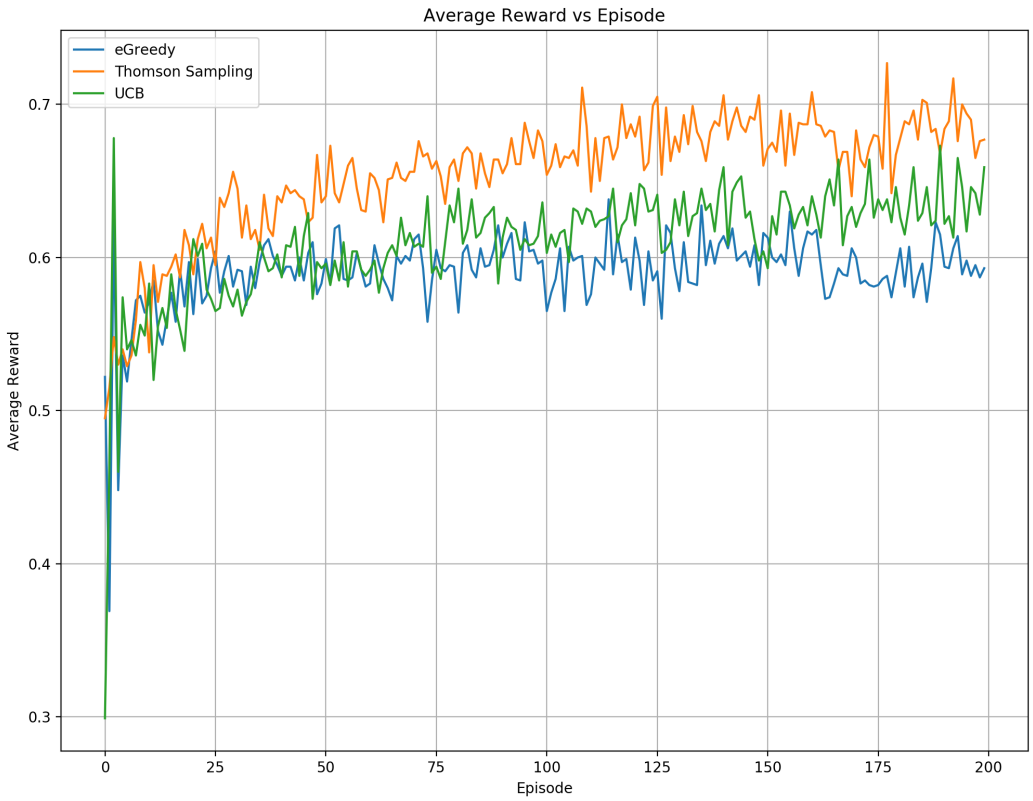
Assignment 2 Solutions

Part 1: 80 marks

Bandits: 40 marks (UCB, Thompson Sampling, epsilon greedy)

8 points per curve (24 total)

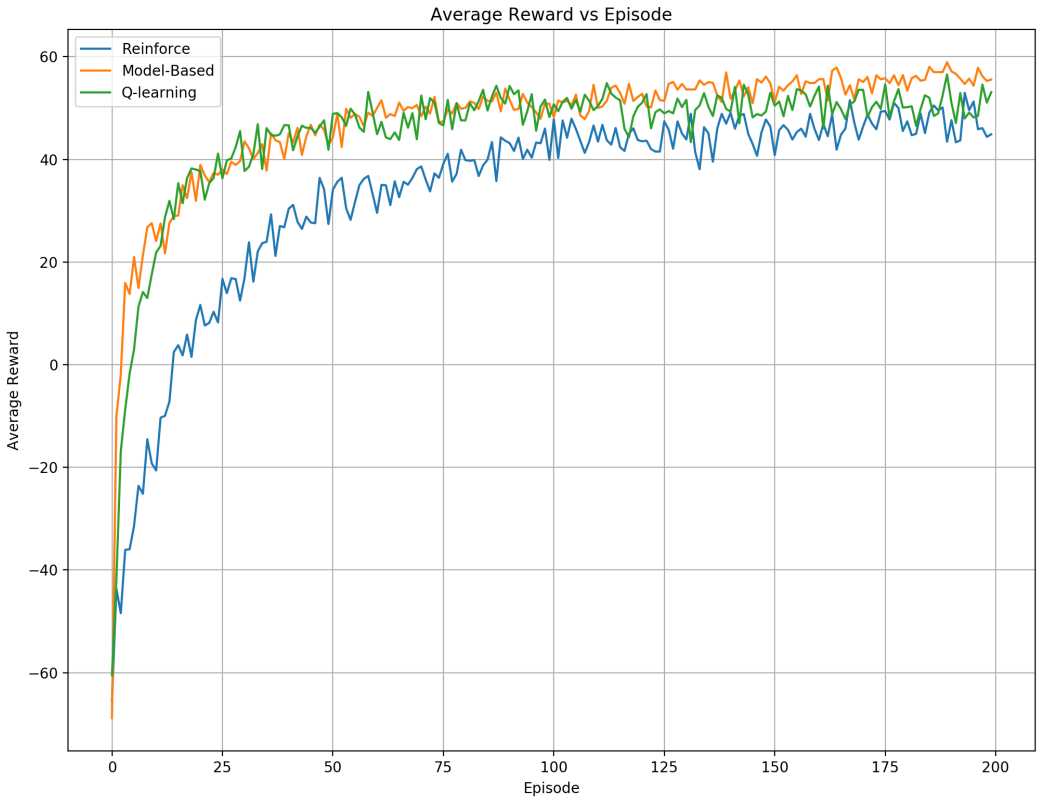


16 points for discussion: We notice that UCB does not perform as well as the other algorithms, but there is no

theory to explain this. All three algorithms exhibit a logarithmic cumulative regret in theory. UCB is a deterministic algorithm (epsilon-greedy and Thompson sampling are stochastic), which is a nice property, but this does not translate in better results.

RL algorithms: 40 marks (model based RL and REINFORCE)

10 points per curve (20 total)



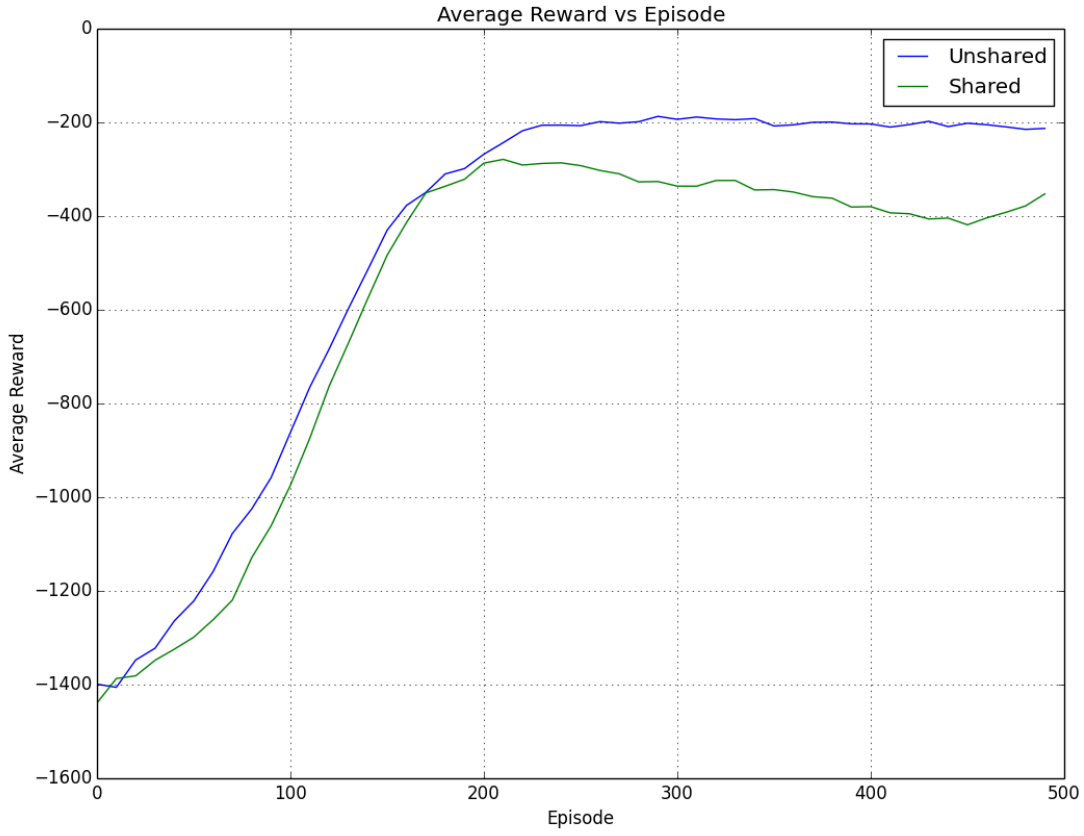
20 points discussion:

model-based is best, then Q-learning and then REINFORCE

Model-based learns faster since it fully uses all the information in each sample. Q-learning is not as data efficient since it only takes a small step in the direction of the gradient at each sample and the gradient is with respect to a single sample. REINFORCE is the worse because it only updates the policy after an entire trajectory and it may get stuck in a local optimum.

Part 2: 20 marks

10 marks for 2 curves



10 marks for explanation:

Sharing can improve the results when the actor and the critic are trying to extract the same features from the state and therefore the parameters can be optimized twice as fast. Sharing can mitigate overfitting.

Sharing can worsen the results when the actor and the critic are trying to extract different features from the state and therefore their updates cancel each other.

NB: The curves in the above graph vary with the random seed. Parameter sharing performs worse in some runs and better in other runs.