Hey Everyone, We've Solved Bipedal Locomotion*

Yesh Godse, Jonah Siekmann, Helei Duan, Jeremy Dao, Kevin Green, Jonathan Hurst, Alan Fern
Collaborative Robotics and Intelligent Systems Institute, Oregon State University
{*godsey, siekmanj, duanh, daoje, greenkev, jhurst, afern*}*@oregonstate.edu*
*certain terms and conditions may apply

I. Motivation

Creating locomotion controllers for legged robots that work in the real world requires producing behaviors with a high degree of robustness to unknown disturbances such as ground height, slopes, and ground friction. Thus, engineering a locomotion controller by hand for this problem requires an intensive process of hand-tuning gains or explicitly handling disturbances such as early or late contact and other edge cases.

Reinforcement learning presents an alternative to this paradigm. Rather than describing an explicit solution for a problem, it may be easier to describe the problem and learn the solution. Describing the problem, however, is not always an easy task. A common approach is using precise reference trajectories, which are effective at guiding learning, but only describe a small space of behaviors. Closely tracking reference trajectories can prevent controllers from exploring more behaviors and reacting well to disturbances.

II. Principled Cost Functions

One key observation that can be used to reduce the scale of the problem description is that any gait can be defined by the timings of its periodic *swing* and *stance* phases for each foot. A principled cost function based on this observation is to penalize foot forces during swing and penalize foot velocities during stance. We can define the continuous spectrum of all 2-phase bipedal gaits with a simple parameterization of the *ratio* between the swing and stance phases and the *shift* in the periodic clocks for each foot. By providing these parameters to the controller, we can smoothly transition between behaviors like walking, running, and hopping. This framework can also define more complex gaits like skipping by parameterizing the timings of four phases instead of two.

III. Sim-to-Real and Robust Controllers

Training control policies in simulation does not necessarily result in policies that work on hardware in the real world, a phenomenon known as the *reality gap*. A solution to this problem, known as dynamics randomization, involves collecting experience from a simulator which is subject to random initializations of dynamics parameters such as joint inertia or ground friction. Under this scheme, policies must learn to be robust to a variety of dynamical conditions, even without explicit knowledge of such conditions.

This presents a problem for conventional reinforcement learning policies, which are often simple feedforward neural networks. These networks have no mechanism for observing or inferring the dynamics parameters, and so may not be able to choose actions appropriate for the dynamical conditions as the world is only partially observed. Instead, we opt to use memory-enabled neural networks, such as LSTMs, so that dynamical conditions can be inferred or approximated in latent form inside the neural network.

IV. Results

Using principled cost functions and our parameterization of all two-phase bipedal gaits, we produced controllers that learn all common bipedal behaviors (walking, running, hopping, galloping, standing) and can smoothly transition between them [1]. Additionally, by extending the dynamics randomization procedure used in [1] with stair-like terrain randomization, we produced controllers which can walk up and down all kinds of real-world stairs despite being completely blind to the external world. [Video Link].

Acknowledgments

This work was supported by DARPA contract W911NF-16-1-0002, NSF Grant No. IIS-1849343. Thanks to Intel's vLab team for computing resources.

References

[1] Jonah Siekmann, Yesh Godse, Alan Fern, and Jonathan Hurst. Sim-to-real learning of all common bipedal gaits via periodic reward composition, 2021.

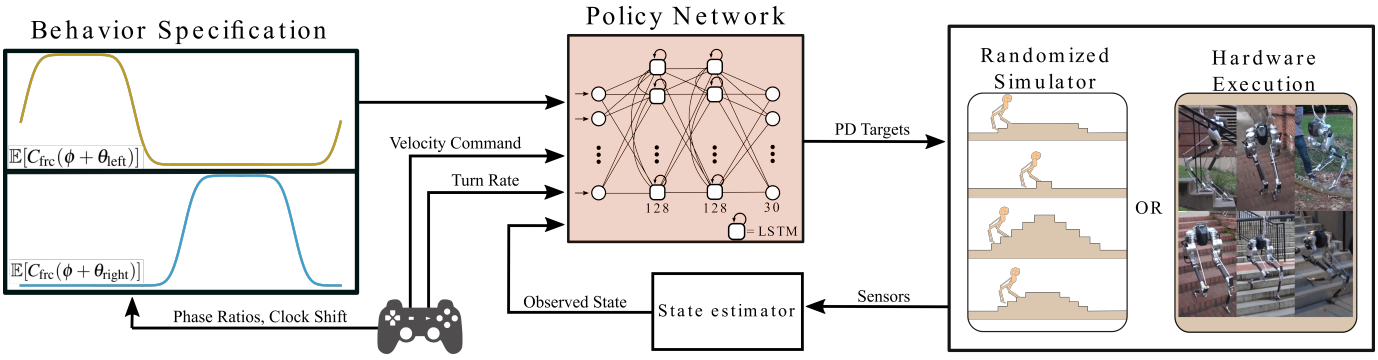

Fig. 1: Our control hierarchy makes use of a recurrent control policy paired with both domain and dynamics randomization, as well as a reference-trajectory-free reward structure allowing for any bipedal gait to be learned.