AI Exam # 2

90pts

Bereket Tekeste

Problem-1

There are several mistakes in this pseudo-code for the SARSA(λ) algorithm:

Missed two errors. 30 pts.

1) The learning rate needs to be bound by 0-1. This means that for alpha to be correct the log must be log base 2 in order to stay within the bounds.

2) The initial action chosen must be chosen according to the policy. If it is chosen from a uniform distribution as is currently done in the algorithm, it will be attempting to do off-policy learning.

3) The update to the Q function does not use the eligibility traces in the update. So the "Update Q function" section can be replaced with the following:

Do for every state action pair:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha_t * n(s_t, a_t) * (y_t - Q(s_t, a_t))$$

$$n(s_t, a_t) = \lambda * n(s_t, a_t)$$

Problem-2

This variant of SARSA is known as expected SARSA. It is similar to the original SARSA algorithm, but the target is computed using a state-dependent policy instead of a fixed policy. This can be useful in situations where the optimal policy is not known in advance, or when the environment is non-stationary.
This variant of SARSA is known as SARSA with policy gradient. The advantage of this algorithm is that it can learn from a sequence of policies, instead of just a single policy. This can be helpful if the environment is non-stationary, or if there are multiple policies that could be beneficial in different situations.

Part A)
The sequence of policies (π_t) that should be chosen in order to make the Expected SARSA variant on-policy will depend on the specific details of the problem. However, one possible sequence of policies that could be used is π_{t+1} where each policy is ε-greedy with respect to the current Q values. This will force all actions to be chosen according to to the current policy.

Part B)
If your friend's algorithm is implemented correctly, then it should still converge to the optimal policy in the long run. However, the rate of convergence may be slower than with SARSA if the policies π_t deviate significantly from the optimal policy.

In this case the limiting Q value will be the Q value of the terminal state. This is because the Q value of the terminal state will never be changed, so Q values of all of the other states will depend on it.

ok. Not great