment timesteps $\sim 3e6$ to $8e6$), the remaining reward term (goal direction) is optimized, until the reward plateaus slightly below a value of 86.0. Noting $r_{max}^{eval}(i)$ the maximum of average evaluation rewards reached during an experiment i , $\text{median}(\{r_{max}^{eval}(i)\}_{i=1}^5) \approx 85.5$, which given our choice of hyperparameters for the reward function (appendix II), translates to an average speed of $\sim 0.5m/s$ in the desired goal direction.

To further asses generalization, we investigated how the average cumulative reward scales beyond the 4000 steps horizon over which the policies were trained. More precisely, policies trained within the 4000 timesteps limit were deployed in previously unseen environments with horizons of $4000 + \psi$, with $\psi \in \mathbb{N}$ (we sampled 30 environments for each value of ψ). As shown in figure 7, the average rewards gathered by the policies scale almost linearly with horizon, indicating that learned policies do indeed generalize.

An example of policy behavior in a previously unseen terrain is given in figure 8 for qualitative illustration. More qualitative examples are available in our repository.

VI. DISCUSSION

Up to this point, we have demonstrated that with appropriate reward shaping and a sufficiently informative observation space, model-free RL algorithms, in particular PPO, can learn policies that allow the ballbot to navigate in uneven terrain and that generalize well to unseen situations. Interestingly, the $8e6$ transitions that were used to obtain those results are equivalent to ~ 4.4 hours of data, which would be reasonable on a real system. Notice however that we have not aimed for data-efficiency, and that this aspect is likely to be improved by using other actor-critic methods (*e.g.* SAC) or model-based methods such as DreamerV3[19].

While those observations set expectations for what can be expected on a real system and may encourage practitioners to experiment with RL methods, a more promising direction is to transfer policies learned in simulation to real ballbots. While our choice to base the simulation on MuJoCo was partly motivated by its physical high fidelity and high transferability [10], evaluating the sim2real performance of the trained agents is left for future work.

VII. RELATED WORK

The large body of work surrounding ballbot control/navigation, is, as discussed in sections §I and §II, rooted in control theory. Seminal research on ballbots [2], [3], [12], [12] as well as recent works (*e.g.* [4], [7]) are based on mathematical models of the ballbot, which make simplifying assumptions such as perfectly planar motion, the absence of slippage between the floor and the ball, and several other ones regarding friction, elasticity, damping, and so on. In contrast with those methods, we aim for a data-driven approach that is not bound by these assumptions. We note in passing that while Zhou *et al.* leverage RL, they only target balance recovery. Furthermore, the use of RL in their work is limited to

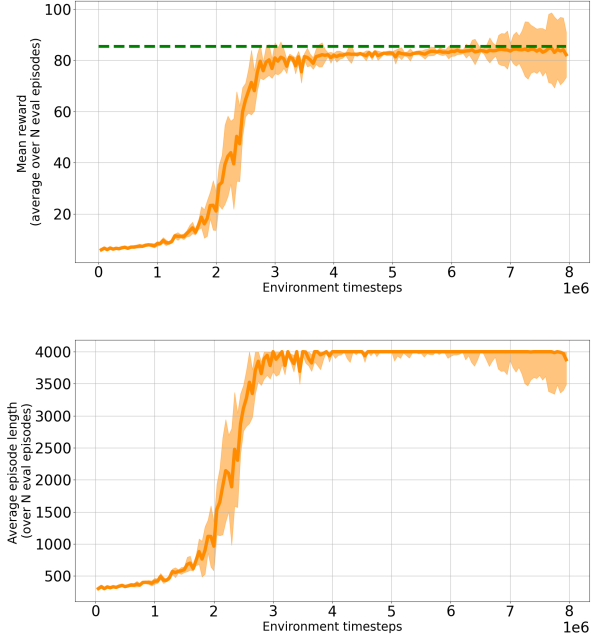


Fig. 6: Evolution of the average rewards and episode lengths on *evaluation* (*i.e.* previously unseen) environments during training, from five different training sessions with five different RNG seeds. Each evaluation is based on $N = 10$ environments. **(top)** Median and standard deviation of the average rewards. The green dashed line indicates the *median* of the maximum average reward reached during the five sessions, and is approximately ~ 85.5 . This reward implies a speed of roughly $0.5m/s$ in the desired direction. **(b)** The evolution of the average episode lengths. The maximum episode length allowed during training episodes was 4000.

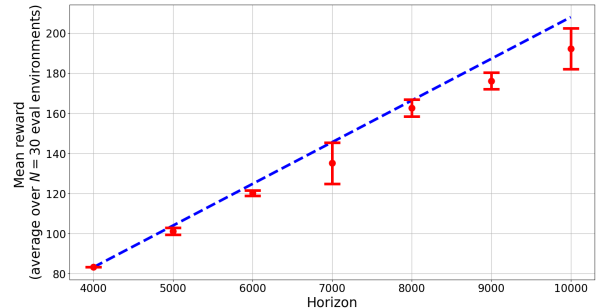


Fig. 7: Policies learned within a horizon of 4000 steps generalize well to longer horizons, as the average reward scales linearly with maximum episode length. Note that each data point (h, r) in the figure results from averaging policy performance over 30 random environments with horizon h , and repeating the experiment with five different RNG seeds.

improving a conventional feedback controller, which still requires a mathematical model of the ballbot.

Several ballbot simulation environments have been developed by different authors (*e.g.* [21], [7], [22], [23]), however, to our knowledge no functional open-source ballbot simulations have been published³. Furthermore, simulations developed as part of prior research projects often rely on physics engines that, unlike MuJoCo,

³A few repositories exists on github, but seem to have been abandoned mid-development.

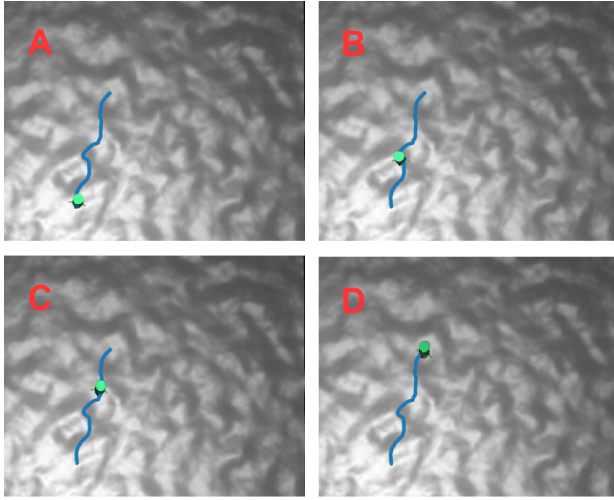


Fig. 8: Top down view from a single navigation episode. The trajectory followed by the robot is illustrated in blue. Note that the task is to move in the general $(0, 1)$ direction, which coincides with the upwards direction in the image. It can be seen that the robot adapts its navigation to the terrain, circumventing risky areas such as sharp local terrain maxima or minima.

are not idea for RL research, mainly due to their computational costs and the difficulty of running parallel environments.

Our research is also related to works in Reinforcement Learning and control that consider operation in difficult terrain [24], [25], [26]. However, algorithms that require adapting the policy based on a short history of observations and actions [24] are not suitable for the ballbot, due to its sensitivity to erroneous controls (the same can be said of meta-learning like approach that require adaptation at deployment). Furthermore, works in this category generally target humanoids or more common embodiments [25]. An exception is the work by Carius *et al.*, which considers ballbot control on uneven terrain as a benchmark to evaluate their proposed approach. However, their observations are only proprioceptive, and their policy operates without any observations about the terrain. Naturally, this leads to lower generalization and additional failure modes compared to our approach.

VIII. CONCLUSION

Prior work has largely overlooked the possibility of applying RL algorithms to ballbot control, despite the fact that their data-driven nature allows them to operate without a mathematical model of the robot, and therefore to avoid unrealistic or limiting assumptions. Furthermore, to our knowledge, no open-source ballbot simulation is available. We have addressed those issues in this work by 1) providing an open-source, RL-friendly simulated ballbot environment based on MuJoCo and 2) demonstrating that model-free RL algorithms, in particular PPO, can train agents that allow the ballbot to navigate in randomly sampled uneven terrain. To our knowledge, none of the previous ballbot control methods have shown this ability. Future directions include improving data-efficiency and evaluating sim2real transferability.

REFERENCES

- [1] U. Nagarajan, G. Kantor, and R. Hollis, "The ballbot: An omnidirectional balancing mobile robot," *The International Journal of Robotics Research*, vol. 33, no. 6, pp. 917–930, 2014.
- [2] T. B. Lauwers, G. A. Kantor, and R. L. Hollis, "A dynamically stable single-wheeled mobile robot with inverse mouse-ball drive," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006*. IEEE, 2006, pp. 2884–2889.
- [3] P. Fankhauser and C. Gwerder, "Modeling and control of a ballbot," B.S. thesis, Eidgenössische Technische Hochschule Zürich, 2010.
- [4] T. Fischer, D. S. Karachalios, I. Zhavzharov, and H. S. Abbas, "Closed-loop identification and tracking control of a ballbot," in *2024 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2024, pp. 337–342.
- [5] C. Xiao, M. Mansouri, D. Lam, J. Ramos, and E. T. Hsiao-Weckler, "Design and control of a ballbot drivetrain with high agility, minimal footprint, and high payload," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 376–383.
- [6] A.-D. Pham, B. H. Thai, P. V. Dang, and N. T. Vo, "Analysis of the parametric configuration impact on ballbot control performance," *International Journal of Mechanical System Dynamics*, vol. 4, no. 4, pp. 446–460, 2024.
- [7] Y. Zhou, J. Lin, S. Wang, and C. Zhang, "Learning ball-balancing robot through deep reinforcement learning," in *2021 International Conference on Computer, Control and Robotics (ICCCR)*. IEEE, 2021, pp. 1–8.
- [8] D. S. Karachalios and H. S. Abbas, "Parameter refinement of a ballbot and predictive control for reference tracking with linear parameter-varying embedding," in *2024 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2024, pp. 688–693.
- [9] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.
- [10] M. Kaup, C. Wolff, H. Hwang, J. Mayer, and E. Bruni, "A review of nine physics engines for reinforcement learning research," *arXiv preprint arXiv:2407.08590*, 2024.
- [11] G. Seyfarth, A. Bhatia, O. Sassnick, M. Shomin, M. Kumagai, and R. Hollis, "Initial results for a ballbot driven with a spherical induction motor," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3771–3776.
- [12] U. Nagarajan and R. Hollis, "Shape space planner for shape-accelerated balancing mobile robots," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1323–1341, 2013.
- [13] K. Perlin, "Improving noise," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 681–682.
- [14] S. Gustavson, "Simplex noise demystified," Linköping University, Tech. Rep., Mar. 2005. [Online]. Available: <http://staffwww.itn.liu.se/~stegu/simplexnoise/simplexnoise.pdf>
- [15] J. Beck, R. Vuorio, E. Z. Liu, Z. Xiong, L. Zintgraf, C. Finn, and S. Whiteson, "A survey of meta-reinforcement learning," *arXiv preprint arXiv:2301.08028*, 2023.
- [16] A. Salehi, S. Rühl, and S. Doncieux, "Adaptive asynchronous control using meta-learned neural ordinary differential equations," *IEEE Transactions on Robotics*, vol. 40, pp. 403–420, 2023.
- [17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [18] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.
- [19] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *arXiv preprint arXiv:2301.04104*, 2023.
- [20] C. Duncan, "noise: Perlin noise library for Python," Python Package Index (PyPI), Mar. 2015. [Online]. Available: <https://pypi.org/project/noise/>
- [21] J. Jo and Y. Oh, "Contact force based balancing and tracking control of a ballbot using projected task space dynamics with inequality constraints," in *2020 17th International Conference on Ubiquitous Robots (UR)*. IEEE, 2020, pp. 118–123.
- [22] S. Y. Song, N. Marin, C. Xiao, R. Okubo, J. Ramos, and E. T. Hsiao-Weckler, "Hands-free physical human-robot interaction and testing for navigating a virtual ballbot," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 556–563.
- [23] A. Nashat, A. Morsi, M. M. Hassan, and M. Abdelrahman, "Ballbot simulation system: Modeling, verification, and gym environment development," in *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE, 2023, pp. 198–204.
- [24] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," *arXiv preprint arXiv:2107.04034*, 2021.

- [25] K. Zakka, B. Tabanpour, Q. Liao, M. Haiderbhai, S. Holt, J. Y. Luo, A. Allshire, E. Frey, K. Sreenath, L. A. Kahrs, *et al.*, “Mujoco playground,” *arXiv preprint arXiv:2502.08844*, 2025.
- [26] J. Carius, R. Ranftl, F. Farshidian, and M. Hutter, “Constrained stochastic optimal control with learned importance sampling: A path integral approach,” *The International Journal of Robotics Research*, vol. 41, no. 2, pp. 189–209, 2022.
- [27] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, “Stable-baselines3: Reliable reinforcement learning implementations,” *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: <http://jmlr.org/papers/v22/20-1364.html>

ACKNOWLEDGMENTS

The author thanks Yuval Tassa for answering their questions regarding anisotropic friction on MuJoCo’s gitlab discussions, and for providing a useful fix.

APPENDIX I NETWORK ARCHITECTURE

The encoder is a vanilla CNN whose details are given in table I. The 56 dimensional feature vector resulting from concatenating proprioceptive observations and depth embeddings (see figure 4) are then passed into a small MLP with 5 linear layers with hidden dimension 128, LeakyReLU and no normalization.

TABLE I: Encoder architecture

Layer	params
Conv1	input_c=1, output_c=32, kernel size=3, stride=2, padding=1
BatchNorm	default
Leaky ReLu	default
Conv2	input_c=32, output_c=32, kernel size=3, stride=2, padding=1
BatchNorm	default
Leaky ReLu	0.01
Flatten	N.A
Linear	num_output_features=20
BatchNorm	default
Tanh	N.A

APPENDIX II TRAINING DETAILS

In our implementation, the coefficient from the reward function (equation 4) where set to $\alpha_1 = 0.01, \alpha_2 = 0.02, \alpha_3 = 0.0001$.

We used the PPO implementation from `stable_baselines3` [27]. The hyperparameters used for training are given in table II. As is usual with PPO, we found that tuning the clip range, entropy coefficient, epochs per updates as well as LR scheduling were critical for training success. Starting with and LR of $1e-4$, we manually scheduled the LR by dividing it by constant factors after a specific number of environment timesteps were reached.

TABLE II: PPO Hyperparameters

Hyperparameter	Value
Discount factor (γ)	0.99
GAE parameter (λ)	0.95
Learning rate	manually scheduled
Clip range	0.015
Entropy coefficient	0.001
Value loss coefficient	2.0
Batch size	256
Epochs per update	5
Steps per rollout	$2048 \times \text{num_parallel_envs}$
Target KL	0.3
number of parallel envs	10
Weight decay	0.01
total timesteps	8e6
Advantage normalization	False

TABLE III: table