

Reinforcement Learning - Exercise 4

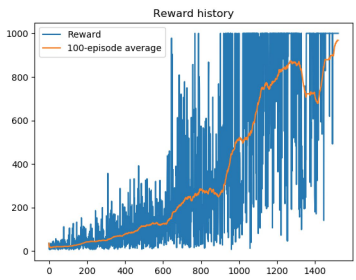
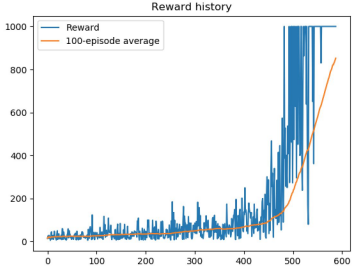
Muhammad Kamal Memon
600442

Task 1

Question 1

During training, in backpropagation the rewards affects the gradients. Hence we would like to keep the rewards in a convenient range so that they don't lead the network weights to extreme values. By preventing the network weights to have really high values and reducing their variance we are stabilizing our training. Hence basically this trick regularize the network.

Question 2

Method	Average episodes to train (5 trainings)	Reward Function
Constant Variance (0.5)	N/A (doesn't converge)	N/A
Decaying Variance	1522	
NN Variance	571	

As evident in the table above, the neural net approach of getting the variance for the gaussian from which we draw our sample is the fastest to converge and works best.

Constant Variance of 0.5 doesn't converge even after 4k episodes. However the Constant Variance approach timestamps on average remain above 900 after around 2k episodes, this

implies that it has explored more but it doesn't really enforce the best behavior but may still be able to keep the pole balanced.

Also we used the $k/(k + \text{episodes number})$ way for decaying the variance where k is a constant value of 1000. Other ways might result in faster convergence but will not explore well enough.

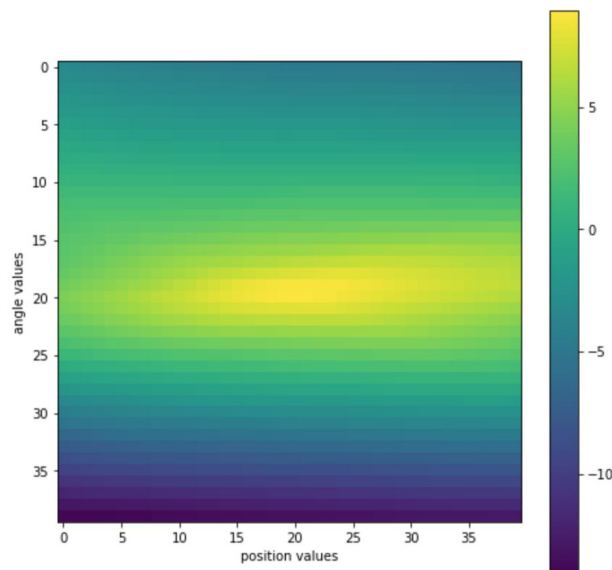
Task 2

Question 3

	Number of episodes to converge	
S#	Actor-critic	Policy Gradient
1.	372	503
2	262	587
3	357	600
4	567	775
5	377	673
Average	387	627.6

Clearly the actor-critic method works faster. The drawback of Value based methods such as Q-Learning tends to have poor convergence trends and Policy based methods tend to stuck to local maximas and have high variance. Actor critic methods takes the best of both these methods by employing an actor (policy gradient update) with a good critic (a value based method). Hence its main benefits are that convergence is guaranteed even for non-linear approximations and it reduces variance.

Task 3



The heatmap represents that irrespective of the position (since the velocity is zero) the highest rewards are in keeping the pole straight so it doesn't fall down. Hence more bright colors in the center of the map corresponding to higher reward values.

Question 5

As described in question 3 above actor-critic methods it is possible to reduce this variance as opposed to conventional policy gradient approach and that makes the learning faster.

Question 6

Policy gradient methods have some major advantages over value based methods such as they can solve large and continuous action spaces without the need of discretisation (as in Q-learning). Moreover, value-based method cannot solve an environment where the optimal policy is stochastic. However Q-learning is beneficial in tabular cases and when you want deterministic policy. Mainly it also depends on the state representation of the problem as in some cases we would want to learn the value function and in others the policy.

Also we can employ an approach which uses the strengths of both methods together, as in Actor-Critic method.

