Intelligent Agents
2022
Quiz 1
20. October 2022

- place your student ID card (carte de legitimation) on the desk in front of you.
- this is a closed-book examination (no documents allowed).
- when choosing the right answer, consider that the given explanation also has to be correct.
- mark the number of your copy on the top of each page to make sure we identify all pages of your exam.
- for questions with a single answer, the correct answer gives you 2 points. For questions with multiple correct answers, each correct answer gives you one point.
- If you give an incorrect answer, the question gives 0 points (even if there one or more of the answers is correct).

Copy No:

1. Which statements are **not** true in a POMDP:

   a) states and actions are known and finite
   b) state transitions are deterministic
   c) the state is observed with certainty
   d) rewards are known with certainty

   c

   a,d: are true in any MDP.
   b: can be true in a POMDP.

2. Why is the discount factor important?

   a) because agents should not be influenced too much by future rewards.
   b) because it makes the sum of discounted rewards over an infinite time horizon computable by a recurrence.
   c) as a parameter to tune agent policies.
   d) to model the uncertainty of future rewards.

   b

   a: the opposite is true - the challenge is to take into account future rewards while still maintaining convergence.
   c: we would like to avoid such hyperparameters.
   d: uncertainty could be better modelled by probabilities.

3. How do we recognize that value iteration has converged?

   Between two successive iterations over all states:

   a) the value function does not change anymore
   b) changes in the value function are bounded by a certain value
   c) the average change in the value function is bounded by a certain value
   d) the optimal policy for this value function does not change.

   b

   a: the value function may only gradually converge to a limit, but never reach it.
   c: the maximum distance from the true value function is determined by the maximum change.
   d: during convergence, there can be multiple steps where the policy does not change, but the value function does.

4. What is the definiton of cumulative regret?

   a) Average difference between the reward of the best possible action and the reward of the action actually taken.

b) For a sequence of actions, difference between the sum of rewards of the best possible policy and the actual policy.

c) For a sequence of actions, difference between the sum of rewards of the best possible fixed policy and the action actually taken.

d) Average difference between the reward of the best possible policy and the reward of the actual policy.

c

a: cumulative refers to a sequence of actions.
b: comparison is only with a fixed action for each state, more complex policies are not allowed. The word "fixed" is important. Also, comparing to the action actually taken is more correct than the actual policy, since the policy could have been updated since the time of the action.
d: cumulative refers to the sum of rewards, not an average.

5. What are the assumptions underlying confidence bounds?

a) rewards are distributed according to a Gaussian and bound holds with probability $(1 - \delta)$

b) samples are statistically independent and bound holds with probablity $(1 - \delta)$.

c) samples are statistically independent and rewards are distributed according to a Gaussian.

d) optimism under uncertainty: reward of an action is equal to the upper confidence bound.

b

a: samples need to be independent.
a,c: the assumption of a Gaussian distribution is not needed.
d: this is an assumption for exploration using the confidence bounds, not the bounds themselves.

6. Which of these learning algorithms require full observability ? (answer all that apply)?

a) Q-learning

b) regret matching

c) multiplicative weight updates

d) exponential weight updates

b,c

a: Q-learning only uses actual observations.
d: exponential weight updates are a modification of multiplicative updates to eliminate the need for counterfactual data.

7. When should we use a deliberative rather than a reactive agent (answer all that apply)?

a) when rewards and state transitions are very uncertain.

b) when only very few of the possible states are ever visited in the lifetime of the agent.

c) when there are strict real-time constraints.

d) when the rewards (goals) are changing frequently.

b,d

a: this situation is very difficult for a deliberative agent, as it has to weigh all the different uncertain transitions and rewards.
c: real-time constraints are much easier to satisfy with a precomputed policy in a reactive agent.

8. When using A* to search for the fastest plan for moving 5 boxes from the same starting point to different destinations in a graph, which of the following heuristics is admissible? Assume that the agent can carry any number of packages and moves at a fixed speed of 1m/s.

   a) The sum of the lengths of the shortest paths from the starting point to the destinations, for the packages not yet delivered.

   b) The maximum of the lengths of the shortest paths from the starting point to the destinations, for the packages not yet delivered.

   c) The minimum of the lengths of the shortest paths from the starting point to the destinations of the packages.

   d) The maximum of the lengths of the shortest paths from the position of the agent to the destination of a package that has not yet been delivered.

   d

   a: this would overestimate the cost if there are several packages with the same starting points and destinations that the agent can carry at the same time.
   b,c: this would overestimate the remaining cost when the agent is close to delivering the last package.

9. How are the actions chosen in counterfactual regret minimization?

   a) play each action a with probability proportional to the average of observed differences in reward for a and the actually played action a'.

   b) play the action a that maximizes the average of observed differences in reward for a and the actually played action a'

   c) play the action a that minmizes the average of observed differences in reward for a and the actually played action a'

   d) play the action a that in the past was most often the optimal one.

   a

   b,c,d: are all deterministic and can be exploited by an adversary.
   c: would not play the optimal but the least promising action.