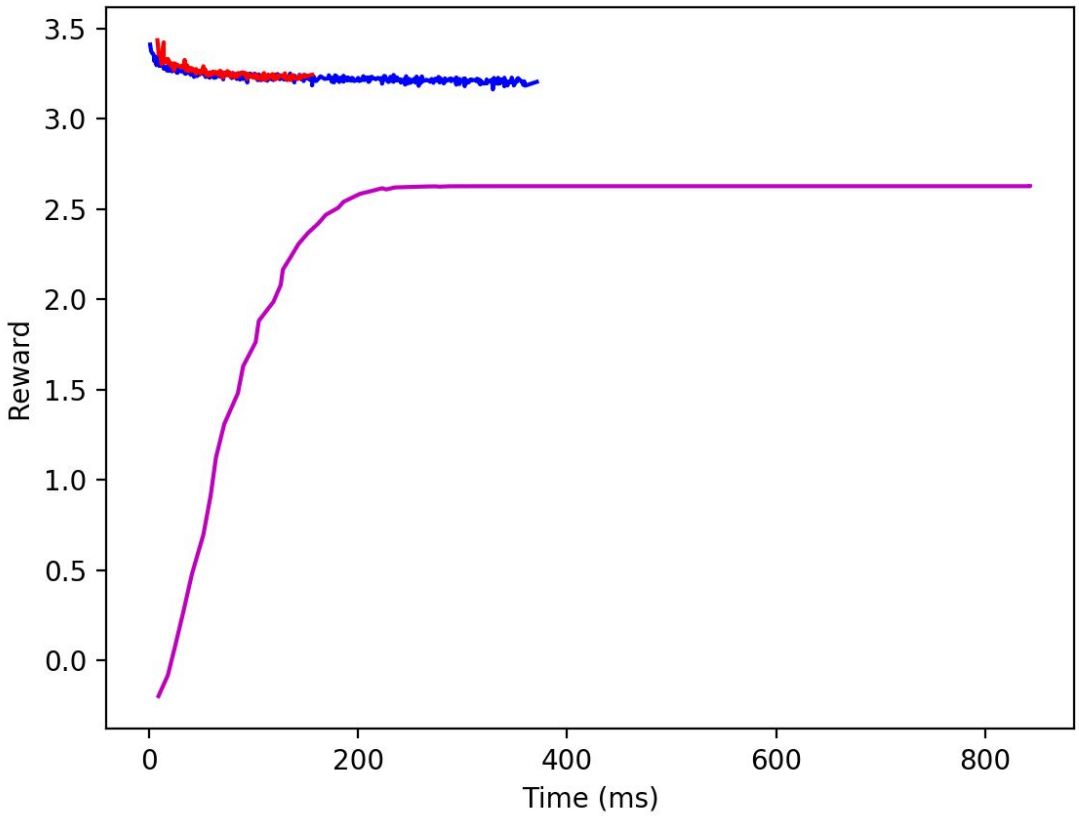
Mikhail Skobov
CSEP573
HW 3



Purple: VI
Blue: RTDP
Red: RTDP-reverse

As expected, VI takes much longer to converge than RTDP. The performance of RTDP appeared to be similar to that of RTDP-reverse. An issue encountered with RTDP-reverse was that it needed to be reset based on the number of iterations (if a local maxima was found - it would get stuck in a loop).

The heuristic chosen was as follows:

- If the state had a reward (positive or negative), the reward was returned
    - This is admissible because the reward is the maximum value for a state
- Otherwise, h(s) = pow(discount, manhattan_distance(s, goal)) * goal_reward
    - This is admissible because if P(s' | s) = 1 then the maximum value at that location would be V(s) = discount * P(s' | s) * V(s') which is maximized by goal

reward, thus it is always an overestimate.