Practice Multiple-Choice Questions for Second Half of Reinforcement Learning 2020

Question 1

In policy gradient methods, what values are put into the input nodes of the policy network, when learning a stochastic policy?

A) The Q-values.
B) The parameters representing the action distribution (e.g. the mean and standard deviation).
C) The state.
D) The state and the action.
E) The state and the Q-values.
F) The action and the Q-values.

Question 2

In policy gradient methods, the policy network is updated on each iteration using gradient ascent. What does this gradient represent?

A) The rate of change of the network's parameters with respect to the objective function.
B) The rate of change of the objective function with respect to the network's parameters.
C) The rate of change of the policy's output action with respect to the objective function.
D) The rate of change of the objective function with respect to the policy's output action.
E) The rate of change of the policy's output action with respect to the network parameters.
F) The rate of change of the network parameters with respect to the policy's output action.

Question 3

A trajectory of length N, can be described as a sequence of states s_t and actions a_t, for $t = 1 \dots N$. Which of the following is the correct expression for the probability of this trajectory, when sampled from policy π_θ?

A) $\sum_{t=1}^{N} p(s_{t+1}|s_t, a_t) \times \pi_\theta(a_t|s_t)$
B) $\sum_{t=1}^{N} p(s_{t+1}|s_t, a_t)$
C) $\sum_{t=1}^{N} \pi_\theta(a_t|s_t)$
D) $\prod_{t=1}^{N} p(s_{t+1}|s_t, a_t) \times \pi_\theta(a_t|s_t)$
E) $\prod_{t=1}^{N} p(s_{t+1}|s_t, a_t)$
F) $\prod_{t=1}^{N} \pi_\theta(a_t|s_t)$

Question 4

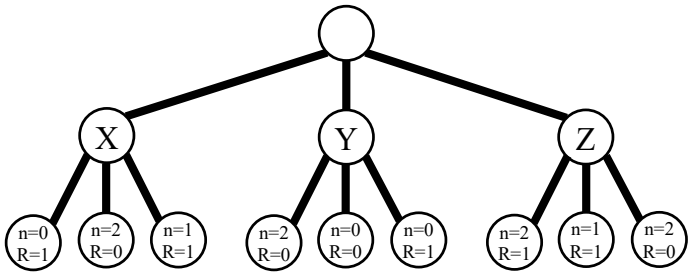

Figure for Question 4

Figure 1 shows the current structure of a tree used in a Monte Carlo Tree search algorithm, after a number of iterations. For the 9 leaf nodes, n represents how many times the node has been visited, and R represents the reward received by reaching that node.

On the next iteration, which of the nodes X, Y, Z will be visited?

A) Node X will be visited.
B) Node Y will be visited.
C) Node Z will be visited.
D) Nodes X and Y have equal probability of being visited, and Node Z will not be visited.
E) Nodes Y and Z have equal probability of being visited, and Node X will not be visited.
F) Nodes X and Z have equal probability of being visited, and Node Y will not be visited.
G) Nodes X, Y, and Z have equal probability of being visited.

Question 5

Monte-Carlo Tree Search (MCTS) and the Cross Entropy Method (CEM) are two ways to perform planning. Both proceed over a number of iterations, and both require computer memory to store data during the planning. Which of the following statements best describes how this memory requirement changes as the number of planning iterations increases?

A) The memory requirement is constant for both MCTS and CEM.
B) The memory requirement increases for both MCTS and CEM.
C) The memory requirement increases for MCTS, but is constant for CEM.
D) The memory requirement increases for CEM, but is constant for MCTS.

Question 6

A robot, whose state is defined as position (x, y), randomly explores a 2-dimensional room. When the robot moves *right*, the x-component of its state increases. The robot moves at a constant speed. Using this exploration data, a dynamics model is trained, $x_{t+1} = f(x_t)$, which represents the dynamics when the robot moves *right*.

The figure below shows 6 graphs, and each is a possible plot of the learned dynamics model. For each graph, the axes are both in the range 0 to 1. Which graph best represents the learned dynamics model?

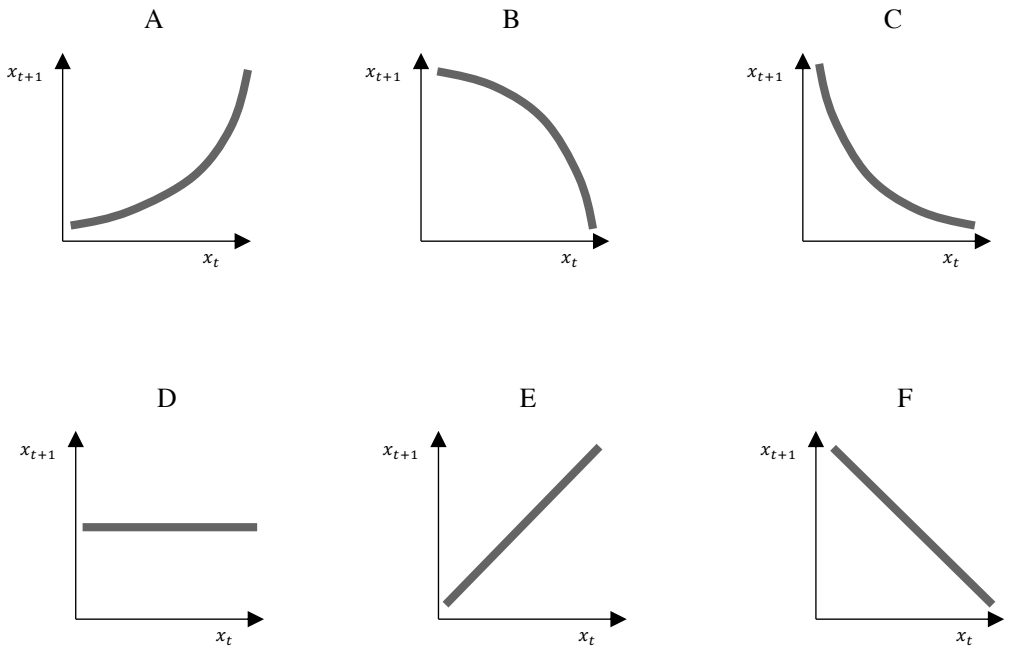

Figure for Question 6

Answers

Question 1, Answer = C
Question 2, Answer = B
Question 3, Answer = D
Question 4, Answer = B
Question 5, Answer = C
Question 6, Answer = E