Reinforcement Learning

1. How is reinforcement learning different from supervised learning?

   Solution: In supervised learning the agent is given correct responses to inputs. In reinforcement learning an agents acts in a word in which the correct resopnses are unknown – the RL agents executes actions in order to receive reward signals that the agent uses to estimate the correct mapping from states to actions.

2. What is a policy?

   Solution: A mapping $\pi(s) : S \rightarrow A$ which specifies the action $a \in A$ an agent should take in state $s \in S$ to maximize goal attainment. A policy is a solution to an MDP.

3. A reinforcement learner solves the temporal credit assignment problem by learning from delayed reward. What does this mean?

   Solution:

4. What is a Markov decision process (MDP)? (List and describe its elements.)

   Solution:

   $$\langle S, A, T(s, a, s'), R(s, a, s')(or R(s))\rangle \tag{1}$$

   where

   - S is a set of states,
   - A is a set of actions,
   - $T(s, a, s')$ is a transition function which gives the probability that executing action a in state s will result in s', and
   - $R(s, a, s')$ which specifies the reward that the world provides to an agent for taking action a in state s and arriving in state s' or, equivalently, a reward for arriving in state s, $R(s)$.

   Some definitions of MDPs include an initialization function, $I(s)$, which specifies the probability the the agent will start in some state $s \in S$, others specify a particular state from S as the start state.

5. What is the difference between solving an MDP using value or policy iteration and learning a policy for an MDP using Q-learning?

Solution: Soving an MDP requires knowledge of the transition function, $T(s, a, s')$, and the reward function $R(s, a, s')$ or $R(s)$. A reinforcement learning algorithm such as Q-learning does not assum prior knowlege of T or R (they are learned implicitly).

6. What is the exploration-exploitation tradeoff?

Solution: How should an agent balance exploration of unseen states which might provide high (or low) reward with explotation of knowledge of states that are known to provide high reward.

Exploration means visiting some state of the world you have not yet visited. Eating at a restaurant or ordering a dish you have not tried to see if you like it is an example of exploration. Exploitation means using knowledge already learned. Eating a dish you know you like at a restaurant you know you like is an example of exploitation. The exploration versus exploitation question is one of risk versus reward: exploration risks wasting effort on a possibly bad outcome, exploitation trades that risk for a known reward but risks missing a discovery of something better.

7. Describe the ϵ-greedy exploration strategy?

Solution: In the case of reinforcement learning exploration means taking an action in a state that may take the agent to a state not yet visited. Exploitation means using the agent's current value estimates to choose an action. We call exploitation *greedy* action selection because it chooses an action that is likely to result in higher reward.

ϵ-greedy action selection means with probability $1 - \epsilon$, choose the action suggested by current value estimates (explotation). With probability ϵ, choose an random action in order to explore the space.

8. In ϵ-greedy exploration, how can the learning algorithm be prevented from "thrashing" after learning a good policy?

Solution: There is a saying in machine learning: the less you know the less you should trust your knowledge, the more you know the more you should trust your knowledge. A reinforcement learning agent should favor exploration early in its learning process and exploitation after its model of the world stabilizes.