Assignment 3

Ajay Subramanian

August 2019

Function Approximation

Question 1

No. It will not be a problem in Monte Carlo since we estimate the return by taking the average of multiple complete trajectories. This return is not dependenton the estimation parameters.

Question 2

No. In the original case, the next state depends only on this one. Therefore, it depends only on the features of this state and does not violate the Markov property

Question 3

All of these will affect generalization. Changing the size of aggregates will affect generalization since the updates vary across tilings.

Question 4

No. The global optimum achieved by the function approximator will be the value function that is closest (in Euclidean distance) to the true value function AND lying on the hyperplane defined by the features.

Question 5

(a) and (b) will be affected since gradient descent methods involve calculating how responsible a feature is for the obtained loss. This will definitely benefit from normalising the feature values since all of them will be brought to the same range and hence, easier to compare. LSPI on the other hand involves solving an argmax equation. This won't be affected since it's relative either way.

Question 6

Yes

Question 7

(c) is wrong. Since LSTDQ uses Q-learning for control, it uses an off-policy update that requires the current and next state's features. The next state features can be obtained from random samples, if we know the policy. In LSTD however, this whole process is on policy and sequences of states must follow the actions taken during policy evaluation.

Question 8

The value of the states can be thought of as increasing linearly from left to right. Hence only one feature is required.

DQN, Fitted Q and Policy Gradient Approaches

Question 1

(c) because a neural network extracts useful features during forward propagation

Question 2

We still need ϵ-greedy for exploration since experience replay picks random transitions that previously been encountered and not new unexplored ones.

Question 3

True. Value function based methods will learn the values of each state/state-action pair. A greedy agent will just take the path through states with highest values. In policy gradient however, we can define the policy as a probability distribution, parameterised by some parameters for say, the mean. There will be some variance as well which will help the agent learn a stochastic policy.

Question 4

(a),(b),(d). Long sample trajectories will only affect the time it takes to converge and not the capability of finding an optimal policy

Question 5

No

Question 6

Slower. Actor critic methods use value function estimates that will reduce the variance and hence lead to faster convergence.

Question 7

Don't know

Question 8

(b) convergence to a locally optimal policy

Hierarchical Reinforcement Learning

Question 1

Yes

Question 2

Recursively optimal

Question 3

No. The options need to only cover the states that are possible. These states can repeat elsewhere in the environment and hence, will only be a subset of the original state space.

Question 4

The options are independent of each other, in this case. Hence, conventional Q-learning should suffice.

Question 5

No. It will depend on the history since the initiation of the option.

Question 6

(b) Initially executing options will give the agent large negative rewards, forcing it to figure out the primitive states first. Hence the speed won't change.

Question 7

True. Intra-option updates are 'off-policy'. Hence we can update one option's estimates using rewards obtained with another one.

Question 8

Both. In options, limit the primitive actions. In HAM, limit the number of choice states.

Hierarchical RL: MAXQ

Question 1

In MaxQ we break down the over state value function into states under a particular subtask. Hence we optimize the policies within subtasks, making it seek recursive optimality

Question 2

Both pseudo rewards and core MDP rewards are available

Question 3

deterministic. We split the state space into non-terminal states and fixed terminal states

Question 4

(c)

Question 5

False. Reward is conditioned on core MDP's policy

Question 6

indirectly yes

POMDPs

Question 1

Yes. sequences of observations differentiate one state from another

Question 2

(b). We will have to take samples to compute some average parameter that differentiates states

Question 3

No. Example: 1001, 1001, 1001, 1001

Question 4

(d) The markovian assumption means that only the previous state, action are required

Question 5

(d) We can use history-based methods to solve this if we formulate is as an MDP

Question 6

(a) observation is unique for all states, in this case