Model-Free Imitation Learning with Policy Optimization: Supplementary Material

Jonathan Ho HOJ@CS.STANFORD.EDU
Jayesh K. Gupta JKG@CS.STANFORD.EDU
Stefano Ermon ERMON@CS.STANFORD.EDU
Stanford University

Here, we give extra information regarding the environment and algorithm setups for our experiments.

Gridworld We used tabular policies for IM-REINFORCE with parameters θ_{sa}, with action probabilities $\pi(a|s) \propto \exp(\theta_{sa})$; we used value iteration to obtain Q values for the gradient formula (8). We solved the linear programs for LPAL with Gurobi 6.5.1, and we defined the policies learned by behavioral cloning as simple lookups into expert data (for states unseen in the expert data, a random action is chosen). All timing tests were performed on an 4-core 3.6GHz Intel i7-4790 CPU.
The gridworlds we used resembled those of Abbeel and Ng (2004). Each was a square grid of states, with five actions (an action to move in each compass direction, and one for staying in place) that fail with 30% probability and result in a random move. Each test consisted of 40 trials. Costs were generated in 8×8 non-overlapping regions in the gridworld, giving one basis function for $\mathcal{C}_{\text{convex}}$ per region.

Waterworld We first ran TRPO for various iteration counts to obtain expert policies achieving various expected costs according to the true cost function, which penalized application of control, and assigned differing cost values to the targets of different colors. Then, we executed each expert policy to yield 25 trajectory samples, and then we ran IM-REINFORCE and IM-TRPO both for 100 iterations to imitate each expert policy. The trajectories were 500 timesteps long, and the discount factor was 0.99. We gave both algorithms 50 rollouts per iteration. Excess costs were computed by averaging over 100 rollouts.

Highway For each driving style, we ran IM-TRPO for 500 iterations, each collecting 20000 state-action pairs with simulation. The datasets and dynamics model were identical to the ones used by Levine & Koltun (2012). We evaluated our policies with the same measurements used by Levine & Koltun, averaged over 50 rollouts.