Specialized Ensembles of Compensator Policies for Environment Mismatch

Anthony Simeonov, Dimitri Schreiber, Daniel Shak

Introduction: Environment Mismatch

Control policies developed in simulation often fail or perform poorly when transitioned to the real-world. This is often caused by a mismatch between the real-world dynamics and the simulated model.

Common sources of mismatch include:
- Inaccurate models
- Uncontrolled disturbances
- Nonstationary dynamics

However, the use of sim-to-real policy transfer has high potential in applying reinforcement learning to robotics by reducing training on physical systems. We therefore are interested in dealing with the issues of environmental mismatch so that policy transfer can be more effective.

Problem Statement: RL Compensator

- Scenario: nominal controller designed to perform well on a particular system (simulation), wish to apply this controller to a similar system with different dynamics (physical system)

$$s_{t+1} = f(s_t, a_t) \qquad s'_{t+1} = \hat{f}(s'_t, a_t) \qquad a_t = F_f(s_t)$$

- Due to the mismatch between environments, the nominal, controller performs poorly on the real system,

$$\|s'_{t+1} - \hat{f}(s'_t, F_f(s'_t))\| > 0$$

- We wish to develop a compensator for the inaccuracies of the nominal controller using a reinforcement learning policy

$$\hat{a} = F_f(s'_t) + \pi_{comp}(s'_t)$$

- We are additionally interested in learning an ensemble of several compensator policies which are weighted according to a mismatch source discriminator

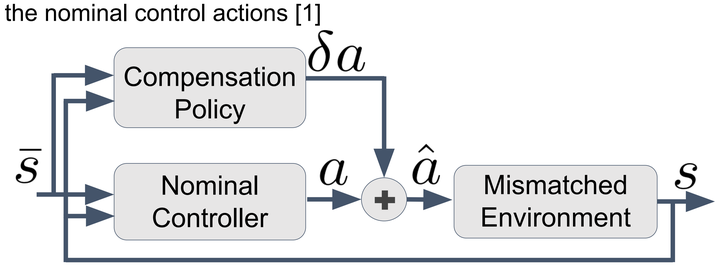
$$\hat{a} = F_f(s'_t) + \sum_{i=1}^{n} w_i \pi_i(s'_t) \qquad \vec{w} = D(\{s'_t\})$$

- The policies will be trained using a reinforcement learning algorithm with reward signals proportional to the negative error on the mismatched system

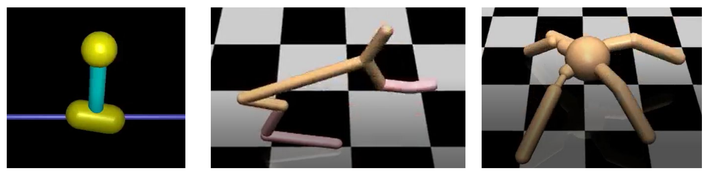
$$R \propto -\|s'_{t+1} - f'(s'_t, F_f(s'_t))\|$$

Single Policy Compensator Baseline

A single policy is trained as a compensator to output actions which sum with the nominal control actions [1]



We evaluate this baseline in OpenAI gym MuJoCo environments, where the dynamics are modified (see below)



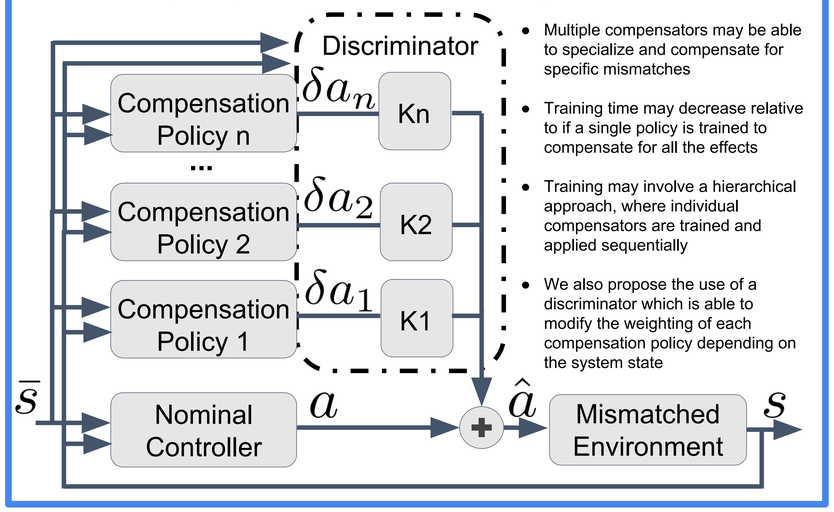
Inverted Pendulum
- Tilted cart
- End effector load
- Friction on hinge/slide
- Mass and inertia distribution

Half Cheetah
- Leg lengths
- Torso size and mass
- Joint friction, stiffness, damping
- Foot-ground friction

Ant
- Leg lengths
- Torso size and mass
- Joint friction, stiffness, damping
- Foot-ground friction

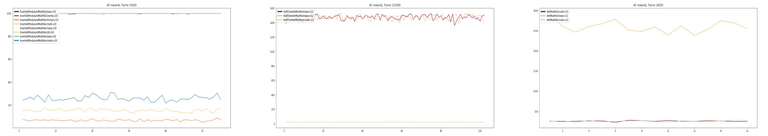
Specialized Compensator Ensemble

Intuition: A set of more specialized compensators have to each make less performance tradeoffs by focusing on specific mismatch sources
This should improve performance and data efficiency.

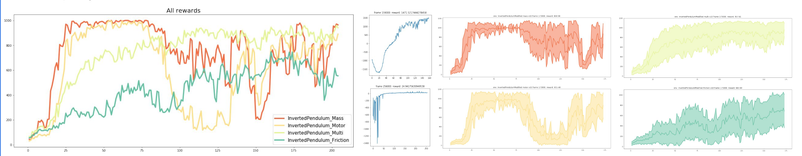


- Multiple compensators may be able to specialize and compensate for specific mismatches
- Training time may decrease relative to if a single policy is trained to compensate for all the effects
- Training may involve a hierarchical approach, where individual compensators are trained and applied sequentially
- We also propose the use of a discriminator which is able to modify the weighting of each compensation policy depending on the system state

Baseline Results



The nominal controller performs well on the unmodified environment but, there are several environments where the performance suffers significantly due to the different dynamics between the environment the controller was developed/trained on and deployed on.



Single compensator trained using PPO algorithm in modified environments

Training process of PPO compensator on InvertedPendulum environments with different modifications

Discussion and Future Experiments

Compensation Performance Tradeoff
- We note that on the inverted pendulum experiment where the compensator was trained on a single modified environment with multiple modifications, but tested on environments with individual modifications, there are performance tradeoffs on the single-mod environments throughout training. This indicates the potential usefulness of incorporating multiple specialized compensators that are decoupled.

Challenging Environments (contacts)
- We also note that this baseline solved the modified Inverted Pendulum, partially solved the modified Half Cheetah (flipping on its back and vibrating forward) and did not solve the modified Ant.

Forcing Specialization Through Hierarchical Ensemble Training
- Individually training for specific errors and disturbances to create a collection of compensating policies.

Bounded Compensator Action Space
- Because the nominal controller should be a better initialization for good performance on the modified environments than starting from scratch, the compensator policies should be constrained in their exploration of states and actions.

Mismatch Source Discrimination
- Different environments may display state dependent mismatch. Therefore, a discriminator trained to properly weight the contributions of the compensators may prove helpful on more challenging dynamical systems

References

[1] I. Koryakovskiy, M. Kudruss, H. Vallery, R. Babuška, and W. Caarls. Model-plant mismatch compensation using reinforcement learning. IEEE Robotics and Automation Letters, 3(3):2471–2477, July 2018.