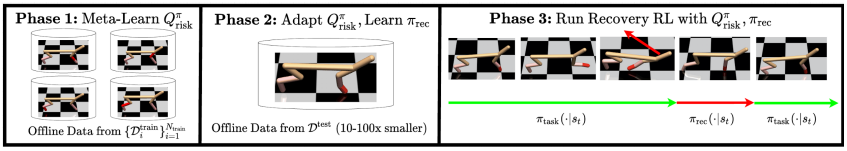
Summer 2021 BAIR Commons Update

MESA: Offline Meta-RL for Safe Adaptation and Fault Tolerance

Safe exploration is critical to deploying reinforcement learning algorithms in risk-sensitive environments. Recent work encourages safety by learning risk measures, which measure the agent's probability of violating constraints. However, learning such risk measures requires significant interaction with the environment, resulting in excessive constraint violations during learning. Furthermore, these measures are not easily transferable to new environment dynamics, which is a frequent issue with real robotic hardware due to issues such as battery capacity losses and motor wear-and-tear when deploying an agent in an environment which differs from that in which unsafe data was collected. In this work, we cast safe exploration as an offline meta-reinforcement learning problem, where the objective is to leverage examples of safe and unsafe behavior across a range of different robot environment dynamics to quickly adapt learned risk measures to modified environments with unseen, perturbed dynamics. For example, consider a legged robot which suddenly loses power in one of its joints: while power loss in this specific joint may not have been previously observed, risk measures learned from prior experience of different power losses can be used to quickly learn how to be safe in this new setting. We propose MESA, an approach for meta-learning a risk measure for safe reinforcement learning. Simulation experiments across 5 continuous control domains suggest that MESA can leverage datasets from prior environments to reduce constraint violations in a similar but unseen environment by up to a factor of 2 while maintaining task performance compared to prior algorithms that do not learn transferable risk measures.



Approach

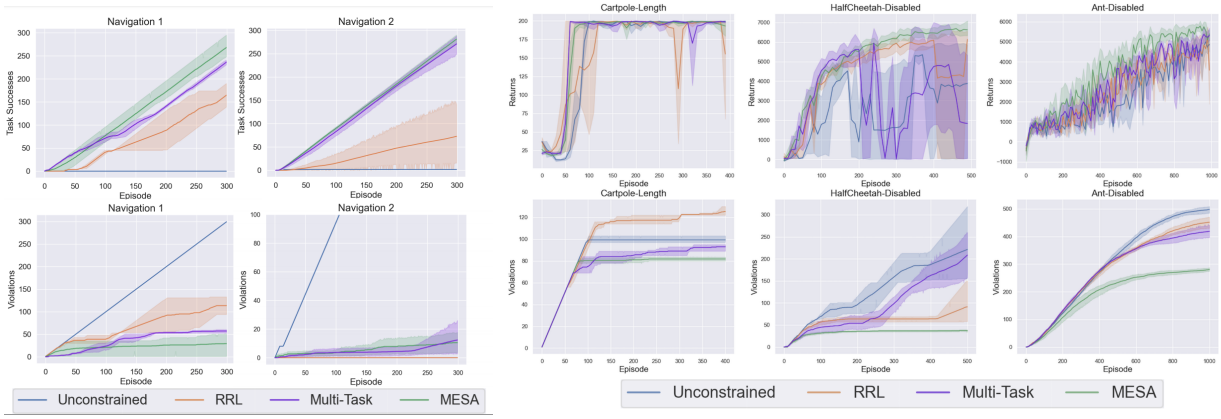
MESA is a meta-RL algorithm for pre-training learned risk measures that can quickly adapt safety measures to new, never-seen-before environments. MESA consists of three phases:

- **Phase 1 (Meta-Learning the Safety Critic):** MESA meta-learns the risk measure (safety critic) across N different datasets, which represent different environments or tasks. We employ a simple offline meta RL algorithm that wraps MAML around DDPG for inner adaptation.
- **Phase 2 (Test Environment Adaptation):** After meta-learning the safety critic, MESA rapidly adapts the risk measure on a very small dataset sampled from the new test environment. In practice, the dataset can be up to 40-100x smaller than that of prior methods.
- **Phase 3 (Online Learning):** After adapting learned risk measures, these risk measures can then be used to help guide a task policy to safely explore and learn in the new test environment.

Results

We investigate how effectively MESA can transfer learned notions of safety by evaluating MESA on 5 environments, roughly partitioned into navigation and locomotion environments:

- **Navigation Environments:** In Navigation 1 and Navigation 2, the agent learns to navigate from a beginning position to the goal while avoiding the obstacles (red walls). Different tasks in Navigation vary by scaling the transition dynamics on each action.
- **Locomotion Environments:** In the Cartpole-Length, the goal is to keep the pole balanced on the cart while minimizing the number of times the pole falls beneath the rail or moves off the rail. Tasks vary by varying the length of the cart pole. Lastly, in the HalfCheetah-Disabled and Ant-Disabled tasks, the objective is to learn how to move forwards while minimizing the number of collisions with the ground of the head (HalfCheetah) or torso (Ant) during training. Tasks vary the joint/limb that is disabled.



Two metrics are employed: task successes (when the agent reaches the goal) and constraint violations committed during adaptation to the test environment.

Experimental results suggest that meta-learning across datasets (MESA) does just as well as using multi-task learning (Multi-Task), while maintaining similar performance. We hypothesize the gap is small because in the Navigation environments, particularly Navigation 2, the space of safe behaviors does not change significantly as a function of the system dynamics, making it possible for the Multi-Task comparison to achieve strong performance by simply learning the safety critic jointly on a buffer of all collected data.
We also evaluate the performance of MESA on the set of 3 locomotion environments. In Cartpole-Length, the Recovery RL comparison exhibits the most constraint violations, which suggests that the set of transitions from the test environment is too small to learn a sufficiently accurate safety critic. The MESA and the MultiTask comparison achieve somewhat similar performance in this environment, but both are able to leverage their learned prior from the training tasks to achieve somewhat fewer constraint violations than the Unconstrained and Recovery RL comparisons. We hypothesize that the difference in the dynamics between the training and test environments is too small for MESA and the Multi-Task comparison to gain sufficient benefit from data in the training environments.
The HalfCheetah-Disabled and Ant-Disabled environments present settings in which the dynamics are much more different between the training and testing environments. Accordingly, MESA significantly outperforms the other methods, including the Multi-Task comparison. We hypothesize that this is because the different training environments are sufficiently different in their dynamics that a safety critic and recovery policy trained jointly on all of them is unable to accurately represent the boundaries between safe and unsafe states. Thus, when adapting to

an environment with unseen dynamics, the space of safe behaviors may be so different than in the training environments that the Multi-Task comparison cannot easily adapt. MESA mitigates this issue by explicitly optimizing the safety critic for rapid adaptation to environments with different dynamics.