Stanford University

AA228/CS238: Decision Making Under Uncertainty
Fall 2021
Prof. Mykel J. Kochenderfer • Online • email: *mykel@stanford.edu*

QUIZ 3 **Due date: November 19, 2021 (5pm Pacific)**

Quizzes will be taken on Gradescope. You may consult any material (e.g., books, calculators, computer programs, and online resources), but you may not consult other people inside or outside of the class. The quiz is designed to be completed in 90 minutes, but we will grant you 120 minutes total to complete and submit your quiz (including uploading any images, handling any logistical issues, etc.) The timing on Gradescope is a hard cutoff. You can start at 5pm PDT on Thursday. To accommodate those in other timezones and complex working situations, the quizzes will be open until 5pm PDT on Friday. Ed will not allow any public posts during that time. **Out of fairness to all students, only material submitted during the allowed time will be graded.**

Question 1. We have an action space $\mathcal{A} = \{1, 2, 3\}$. Suppose that at state s we have an action value function that assigns $Q(s, 1) = 0.6$, $Q(s, 2) = 0.1$, and $Q(s, 3) = 0.1$. For what precision parameter setting would softmax exploration select the first action with probability 0.5?

Solution: The softmax exploration strategy selects action i with a probability proportional to $e^{\lambda Q(s,i)}$, where λ is the precision parameter. Thus, we can write

$$\begin{aligned} P(i = 1) &= \eta\, e^{0.6\lambda} \\ P(i = 2) &= \eta\, e^{0.1\lambda} \\ P(i = 3) &= \eta\, e^{0.1\lambda}. \end{aligned}$$

We can determine the normalizing constant by solving

$$\begin{aligned} &\sum_i P(i) = 1 \\ \iff\ &\eta(e^{0.6\lambda} + e^{0.1\lambda} + e^{0.1\lambda}) = 1 \\ \iff\ &\eta = \frac{1}{e^{0.6\lambda} + 2\, e^{0.1\lambda}} \end{aligned}$$

and finally express λ

$$\begin{aligned} P(i = 1) &= 0.5 \\ \iff \frac{e^{0.6\lambda}}{e^{0.6\lambda} + 2\, e^{0.1\lambda}} &= \frac{1}{2} \\ \iff 1 + 2\, e^{-0.5\lambda} &= 2 \\ \iff e^{-0.5\lambda} &= \frac{1}{2} \\ \iff -0.5\lambda &= -\ln(2) \\ \iff \lambda &= 2\ln(2). \end{aligned}$$

We numerically find $\lambda \approx 1.37$.

Question 2. Suppose we want to apply Bayesian reinforcement learning to a discrete problem where the rewards we receive while interacting in the environment are either 1 or 0 with some probability that depends on our state and action taken. In other words, rewards are similar to binary multiarmed bandits, but we have multiple states. We do not know the probabilities we will receive a reward of 1 for the various states and actions, but we start with uniform priors. We know the transition model $T(s' \mid s, a)$. How would we apply posterior sampling to determine what action we should select from our current state? Please specify what kind of distribution we would use to represent the posterior.

Solution: We can model the probability of obtaining rewards at each state with Beta distributions. Then, we apply posterior sampling by sampling probabilities from the Beta distributions and solving the corresponding MDP.

Question 3. We want to apply reinforcement learning to a problem where both the state and action spaces are in $\mathbb{R}^2$. Suppose we use a parametric approximation of the action value function $Q_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) = \theta_1 s_1 + \theta_2 s_2 + \theta_3 a_1 + \theta_4 a_2$. We want to update our original estimate of $\boldsymbol{\theta} = [0, 0, 0, 0]$ after taking action $a = [1, 1]$ from state $s = [0, 0]$ and receiving reward $r = -1$ using gradient descent on a squared error loss function. Assume a step factor of $\alpha = 1$. Write down our updated $\boldsymbol{\theta}$.

Solution: The update rule for the squared error loss function is given by

$$\theta \leftarrow \theta + (r + \gamma \max_{a'} Q_\theta(s', a') - Q_\theta(s, a))\nabla Q_\theta(s, a)$$

As $Q_\theta(s, a)$ is 0 for all states and action pairs, this reduces to

$$\theta \leftarrow \theta + r\nabla Q_\theta(s, a)$$

Taking the gradient $\nabla Q_\theta(s, a) = [0, 0, 1, 1]$, we obtain the updated value

$$\theta' = [0, 0, -1, -1]$$

Question 4. We want to estimate the probability of whether we have a disease over time. Once we have the disease, we continue to have the disease indefinitely. If we do not currently have the disease, we will develop the disease with probability 0.1. We test whether we have the disease at each time step. With probability 0.1 the test will say that we do not have the disease when we actually have it, and with probability 0.2 it will say that we do have the disease when we do not. Assume that we start in a belief state that assigns equal probability to having and not having the disease. At the next time step, after observing a test that says that we have the disease, what is our updated belief state?

Solution:

We recall the expression of the discrete state filter,

$$b'(s') \propto O(o \mid s', a) \sum_s T(s' \mid s, a) b(s)$$

and compute terms for each of the possible states:

$$\begin{aligned} b'(s^0) &\propto O(o^1 \mid s^0) \sum_s T(s^0 \mid s) b(s) \\ &\propto 0.2\,(0.9 \times 0.5 + 0 \times 0.5) \\ &\propto 0.09, \end{aligned}$$

and

$$\begin{aligned} b'(s^1) &\propto O(o^1 \mid s^1) \sum_s T(s^1 \mid s) b(s) \\ &\propto 0.9\,(0.1 \times 0.5 + 1 \times 0.5) \\ &\propto 0.495. \end{aligned}$$

Finally, we normalize the belief vector

$$b'(s') = \eta \begin{bmatrix} 0.09 \\ 0.495 \end{bmatrix}$$

where

$$\eta = \frac{1}{0.495 + 0.09} \approx 1.71$$

and find

$$b'(s') = \begin{bmatrix} 0.15 \\ 0.85 \end{bmatrix}.$$

Question 5. Suppose we are building a system that will help an aircraft recover from a stall. The action space consists of two actions: do nothing or pitch the nose down. The state corresponds to the angle of attack of the aircraft, which we discretize into into 40 bins, equally spaced from −20 to 20 degrees. There are two angle of attack sensors, each providing independent measurements of the angle of attack, each discretized into 40 bins, resulting in 1600 possible observations. We solve for the optimal five-step policy to assist in stall recovery. What is the dimensionality of our resulting alpha-vectors? (No knowledge of aircraft dynamics, stalls, and angle of attack is required to answer this question.)

Solution: The cardinality of the state space is 40, resulting in a belief vector of dimension 40. Thus, the alpha vector will be dimension 40.

Question 6. How many alpha vectors are needed to represent the fast-informed bound for the problem introduced in the previous question?

Solution: The Fast Informed bound computes one alpha vector for each action. Thus, 2 alpha vectors are needed.

Question 7. We want to take a branch and bound approach to solving the problem sketched in Question 5. How might we obtain a lower bound for this problem that would encourage pruning?

Solution: To encourage pruning, our solution must provide high lower bound utility values. Possible answers include:

- A best-action-worst-state lower bound
- A blind lower bound
- A lower bound resulting from point-based value iteration