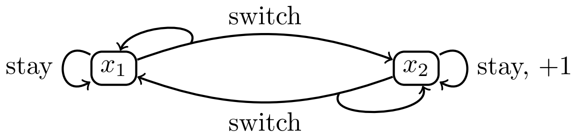
Question 1. (10 points)

Consider the following MDP. Assume that reward is in the form $r(x, y)$, i.e., $r : X \times Y \mapsto \mathbb{R}$. Set $\gamma = \frac{1}{2}$.



Suppose that you have seen the following sequence of states, actions, and rewards:

$$x_1, \text{switch}, x_2, \text{stay}, +1, x_2, \text{stay}, +1, x_2, \text{switch}, x_1, \text{stay}, x_1, \text{switch}, x_1, \text{switch}, x_1, \text{stay}, x_1, \text{switch}, x_2, \text{stay}, +1, x_2$$

1. (4 points) What is $\widehat{U}^{\pi}(x_i)$ calculated by the Direct Utility Estimation algorithm?
2. (2 points) What is transition model P estimated by the Adaptive Dynamic Programming algorithm?
3. (2 points) In the ADP estimates, some of the rare events might have zero probability, even though they are possible. Provide a solution in which the rare events that the algorithm misses during learning have a non-zero probability.
4. (2 points) What are state values estimated by a Temporal Difference learning agent after two steps? Assume that $\alpha = 0.1$ and all values are initialized to zero.

[adapted from Richard Sutton's 609 course, see `http://www.incompleteideas.net/book/the-book-2nd.html`]

Question 2. (3 points)

Decide whether the following statement is true or false: *If a policy π is greedy with respect to its own value function U^{π}, then this policy is an optimal policy.* Explain your decision.

[adapted from Richard Sutton's 609 course, see `http://www.incompleteideas.net/book/the-book-2nd.html`]

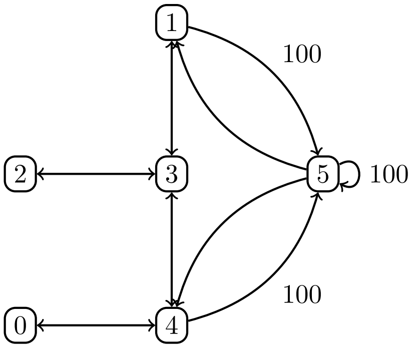
Question 3. (5 points)

Do the following exploration/exploitation schemes fulfill the 'infinite exploration' and 'greedy in limit' conditions? Which lead to the convergence of Q-values in Q-learning and which lead to the convergence of Q-values in SARSA. Does anything change if we are interested in the convergence of policy? $n_{x,y}$ denotes the number of times when action y was taken in state x. n_y is defined similarly.

1. a random policy
2. $$\pi(x) = \begin{cases} y, & \text{if } n_{x,y} \leq 100, \\ \arg\max_y Q(x, y), & \text{otherwise.} \end{cases}$$
3. ε-greedy policy with $\varepsilon = \frac{1}{n_x^2}$
4. ε-greedy policy with $\varepsilon = \frac{1{,}000}{999 + n_x}$
5. ε-greedy policy with $\varepsilon = \frac{1}{\sqrt{n_x}}$

Question 4. (5 points)

Consider the following MDP with $\gamma = 0.8$, $r(5) = 100$, $r(\cdot) = 0$.



The initial matrix of Q-values is

$$\widehat{Q}(x,y) = \begin{bmatrix} - & - & - & - & 0 & - \\ - & - & - & 0 & - & 0 \\ - & - & - & 0 & - & - \\ - & 0 & 0 & - & 0 & - \\ 0 & - & - & 0 & - & 0 \\ - & 0 & - & - & 0 & 0 \end{bmatrix}.$$

Consider path $1 - 5 - 1 - 3$ and constant learning rate $\alpha = 0.1$. Show changes in Q values after the agent-environment interaction for the Q-learning algorithm.

[adapted from Richard Sutton's 609 course, see `http://www.incompleteideas.net/book/the-book-2nd.html`]

Question 5. (10 points)

Consider an active reinforcement learning algorithm implemented by SARSA or Q-learning.

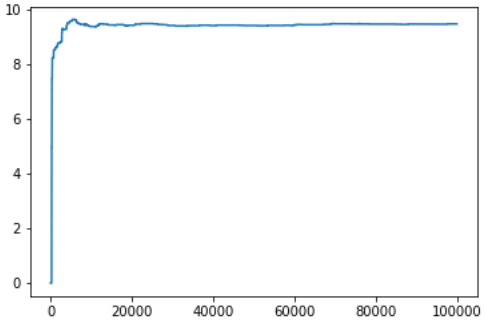
1. (2 points) Unlike the temporal difference learning, SARSA and Q-learning algorithms learn Q values instead of U. Why is U not enough?
2. (3 points) Explain why those algorithms need to balance exploration vs. exploitation. What those terms mean, and which of those is preferred early in the learning.
3. (2 points) SARSA and Q-learning are guaranteed to converge to an optimal policy if both:
   - convergence criteria for learning rate α known from TD-learning are met, and
   - convergence criteria on the explore-exploit policy are met.

   What are those criteria placed on the explore-exploit policy?
4. (1 point) Provide an example of an explore-exploit policy that guarantees policy convergence for SARSA and Q-learning.
5. (2 points) Will one of the algorithms learn Q-values even if one of the conditions is not met? If yes, which and why, if not, explain.

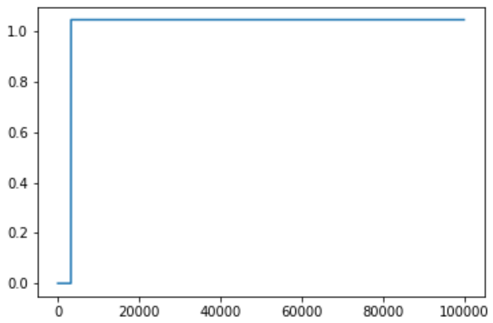
Question 6. (10 points)

Consider an active reinforcement learning algorithm. You are not told whether it is an instance of SARSA or Q-learning. The implementation met all convergence criteria. All plots shown below are related to the same state-action pair Q-value, i.e., $\widehat{Q}(x,y)$. The action y is **suboptimal** in state x. The used explore-exploit policy was the ε-greedy policy, i.e., with probability ε a random action is selected; otherwise, the agent behaves greedily.

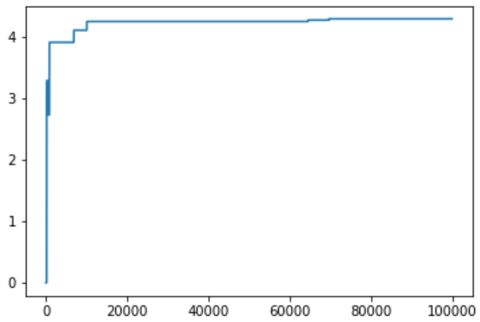
Now, consider four different situations of learning Q-values over 100 000 episodes.



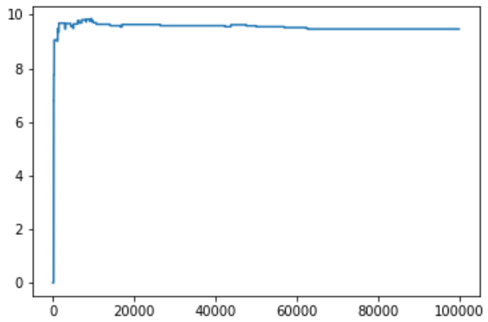
1.



2.

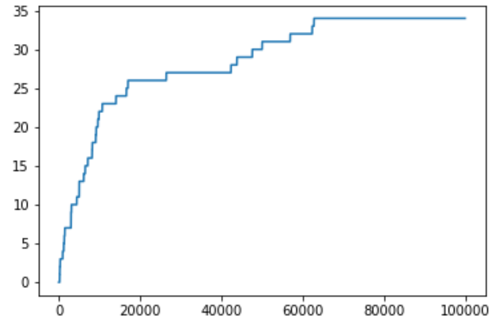


3.

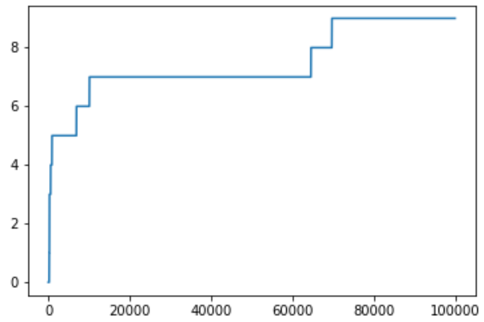


4.

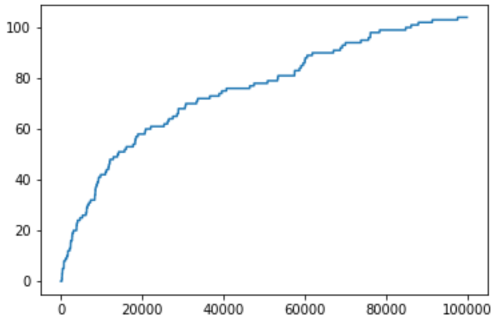
1. (4 points) Four plots below show how many times action y was selected by the agent in state x. For example, point $(1000, 6)$ means that the action y was selected 6 times in state x over the first 1000 episodes.



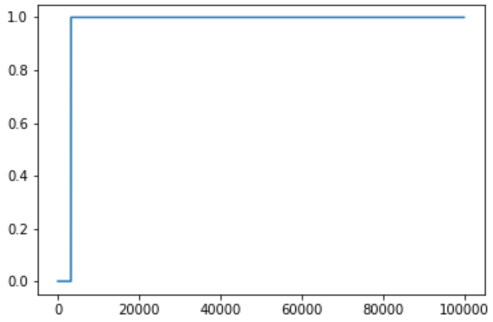
a.



b.



c.



d.

Match figures a-d to figures 1-4. Explain your decision.

2. (2 points) The ε was set as a function of the number of visits of state x. Relate the following four functions to the figures 1-4.

$$\text{i.}\quad \varepsilon(n_x) = \frac{8}{7+n_x} \qquad \text{ii.}\quad \varepsilon(n_x) = \frac{3}{2+n_x} \qquad \text{iii.}\quad \varepsilon(n_x) = \frac{100}{99+n_x} \qquad \text{iv.}\quad \varepsilon(n_x) = \frac{1000}{999+n_x}$$

Match those policies to figures 1-4. Explain your choice.

3. (1 point) Why should agents use different epsilon for different states.
4. (2 points) Decide whether the learning algorithm used was SARSA or Q-learning. Explain your decision.
5. (1 point) What is $Q(x, y)$? Explain your answer.