Reinforcement Learning - Continuous Cartpole

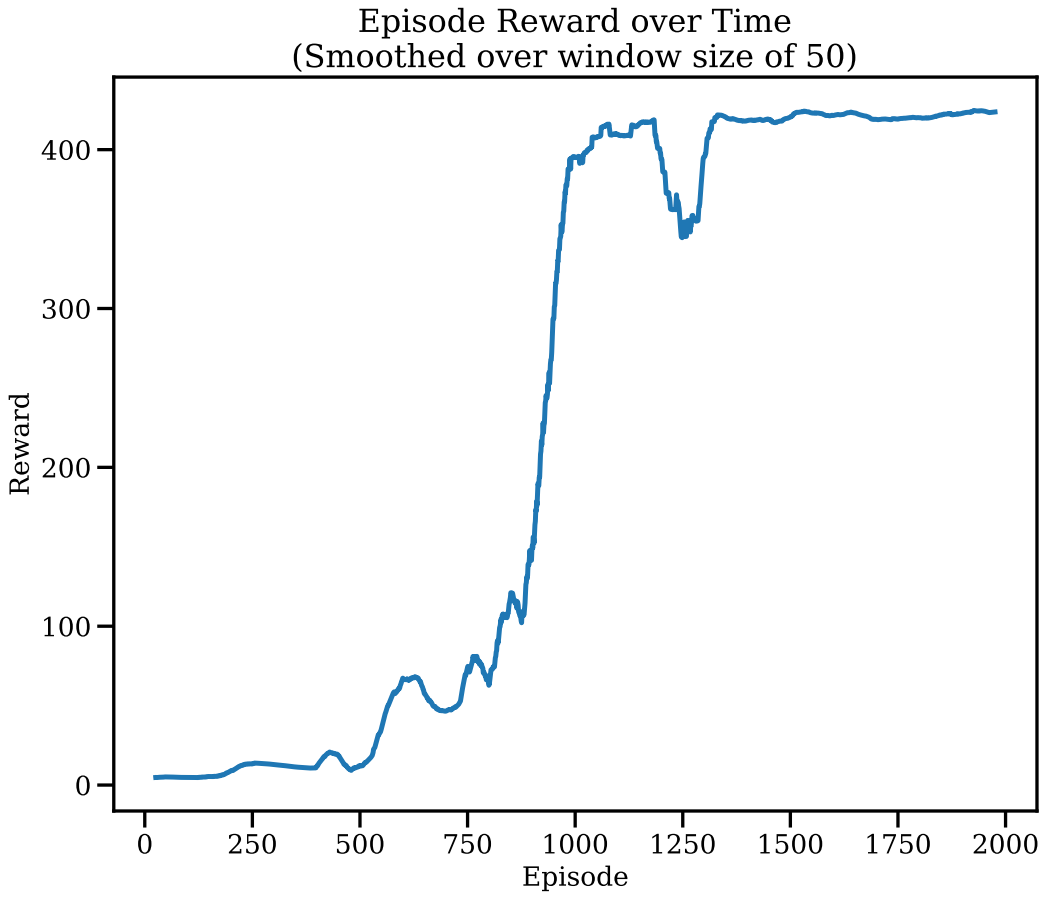
Aron Distelzweig

Project

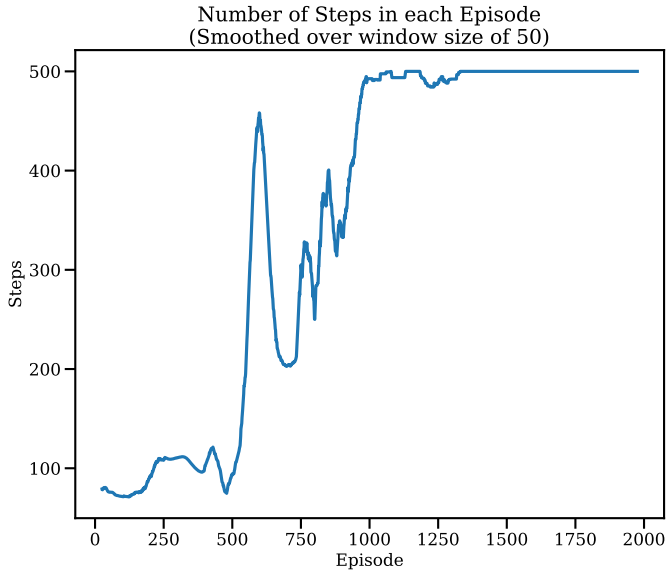
Approach

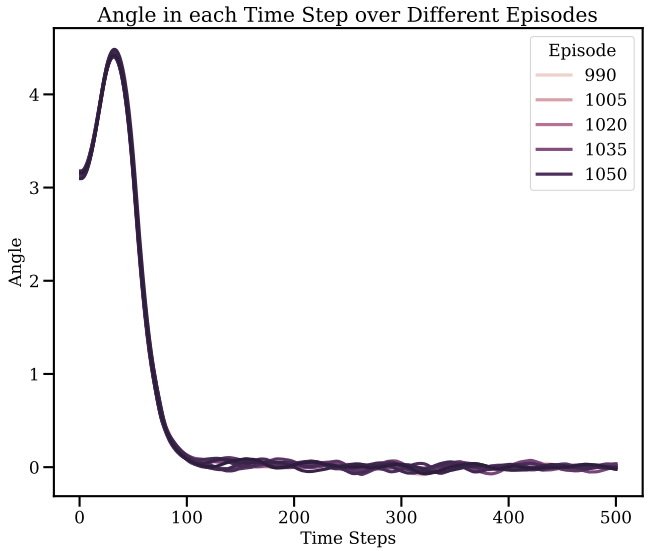
- Actor Critic method because of continuous action and state space
- Reward function: The smaller the angle, the higher the reward
- First approach: Deep Deterministic Policy Gradient (DDPG)
- DDPG shows unstable performance
- Turn DDPG into TD3 by adding the following three adjustments
  - Clipped Double Q-Learning
  - Delayed Policy Updates
  - Target-policy smoothing

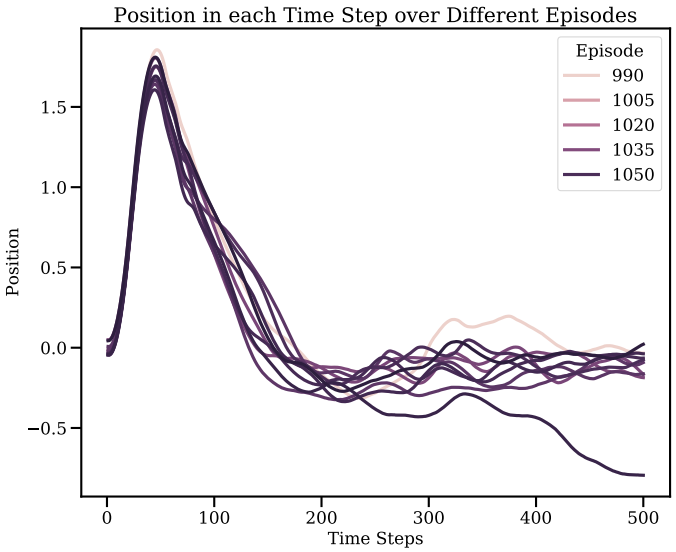
Results - TD3

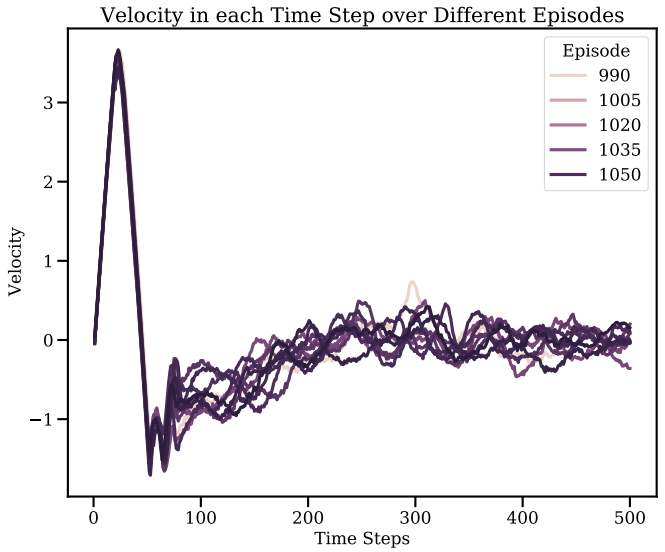

Episode Reward over Time
(Smoothed over window size of 50)
Reward
0
100
200
300
400
0
250
500
750
1000
1250
1500
1750
2000
Episode

Results - TD3

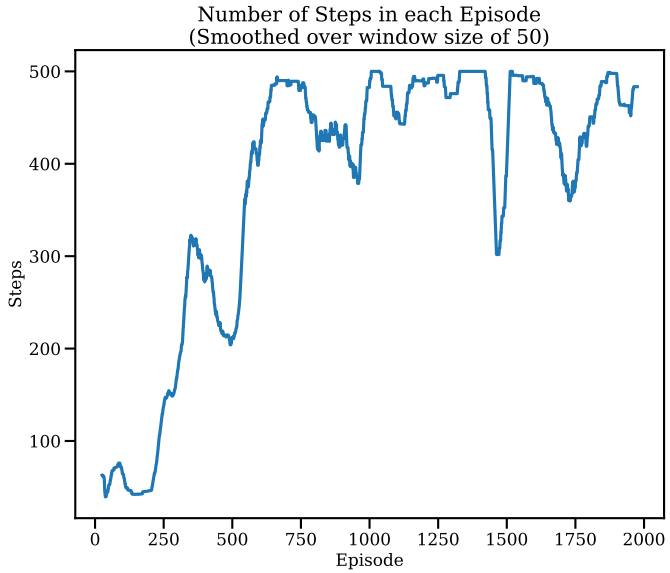

Number of Steps in each Episode
(Smoothed over window size of 50)
Steps
Episode
500
400
300
200
100
0
250
500
750
1000
1250
1500
1750
2000

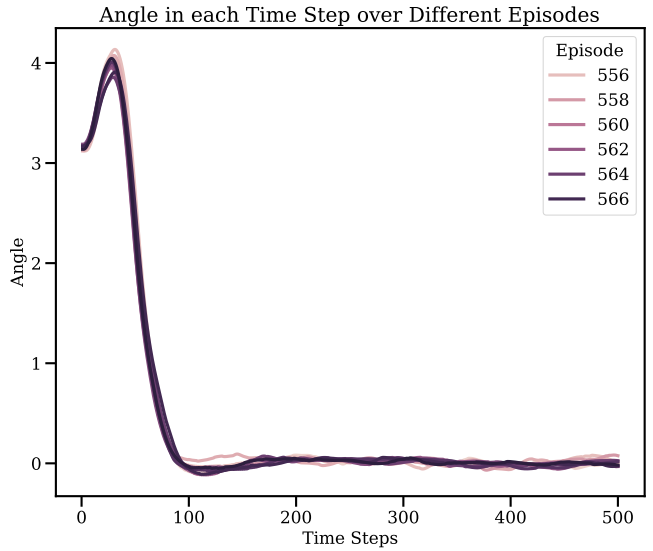

Angle in each Time Step over Different Episodes
Episode
990
1005
1020
1035
1050
Angle
Time Steps
4
3
2
1
0
0
100
200
300
400
500

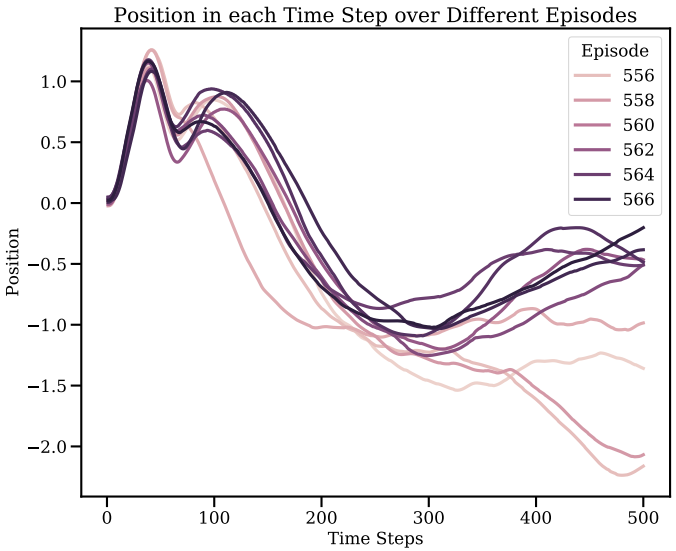

Position in each Time Step over Different Episodes
Episode
990
1005
1020
1035
1050
Position
Time Steps
1.5
1.0
0.5
0.0
−0.5
0
100
200
300
400
500

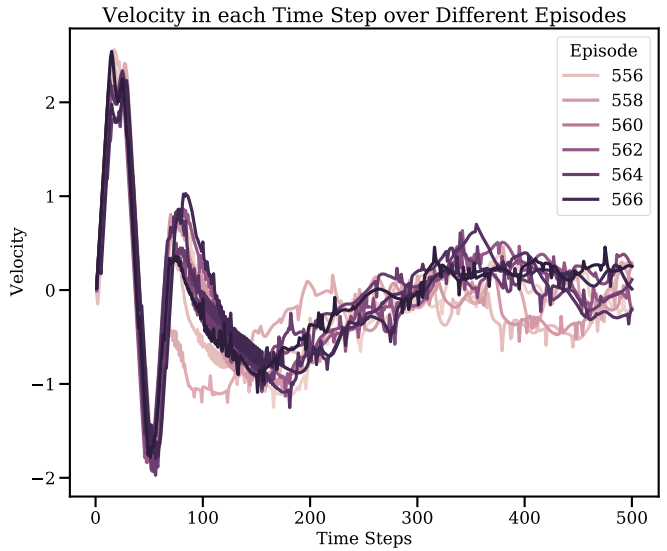

Velocity in each Time Step over Different Episodes
Episode
990
1005
1020
1035
1050
Velocity
Time Steps
3
2
1
0
−1
0
100
200
300
400
500

Results - DDPG

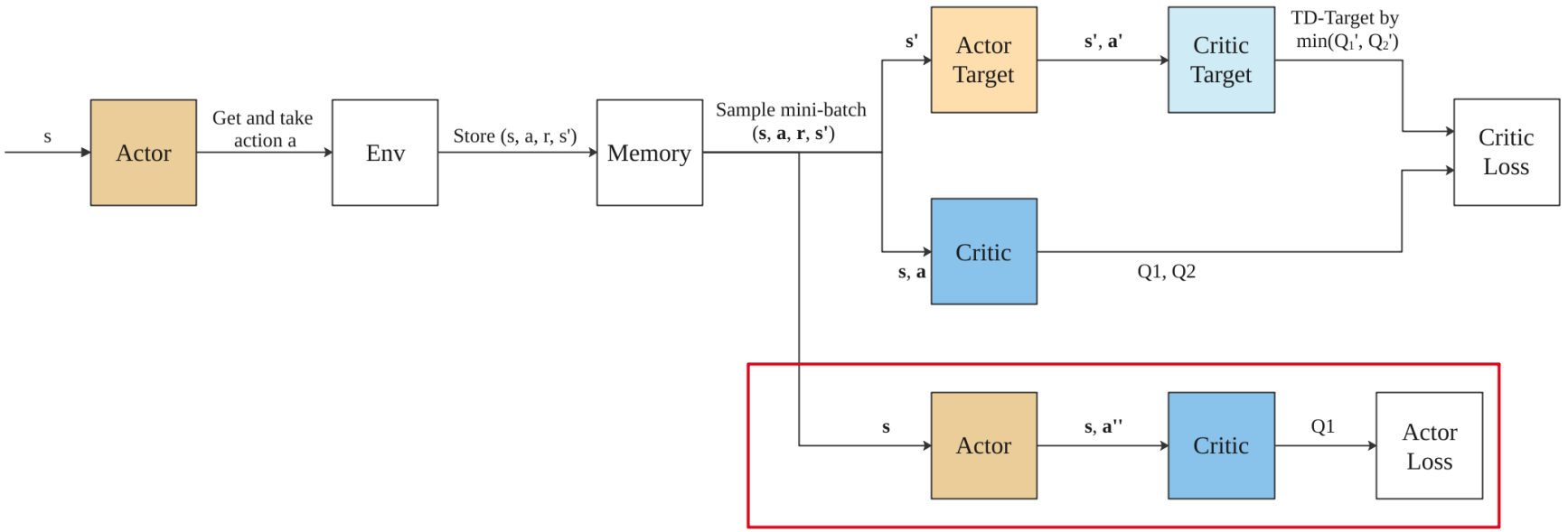








TD3



Every 2. Step (+ Update Target Networks)