Learning Locomotion Controllers via Trajectory Optimization

Sergey Levine and Pieter Abbeel
Department of Electrical Engineering and Computer Sciences, UC Berkeley
E-mail: svlevine@eecs.berkeley.edu, pabbeel@cs.berkeley.edu

I. Introduction

Policy search methods can in principle learn controllers for a wide range of locomotion tasks automatically [8, 3, 7, 9, 1]. However, these algorithms typically require a carefully engineered policy class for each locomotion task. A policy class designed for one task, such as fast running, may not be effective for learning another task, such as rough terrain traversal. Recently developed guided policy search methods can learn general-purpose policies represented by neural networks, without task-specific engineering, by using trajectory optimization to find successful task executions [4, 5, 6]. We summarize our recent results on learning locomotion controllers with guided policy search, and present a novel trajectory optimization algorithm that can be used with guided policy search to learn policies even under unknown system dynamics.

II. Guided Policy Search

Guided policy search (GPS) methods optimize the parameters θ of a policy $\pi_\theta(\mathbf{u}_t|\mathbf{x}_t)$ with respect to a cost $\ell(\mathbf{x}_t, \mathbf{u}_t)$ by using trajectory optimization to guide the policy toward good solutions. A sketch of the guided policy search method is provided in Algorithm 1. The key component of GPS is the use of samples around optimized trajectories to improve the policy. These samples serve to illustrate successful task executions, and allow the difficult temporal aspects of the control problem to be handled with trajectory optimization. A second key component is the iterative reoptimization of the trajectories with an objective that encourages low cost and agreement with the current policy π_θ. This adaptation procedure gradually forces trajectory optimization to converge to a solution that is realizable under the policy.

In Figure 1, we show some simulated locomotion controllers trained with GPS under known dynamics. These results include 3D humanoid running on uneven terrain and recovery from strong lateral pushes. The push recovery controller is trained on four different pushes to capture a variety of recovery strategies, and the learned policy can recover from pushes of 250 to 500 Newtons delivered over 100 ms at different points in the gait. All experiments used a single example demonstration to initialize the trajectory, and a simulator of the dynamics was used during trajectory optimization. Videos are available on the websites associated with these papers.[1] [2]

[1] http://graphics.stanford.edu/projects/cgpspaper/index.htm
[2] http://graphics.stanford.edu/projects/gpspaper/index.htm

Algorithm 1 Guided policy search sketch

1: Initialize the trajectories $\{\tau_1, \ldots, \tau_M\}$ with trajectory optimization and/or examples
2: **for** iteration $k = 1$ to K **do**
3: Generate samples $\mathcal{S}$ around $\{\tau_1, \ldots, \tau_M\}$
4: Use samples $\mathcal{S}$ to optimize θ and improve the policy
5: Reoptimize $\{\tau_1, \ldots, \tau_M\}$ to agree with π_θ
6: **end for**

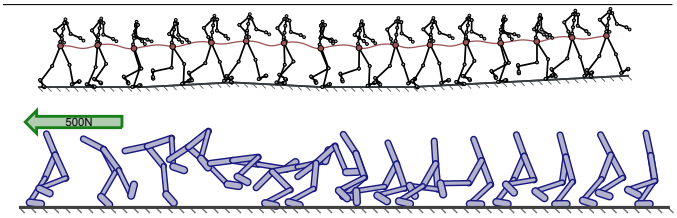

Fig. 1. Controllers trained with guided policy search for running on uneven terrain and push recovery. Adapted from [4, 6].

III. Learning with Unknown Dynamics

Using simulated dynamics models presents serious challenges, since even an accurately modeled robotic platform may respond differently from the simulation in the presence of contacts. We are currently developing a trajectory optimization method for guided policy search that does not rely on known dynamics. Similarly to differential dynamic programming [2], we use linearized dynamics and an LQR backward pass to obtain time-varying linear feedbacks. Rollouts using these feedbacks can then define a new trajectory. To work with unknown dynamics, we construct a stochastic linear Gaussian controller, which induces a distribution over trajectories. We then sample trajectories from this distribution using rollouts of the stochastic controller, use them to refit the time-varying linear dynamics model, and repeat the process.

One challenge with such local dynamics models is that the LQR update can drastically change the trajectory, leading to instability and divergence when new samples are generated from the new trajectory distribution. Inspired by work in model-free policy search [7], we bound the KL-divergence between the new and old trajectory distributions, allowing the method to make consistent, stable progress. The constrained problem can still be solved by an LQR-like method, and can be used with GPS to learn neural network policies under unknown dynamics. Preliminary results show that this method can learn simple walking controllers with 30 minutes of experience, and we are working to further improve sample efficiency.

REFERENCES

[1] M. Deisenroth and C. Rasmussen. PILCO: a model-based and data-efficient approach to policy search. In *International Conference on Machine Learning (ICML)*, 2011.

[2] D. Jacobson and D. Mayne. *Differential Dynamic Programming*. Elsevier, 1970.

[3] Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2004.

[4] S. Levine and V. Koltun. Guided policy search. In *International Conference on Machine Learning (ICML)*, 2013.

[5] S. Levine and V. Koltun. Variational policy search via trajectory optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[6] S. Levine and V. Koltun. Learning complex neural network policies with trajectory optimization. In *International Conference on Machine Learning (ICML)*, 2014.

[7] J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008.

[8] R. Tedrake, T. Zhang, and H. Seung. Stochastic policy gradient reinforcement learning on a simple 3d biped. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004.

[9] E. Theodorou, J. Buchli, and S. Schaal. Reinforcement learning of motor skills in high dimensions: a path integral approach. In *International Conference on Robotics and Automation (ICRA)*, 2010.