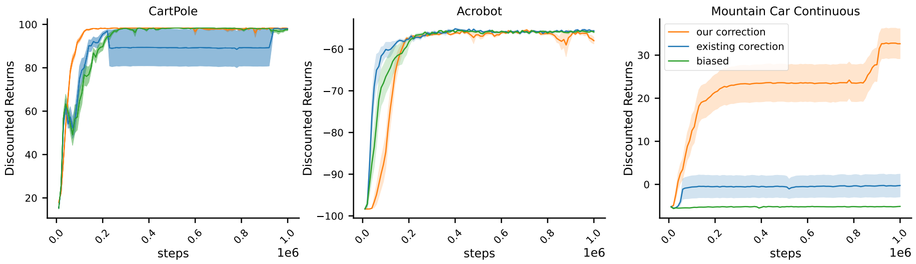

Figure 1: Learning with discount factor 0.99.

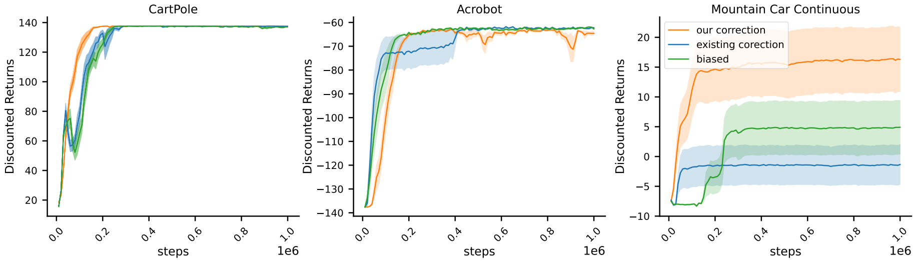

Figure 2: Learning with discount factor 0.993.

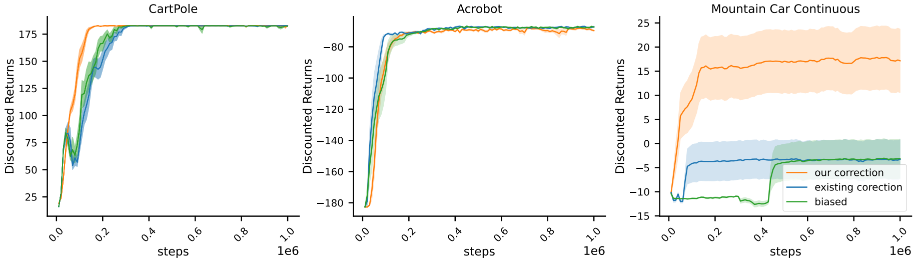

Figure 3: Learning with discount factor 0.995.

First, we show the learning curves for our correct batch-actor-critic and corresponding baselines in Figure 1-3.

The results for three discount factor values are similar. All three algorithms show similar performances on Acrobot. But BAC with our correction in orange dominates CartPole and continuous MountainCar.

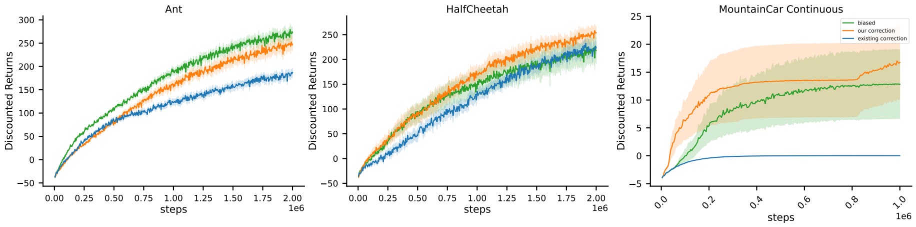

Figure 4: Learning with discount factor 0.99.

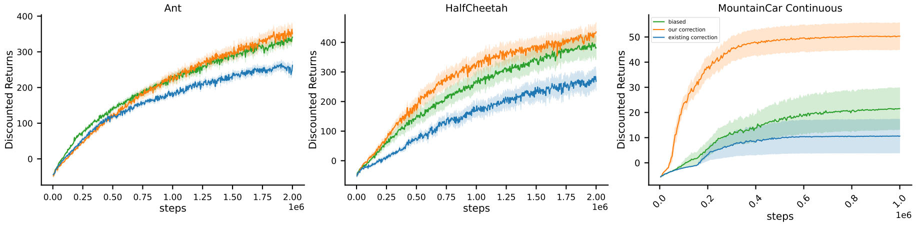

Figure 5: Learning with discount factor 0.993.

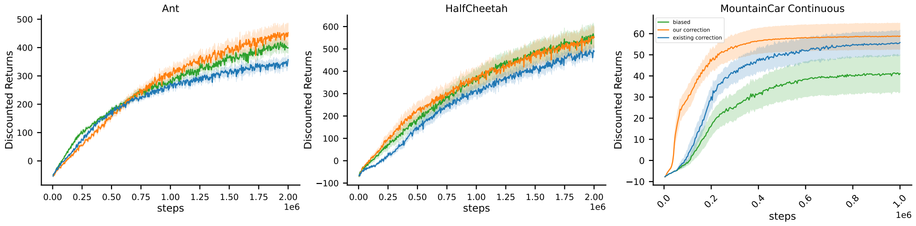

Figure 6: Learning with discount factor 0.995.

The existing correction hinders the final performance of the agent as shown in blue. It causes worse returns than the original algorithm, except for MountainCar Continuous with discount factor 0.995. However, PPO with our averaging correction in orange successfully matches the biased algorithm's performance and can even improves the learning speed and final performance in MountainCar Continuous.