Reinforcement learning for bipedal walking robot

Patrick Behr
Technical University of Munich
Chair of Applied Mechanics
Munich, Germany
patrick.behr@tum.de

Daniel Renjewski
Technical University of Munich
Chair of Applied Mechanics
Munich, Germany
daniel.renjewski@tum.de

I. Introduction

For a long time, scientists and engineers have studied the complex interplay of bipedal locomotion mechanics and control using models of different complexity [1] and aiming to identify control objectives [2]. Humans, on the other hand, learn to walk from trial and error. Basic locomotion behaviors have already been learned by computer models in end-to-end reinforcement learning (RL) frameworks [3] [4]. The underlying models, however, often abstract the real-world physics, especially in their contact models. In this study, we aim at developing a self-learned walking policy through RL on an accurate and detailed multi-body model of the JenaFox robot.

II. Methods

In this study, a standard RL setting is considered with an environment E and discrete timesteps in which the agent receives an observation $s_t \in \mathbb{R}^n$, *with* n *number of features* which represents the visible features (sensor data, e.g. joint angles, joint velocity and ground contact signals) of the current state s_t. Based on this state, the agent performs an action $a_t \in \mathbb{R}^4$, representing the voltages for the two hip and two knee joint motors. After each discrete timestep with a sampling time of 2.5 ms, the agent receives new observation s_t and a reward $r_t(s_t, a_t)$ calculated with the current action and observation. This reward scalar is then used in training the agent in an RL setting. The agent is a neural network trained with a Twin Delayed Deep Deterministic policy gradient (TD3) algorithm [4]. Same as the simulation, the training algorithm and agent are realized with the MATLAB environment through the MATLAB Reinforcement Learning Toolbox. For stable bipedal walking, different subgoals need to be achieved [6]. For example:

- Swinging the swing leg
- Transfer support from one leg to the other
- Control Center of Mass velocity

The reward (objective) function can be tuned to achieve the desired behavior. The agent learns with the reward and tries to maximize. In this study, reward shaping is the primary tool to include expert knowledge and guide the agent to the desired behavior. Moreover, the disadvantage of a more accurate but slower model is compensated by using a complex reward function to achieve stable walking through fewer training episodes. The reward function can be divided into positive rewards and negative rewards Eq. 1.

$$r = r_{positive} - r_{punishment} \tag{1}$$

While positive rewards result from the forward speed and a constant small reward for each time step, the negative rewards punish all unwanted behavior. One such unwanted behavior is the boom-bang control, which drives the motors only on its positive and negative voltage limits. This can be prevented by punishing the usage of large action (voltage) of the agent. Other punishments include punishment for violating angle thresholds for the hip and knee joints or penalty for touching down the swing leg before the supporting legs.

III. First Results

Initial surprising solutions have been found which produce unusual yet continuous gait patterns. With adjustments to the initial reward function, the robot hopped on one leg using the other leg as an oscillating weight for balancing.

IV. Conclusion and Outlook

The initial results indicate the algorithm's ability to discover meaningful behavior. With today's computational resources, reinforcement learning becomes viable for complex real-world tasks. We expect an adequately shaped reward function to result in policies that can ultimately be deployed on the physical robot.

References

[1] Full, Robert J., and Daniel E. Koditschek. "Templates and anchors: neuromechanical hypotheses of legged locomotion on land." Journal of experimental biology 202.23 (1999): 3325-3332.

[2] Birn-Jeffery, Aleksandra V., et al. "Don't break a leg: running birds from quail to ostrich prioritise leg safety and economy on uneven terrain." Journal of Experimental Biology 217.21 (2014): 3786-3796.

[3] Kumar, Arun, Navneet Paul, and S. N. Omkar. "Bipedal walking robot using deep deterministic policy gradient." arXiv preprint arXiv:1807.05924 (2018).

[4] Fujimoto, Scott, Herke Hoof, and David Meger. "Addressing function approximation error in actor-critic methods." International Conference on Machine Learning. PMLR, 2018.

[5] Renjewski, Daniel. "An engineering contribution to human gait biomechanics". TU Ilmenau, 2012.

[6] Pratt, Jerry E., and Russ Tedrake. "Velocity-based stability margins for fast bipedal walking." Fast Motions in Biomechanics and Robotics. Springer, Berlin, Heidelberg, 2006. 299-324.