Bridging the Gap Between Simulation and Reality (Doctoral Consortium)

Josiah P. Hanna
Department of Computer Science
The University of Texas at Austin
Austin, Texas, U.S.A.
{jphanna}@cs.utexas.edu

ABSTRACT

Transferring robotic control policies — learned in simulation — to physical robots is a promising alternative to learning directly on the physical system. Unfortunately, policies learned in simulation often fail in the real world due to the inevitable discrepancies between the real world and simulation. This thesis aims to bridge the gap between simulation and reality by developing methods for grounding simulation to reality and developing methods for assessing how well a policy learned in simulation will perform *before it is executed in the real world*. We discuss completed work towards a simulation-transfer method and methods of safe policy evaluation. We then present directions for future work in these areas.

Keywords

Reinforcement learning; off-policy evaluation; simulation-transfer

1. INTRODUCTION

A key limitation for widescale deployment of robots is the necessity of expert-designed control software for any situation the robot could find itself in. This approach has limited robotics to controlled, structured environments such as factory assembly lines. If robots are going to be able to leave the factory floor and enter unstructured environments such as homes or workplaces then they must have the capability to autonomously acquire new skills.

Reinforcement learning (RL) provides a promising alternative to hand-coded skills. Unfortunately, the amount of experience required by state-of-the-art RL algorithms is orders of magnitude higher than what is obtainable on a physical robot. Aside from the time it would take, collecting the required training data may lead to substantial wear on the robot. Furthermore, as the robot explores different policies it may execute unsafe actions which could damage the robot. For these reasons, recent empirical successes of reinforcement learning have taken place within simulation. This thesis research proposes to side-step the challenges of robotic reinforcement learning by learning skills in simulation and then transferring the skills to the physical robot.

In theory, the transfer of skills learned in simulation makes state-of-the-art RL immediately applicable to physical robots. Unfortunately, even small discrepancies between simulated physics and reality cause learning in simulation to find policies that fail in the real world. As an illustrative example, consider a robot learning to walk in a simulator where frictional forces are under-modeled. The robot learns it can move its leg joints very quickly to achieve a fast walk. When the same controls are applied in the real world, the walk is jerky and the robot falls over.

An additional limitation of learning in simulation is that policies learned in simulation lack guarantees about their performance in the real world. For any policy learned in simulation, the policy should only deployed if its expected performance is above a predefined threshold with high confidence. Current methods exists for this problem but their data requirements preclude their use in data-scarce settings such as robotics. Thus simulation-transfer methods have no practical method for determining if a proposed policy will work when deployed in the real world.

This thesis research aims to close the gap between simulation and reality through the transfer of simulated robot skills to physical robots. Specifically, this research answers the question, “How can reinforcement learning be applied to learning robot skills in simulation such that those skills can be deployed on a physical robot with high confidence that learning will improve performance?"

2. COMPLETED WORK

To address the inevitable discrepancies between simulation and reality, we have proposed the grounded action transformation (GAT) algorithm for grounded simulation learning. The proposed approach is to augment the simulator with a differentiable *action transformation function*, g, which transforms the robot's simulated action into an action which — when taken in simulation — produces the same transition that would have occurred in the physical system. The simplest instantiation of GAT learns two functions: f which predicts the effects of the physical robot's dynamics and $f_{\texttt{sim}}^{-1}$, which predicts the action needed in simulation to transition from one specific state to another. The transformation function g is specified as $g(s_t, a_t) := f_{\texttt{sim}}^{-1}(s_t, f(s_t, a_t))$ where s_t is the state of the environment and a_t is the action the robot's policy chooses at time-step t. When the robot is in state s_t in simulation and takes action a_t, the augmented simulator replaces a_t with $g(s_t, a_t)$ and the simulator returns s_{t+1} which is the next state that would have occurred on the physical robot. The advantage of GAT is that learning f and $f_{\texttt{sim}}^{-1}$ is a supervised learning problem which can be solved with a variety of techniques such as artificial neural networks trained with backpropagation. Figure 1 illustrates the augmented simulator induced by GAT. Initial results with GAT have shown that grounding the simulator leads to better learning for physical robots — in one instance increasing the walking speed of a bipedal humanoid robot by over 40% [3].

We also present two methods for providing safety guarantees for policies proposed in simulation. This problem falls into the research area known as *high confidence off-policy evaluation* (HCOPE).

Appears in: *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

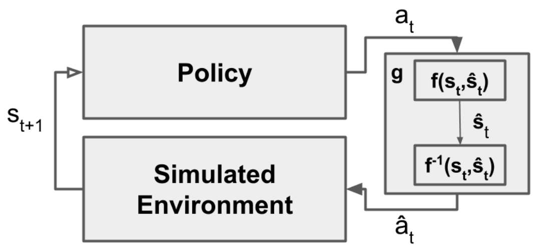

Figure 1: The augmented simulator induced by GAT.

HCOPE methods attempt to place guarantees on the performance of an untested policy using data produced by executing a known, safe behavior policy. Current methods [1, 5] require too much data for a data-scarce setting such as robotics. Towards HCOPE methods for robotics settings, we have investigated the novel combination of bootstrapping with two different model-based off-policy estimators and shown it provides a data-efficient approximate solution to this problem. Bootstrapping is a statistical technique for estimating the distribution of an estimator from which approximate confidence intervals can be derived [2]. The use of bootstrapping means forfeiting strict safety guarantees but can increase data-efficiency. In our work, we propose bootstrapping with the model-based off-policy estimator and bootstrapping with the weighted-doubly robust estimator [6]. Empirical evaluation of the proposed methods showed the combination of bootstrapping with model-based off-policy estimators significantly reduces the amount of data needed to produce tight confidence bounds on policy performance [4].

3. DIRECTIONS FOR FUTURE WORK

In [3], we introduced the grounded action transformation (GAT) method for simulation-transfer. The current instantiation of GAT implements an action-transformation module with maximum likelihood models that predict state changes and the actions needed to produce these state changes in simulation. A limitation of this approach is that the models are not robust to mistakes made at previous time-steps — small errors in prediction can accumulate and lead to the models making worse predictions the longer the robot interacts with the modified simulator. One way to account for the temporal dependencies of actions is to use reinforcement learning to train the action transformation module to choose actions that result in more realistic trajectories over the entire course of interaction. If the action transformation module is represented by a differentiable function approximator then this problem can be solved with policy gradient methods such as REINFORCE [7]. Learning the action transformation module in this way should increase the effectiveness of GAT and help extend its applicability to more tasks.

In [4], we introduced two methods for safe policy evaluation. While empirical evaluation showed the proposed methods decrease data requirements relative to existing methods, so far these gains have only been shown on simple reinforcement learning tasks. The goal of this research is a safety test for simulation-transfer methods and thus the proposed methods need to be evaluated in this context. In robotics, off-policy challenges may arise from data scarcity, deterministic policies, or unknown behavior policies (e.g. experience collected via demonstration). Additionally, robots may exhibit complex, non-linear dynamics that are hard to model. All of these problem characteristics present challenges to existing high confidence off-policy evaluation methods. Understanding and finding solutions for high confidence off-policy evaluation in robot tasks may inspire innovation that can be applied to other domains as well.

Finally, a crucial part of this work is evaluation on challenging and realistic robotic domains. This research will introduces a set of motion tasks for the NAO robot which are applicable to the robot soccer domain: bipedal walking, kicking, and getting up from the ground. These skills are challenging to learn on the physical robot since they involve unstable, dynamic motions which risk damage to the robot if executed poorly. Learning in simulation allows the robot to explore the space of possible motions and learn which ones are unsafe *without* executing them on the physical robot. Furthermore, extensive use of the NAO results in substantial wear on the joints. Thus learning may be intractable for this platform without the assistance of simulation. Evaluating all proposed methods on these tasks is an important step towards establishing their applicability to a wide range of real world robotics problems.

4. CONCLUSION

In summary, this thesis research proposes a simulation-transfer method which allows robotic skills — learned in simulation — to transfer to the real world. In addition, we propose a method for lower bounding the expected performance of skills learned in simulation. These methods will be empirically evaluated across several high-dimensional, continuous control tasks from the robot soccer domain. Taken together, these methods narrow the gap between simulation and reality for reinforcement learning, open up many new promising directions for research pertaining to off-policy evaluation and could dramatically improve the applicability and usefulness of robots in the real world.

Acknowledgements

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by NSF (CNS-1330072, CNS-1305287, IIS-1637736, IIS-1651089), ONR (21C184-01), AFOSR (FA9550-14-1-0087), Raytheon, Toyota, AT&T, and Lockheed Martin. Josiah Hanna is supported by an NSF Graduate Research Fellowship.

REFERENCES

[1] Yinlam Chow, Marek Petrik, and Mohammad Ghavamzadeh. Robust policy optimization with baseline guarantees. *arXiv preprint arXiv:1506.04514*, 2015.

[2] B et al. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[3] Josiah Hanna and Peter Stone. Grounded action transformation for robot learning in simulation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017.

[4] Josiah Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th International Conference on Autonomous Agents and Multi-Agent Systems*, 2017.

[5] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence off-policy evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

[6] P.S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning, ICML*, 2016.

[7] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.