PM 1 Part 2: Experimentation

Zehua Li
NetID: zehuali2
September 26th, 2014

Problem 1. In WorldWithoutThief :

(a) **Observation**: The rewards of episodes quickly converged to 0, and during policy simulation the bot tend to stay at the base (hitting the wall) or pace around the tiles on the left side of the slippery tiles.
Explanation: Since ϵ is 0, there will not be any random explorations(tie breaking is the only random element). In addition, once the bots get penalized for dropping a package from trying to pass the slippery tiles, the Q-value(utility) of the action is decreased. Since there are no random explorations, once the Q-value drops below that of other available actions in the robot's current state that lead to other tiles, the bot will not attempt it again. Since the customers are totally separated from the company by slippery tiles, Over the course of a significant number of episodes, the accumulation of this phenomenon will make it likely for the bot to stick with 0 reward, and highly unlikely for the bot have a policy instructing it to travel across the slippery tiles and to reach the customers.

(b) **Observation**: The rewards of episodes gradually raised from being generally negative to being generally positive, with a noticeable oscillation. Thus the bot would perform well majority of the time, but can still be as inactive as when ϵ was 0, or even worse at times.
Explanation: Since ϵ is now 0.1, 10% of the time the bot will choose a random action rather than be greedy with the current best action. This means that it will attempt to cross over slippery tiles even when the action has a relatively lower Q-value(utility), resulting in a higher probability for the robot to arrive at the customers with packages, gain a positive reward, and increase the Q-values for certain actions. Over a significant number of episodes, this will more likely to result in the bot having a policy that will allow it to gain positive rewards. However, the randomness also increases the possibility of the policy containing low utility actions, and result in the bot getting negative rewards.

(c) **Observation**: The rewards of episodes varies among negative values and rare, low positive values. The bot tends to pace around without aiming for the customer and lose the packages.
Explanation: With a higher ϵ, 0.5, 50% of the time the bot will choose a random action rather than be greedy with the current best action. This means that the policy will contain a significant number of actions that does not have a relatively high utility, making it less likely for the bot to gain positive reward or maintain it's current reward, but more likely to gain penalties.

Problem 2. In WorldWithThief :

(a) **Observation**: The rewards of episodes varies among negative values and rare, low positive values. The bot tends to stay at the company (hitting the wall) and rarely even pace back and forth.
Explanation: Without knowing the existence (and thus movement and location) of the thief, the Q-values(utilities) of actions approaching the middle tiles will not be affected by whether the thief is actually there, and are rather treated just like the actions approaching slippery tiles. Thus, similar to the situation of 1.(a), after a significant number of episodes, the bot will tend to say idle rather than crossing over the middle tiles to deliver the packages. Combined with the fact there are tiles next to the company in this world, the utility of actions in the states around the company are more likely to be very low (negative), resulting in the bot being more idle than in the situation of 1.(a), even though there is a 0.05 ϵ.

(b) **Observation**: The rewards of episodes quickly raised from negative values to positive values and slightly varies around 280. The bot tends to successfully deliver the packages.
Explanation: Knowing the existence (and thus movement and location) of the thief, the number of new states is now five times the original, each depending on which row the thief is

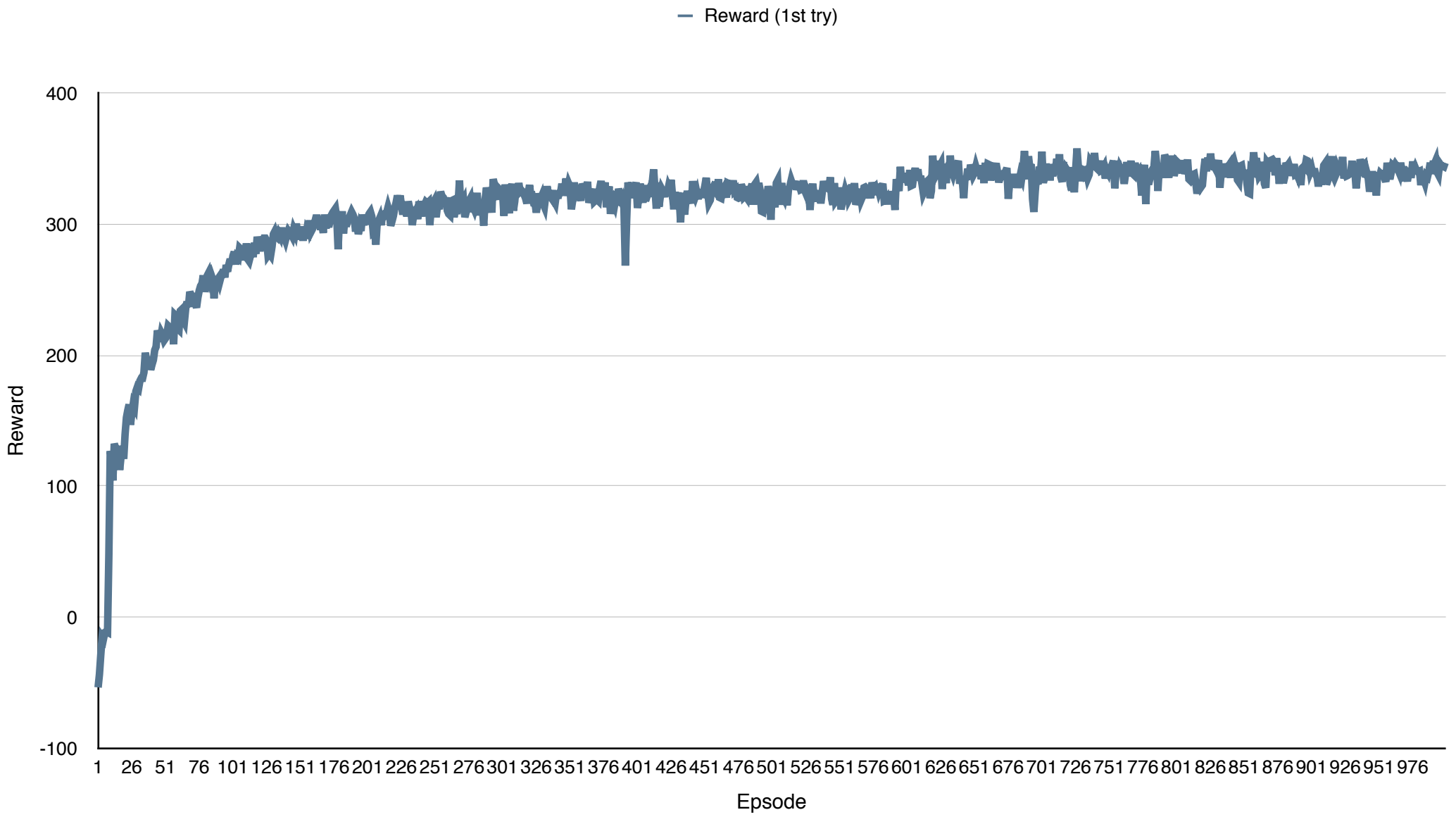
in. Thus the Q-values(utilities) of actions approaching the middle tiles (and other ones) are now distinguishable by whether the thief is actually there. Therefore, instead of treating the 5 tiles in the middle column slippery tiles all at ones, there is only one of them in each set of the states. Consequentially the utilities of actions in each of the states will be much higher, allowing the bot to move around and deliver the packages to the customers.

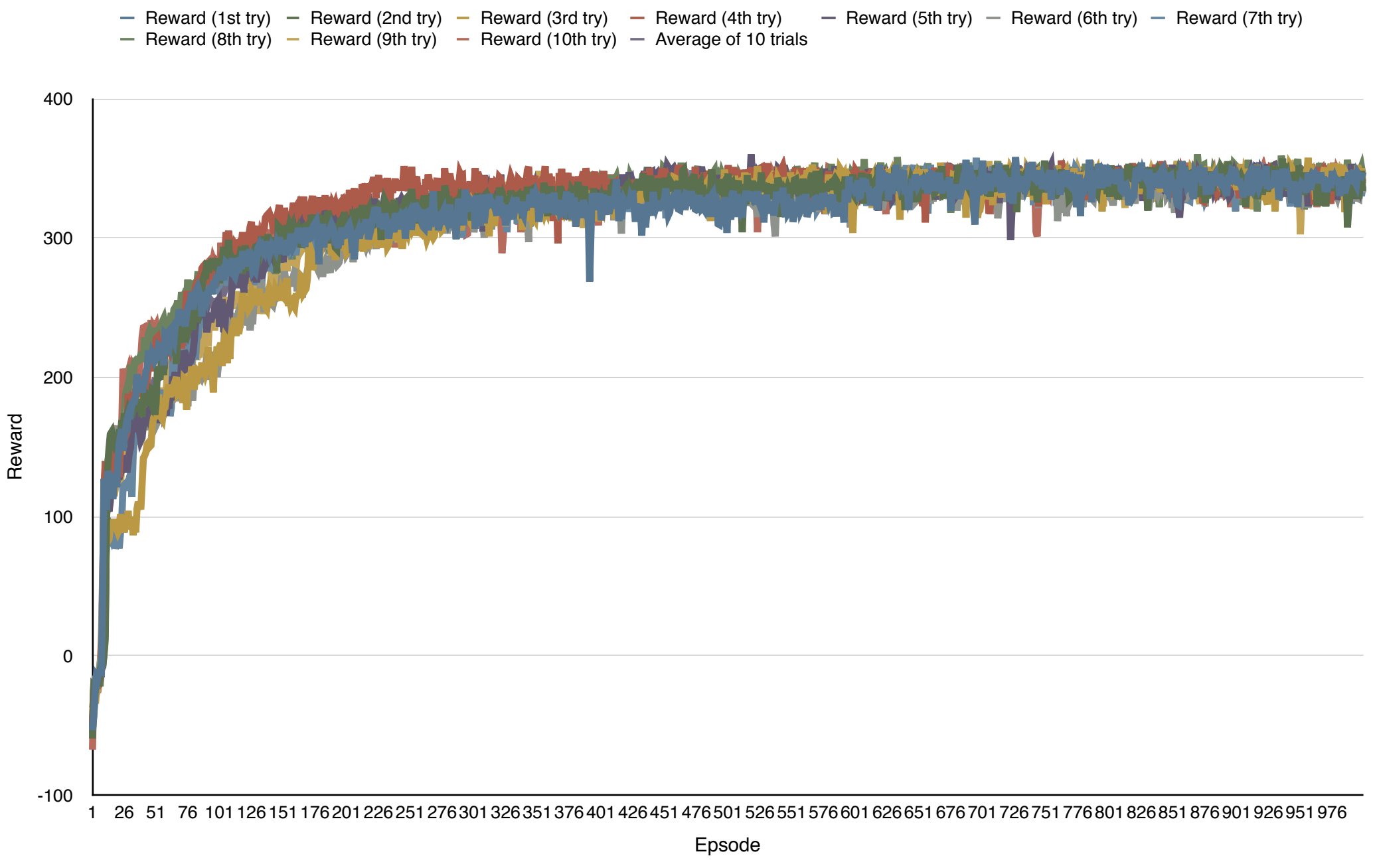
(c) After a first round of evenly distributed sampling of average rewards (rewards of the last 200 reward, where averages have stabilized) produced with pairs of ϵ and learning rate both within [0:1], the results showed a clear area of maximum when epsilon value is around 0.05 and when learning rate is around 0.1 to 0.2. After a few more rounds of detailed sampling, the best ϵ's and learning rate's resolution got a bit clearer, with ϵ around 0.0045 and learning rate around 0.15, where the average reward is slightly higher.

Problem 3. :

i) First Plot:
The increasing of total reward in the beginning indicates the improvement of the policy, as the agent updates the q-value every time, the policy is also advanced through the trials. In the latter portion, the reward fluctuates around a average value, because the learning agent has found a local maximum and the policy is close to optimal, except that the randomness is still effective, so the policy in each episode will be slightly different, and the resulting reward does not stay at a single value.

ii) Second Plot:
All 10 plots converge to certain average values, and since the parameters are the same, they tend to converge to values in a close range (having similar result) in the end, even though while earning, they may explore different paths and end up with slightly different local maximums.


Reward (1st try)
400
300
200
100
0
-100
Reward
1 26 51 76 101 126 151 176 201 226 251 276 301 326 351 376 401 426 451 476 501 526 551 576 601 626 651 676 701 726 751 776 801 826 851 876 901 926 951 976
Epsode


Reward (1st try)
Reward (2nd try)
Reward (3rd try)
Reward (4th try)
Reward (5th try)
Reward (6th try)
Reward (7th try)
Reward (8th try)
Reward (9th try)
Reward (10th try)
Average of 10 trials
400
300
200
100
0
-100
Reward
1 26 51 76 101 126 151 176 201 226 251 276 301 326 351 376 401 426 451 476 501 526 551 576 601 626 651 676 701 726 751 776 801 826 851 876 901 926 951 976
Epsode