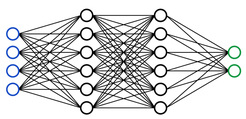
Bridging the Simulation-to-Real Gap in Deep Reinforcement Learning

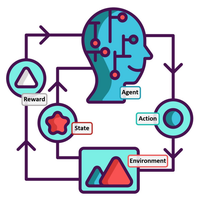
Nathan Van Utrecht, Prajwal Koirala, Cody Fleming Ph.D.

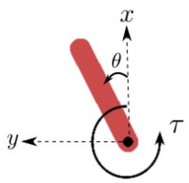
Department of Mechanical Engineering | Iowa State University

Background

Deep reinforcement learning (RL) is a **powerful control paradigm** which has proven to be a promising alternative to traditional algorithms. For complex environments, deep RL relies on **simulated environments** for training since it can take millions of interactions to learn an optimal policy. However, simulations are **imperfect representations** of real-world systems, so a gap between performance in simulation and performance in the real-world is formed. This performance disparity is known as the **simulation-to-real gap**, and solving it is the focus of this research.







Methods

The current state of the art algorithms, like **Soft Actor-Critic (SAC)** [1] and **Proximal Policy Optimization (PPO)** [2], are model-free meaning they do not learn the dynamics of the system, but rather the actions that are most likely to maximize the reward. This work investigates how using the model-based **Short Horizon Actor-Critic (SHAC)** [3] algorithm with a differentiable simulator compares to these model-free algorithms. To emulate a real-world environment, models were tested for their **robustness to observation noise and system parameter tweaks**. All experiments were averaged over **five different seeds** on the classical **pendulum control** problem.

Acknowledgements

Thanks to Prajwal Koirala for his helpful feedback and ideas, Cody Fleming for his guidance and support, and to the TrAC REU for making this research possible.

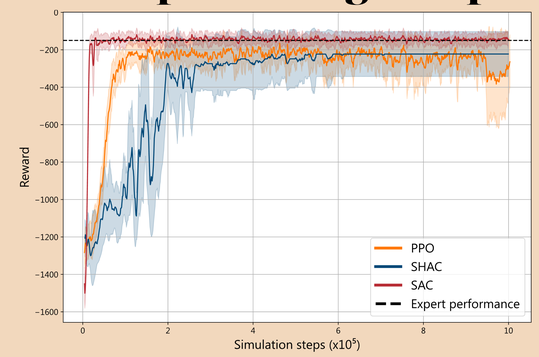
Auxiliary Information

All graphs created from the data, hyperparameters used for the experiments, and references used for this research.

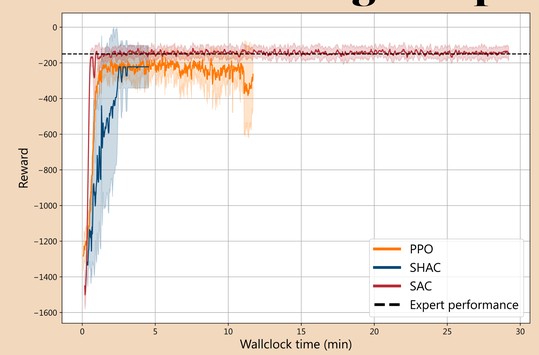
Results

Across every test, **SAC outperformed** both **PPO and SHAC**. It converged to expert level performance with **5x** less timesteps, and its policies were **more robust** to **observation noise and system parameter tweaks** across the board. However, its sample efficiency comes at the cost of **increased training time** relative to PPO and SHAC (as shown in the auxiliary information).

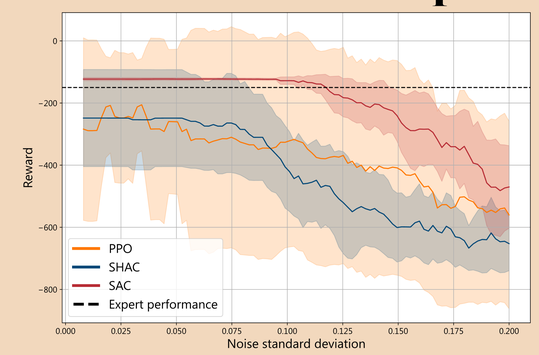
Timestep Learning Comparison



Wallclock Learning Comparison



Noise Robustness Comparison



Conclusions

Why does this matter?

- These results **go against** the findings in the SHAC paper, but this is likely due to **optimizations in their simulator** that were not used for this environment
- SHAC can obtain **expert performance**, but fails with certain seeds likely due to getting stuck in **local minima**
- **Model-free algorithms** are likely more robust and efficient than **model-based algorithms**

Future Work

These findings are **preliminary results** for later experiments which will use the more complex **double pendulum environment**. Similar methods will be used to test the algorithms in simulation, and a **physical system of the environment** will be developed to test the **transferability of each algorithm's policy**. This will give a better representation of how well each model is able to **bridge the simulation-to-real gap**.

Translational AI Center