Stanford University

AA228/CS238: Decision Making Under Uncertainty
Fall 2023
Anil Yildiz • Bishop Auditorium • email: *yildiz@stanford.edu*

PROBLEM SESSION 7: REINFORCEMENT LEARNING **November 8, 2023**

1 Introduction

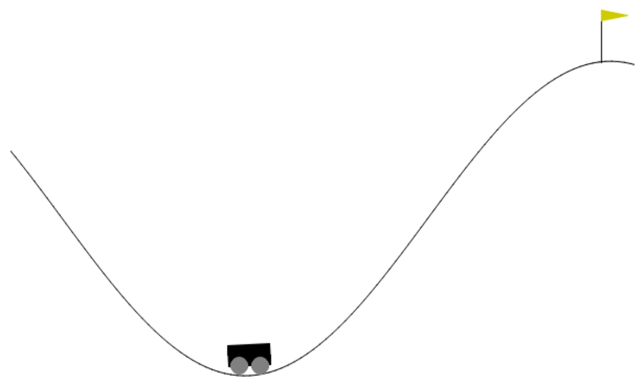

Figure 1: Mountain car toy problem. Source.

For this problem session, we will go over reinforcement learning (RL) methods. We will cover both model-based (Chapter 16) and model-free (Chapter 17) methods. The MDP that we will use as the running example is the mountain car toy problem that you are familiar from Project 2. However, in Project 2, you were given a CSV file that has already simulated the model (based on some exploration/exploitation strategy not known to you) and lists the resulting (s, a, r, s') tuples. In this problem session, we will also cover how to choose which state transitions to simulate. As in Project 2, for this toy problem, we have 50,000 possible states and 7 possible actions.

2 Model-Based RL Methods

In model-based methods, we are not only concerned with the value of each action, but also would like to model the transition $T(s' \mid s, a)$ and reward $R(s, a)$ functions. We might be interested in modeling T and R to further analyze the system afterwards, e.g. to benchmark the model's accuracy, to run sensitivity analysis on the model, visualize where the highest rewards are, etc. We categorize model-based RL methods into two: maximum likelihood methods, and Bayesian methods.

2.1 Maximum Likelihood Methods

Question 1. Approximating the model.

a) Given the data in Table 1, perform maximum likelihood learning to estimate the transition and reward functions. Assume zero initialization for all $N(s, a)$.

Table 1: Data for s, a, r, s'.

s	a	r	s'
20017	6	-100	20018
20018	7	-100	20020
20020	4	0	20018
20018	7	-225	20019
20019	3	-225	20017

b) Suppose we have obtained the results above for T and R functions. Our friend has simulated the same problem and has obtained the results in Table 2. But they have forgotten to record the rewards obtained! Can we still combine our result from part a) with theirs? If so, how?

Table 2: New data from our friend.

$T(s' = 20019 \mid s = 20017, a = 6) = 0.5$
$T(s' = 20018 \mid s = 20017, a = 6) = 0.5$
$T(s' = 20020 \mid s = 20018, a = 7) = 0.33$
$T(s' = 20021 \mid s = 20018, a = 7) = 0.67$
$N(s = 20017, a = 6) = 2$
$N(s = 20018, a = 7) = 3$

2.2 Update Schemes

Update schemes answer the following question: Given that we have approximated $T(s' \mid s, a)$ and $R(s, a)$, how do we compute the action values (and therefore the optimal policy)?

Question 2. Computing the optimal policy.

a) Explain the differences between full and randomized updates. When might you prefer one over the other?

2.3 Exploration

Now let us talk about the exploration strategies for model-based RL methods. I.e. how do we choose which (s, a, r, s') tuples to simulate in the first place? There are two main techniques that we utilize: ϵ-greedy, and RMAX exploration.

Question 3. Exploration strategies.

a) Explain how you would perform ϵ-greedy exploration.

b) Explain how you would perform RMAX exploration.

2.4 Bayesian Methods

Question 4. Approximating the model (Bayesian).

a) Apply Bayesian RL to the mountain car problem to approximate the transition dynamics. Use the data from Table 1.

b) In the Bayesian RL above, how many independent parameters would we need to represent the transition dynamics?

3 Model-Free RL Methods

In model-free methods, we are not concerned with modeling $T(s' \mid s, a)$ or $R(s, a)$. We aim to learn the action values (and therefore the optimal policy) directly. Avoiding explicit representations is attractive, especially when the problem is high dimensional. There are two main techniques to perform model-free RL: Q-learning and SARSA.

Question 5. Q-learining and SARSA.

a) Assume that we have been applying Q-learning and have reached the following values in Table 3.

Table 3: Latest results from training, abbreviated.

$Q(s = 5006, a = 3) = +20$
$Q(s = 5006, a = 4) = -30$
$Q(s = 5007, a = 4) = -75$
$Q(s = 5007, a = 5) = -50$

Our current state is $s = 5006$, and our exploration strategy is telling us to take action $a = 4$, and by doing so, we reach next state $s' = 5007$ and receive reward $r = -100$. Perform one step of Q-learning. Use $\gamma = 1.0$ and $\alpha = 0.4$.

b) At the next state $s' = 5007$, we know that our exploration policy will recommend the action $a' = 4$. Given the information above, apply one step of SARSA instead.

c) Are the results from part a) and b) the same? If not, how does the learning rate α mitigate for errors?