RBE 595 — Reinforcement Learning
Assignment #6
Model-Based Reinforcement Learning

Arjan Gupta

Problem 1

What is "planning" in the context of Reinforcement Learning?

Answer

In the context of Reinforcement Learning, planning is the process of using a model of the environment to improve the policy.

A model of the environment is a representation of the environment that is used to predict what next state and reward will be given a current state and action. Once we have a model, we can use it to simulate the environment and produce a simulated experience, in the form of an episode. Then we can use this simulated experience to improve the policy. This is the process of **planning**.

Problem 2

What is the difference between Dyna-Q and Dyna-Q+ algorithms?

Answer

The problem with Dyna-Q is that it does not balance exploration and exploitation well in the planning phase. This is because the planning phase is greedy. Only the learning phase is ϵ-greedy. The reason that we want to also explore in the planning phase is that the model may need to change if the environment changes over time.

Dyna-Q+ is like Dyna-Q, except that it adds an exploration 'bonus' to the planning phase. What this means is that we essentially provide a bonus reward in the planning phase for states that have not been visited in a long time. This encourages the agent to explore in the planning phase. Specifically, the bonus reward is given by:

$$R = r + \kappa\sqrt{\tau(s,a)}$$

Where r is the reward received, κ is a constant of our choice, and $\tau(s,a)$ is the number of time steps since the last visit to state s after taking action a. It is important to note that the bonus reward is only given in the planning phase, and not during regular interaction with the real environment.

Problem 3

Model-based RL methods suffer more bias than model-free methods. Is this statement correct? Why or why not?

Answer

This statement is correct. Model-based RL methods suffer more bias than model-free methods because the design of the model introduces bias. The model is a representation of the environment, and it is not possible to represent the environment perfectly. Therefore, the model will always introduce some bias.

Problem 4

Model-based RL methods are more sample efficient. Is this statement correct? Why or why not?

Answer

The n-step return of Sarsa (7.4) is as follows:

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}), \quad n \geq 1, \quad 0 \leq t < T - n$$

We can rewrite this as follows:

$$G_{t:t+n} = \sum_{i=1}^{n} \gamma^{i-1} R_{t+i} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) \qquad \text{(Equation 1, Problem 4)}$$

Now we take two cases of equation 7.6 — one where $t + n < T$, and one where $t + n \geq T$.

Case 1: $t + n < T$

In this case, we have $min(t + n, T) = t + n$. Therefore, equation 7.6 becomes:

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{t+n-1} \gamma^{k-t} [R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)]$$

Let us expand the summation in the above equation:

$$\begin{aligned}
G_{t:t+n} &= Q_{t-1}(S_t, A_t) + \gamma^0 [R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t)] \\
&+ \gamma^1 [R_{t+2} + \gamma Q_{t+1}(S_{t+2}, A_{t+2}) - Q_t(S_{t+1}, A_{t+1})] \\
&+ \gamma^2 [R_{t+3} + \gamma Q_{t+2}(S_{t+3}, A_{t+3}) - Q_{t+1}(S_{t+2}, A_{t+2})] \\
&+ \gamma^3 [R_{t+4} + \gamma Q_{t+3}(S_{t+4}, A_{t+4}) - Q_{t+2}(S_{t+3}, A_{t+3})] \\
&+ \ldots \\
&+ \gamma^{n-1} [R_{t+n} + \gamma Q_{t+n-1}(S_{t+n}, A_{t+n}) - Q_{t+n-2}(S_{t+n-1}, A_{t+n-1})] \\
&= Q_{t-1}(S_t, A_t) \\
&+ R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t) \\
&+ \gamma R_{t+2} + \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) - \gamma Q_t(S_{t+1}, A_{t+1}) \\
&+ \gamma^2 R_{t+3} + \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) - \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) \\
&+ \gamma^3 R_{t+4} + \gamma^4 Q_{t+3}(S_{t+4}, A_{t+4}) - \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) \\
&+ \ldots \\
&+ \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) - \gamma^{n-1} Q_{t+n-2}(S_{t+n-1}, A_{t+n-1})
\end{aligned}$$

However, we can see that the terms in the above equation cancel out. The cancellation pattern is, $Q_{t-1}(S_t, A_t)$ from the first line gets cancelled by the last term in the second line, $\gamma Q_t(S_{t+1}, A_{t+1})$ gets cancelled by the last term in the third line, and so on.

Therefore, we are left with the following:

Problem 4 continued on next page. . .

$$\begin{aligned}
G_{t:t+n} &= R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) \\
&= \sum_{i=1}^{n} \gamma^{i-1} R_{t+i} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})
\end{aligned}$$

Which is the same as the rewritten form of the n-step return of Sarsa (7.4), as shown in equation 1 of this Problem. Therefore, this case is proven.

Case 2: $t + n \geq T$

In this case, we have $min(t + n, T) = T$. Let us assume that $t + n$ overshoots T by x steps, where $x \geq 0$. So, $t + n = T + x$, or $T = t + n - x$. Therefore, equation 7.6 becomes:

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{t+n-x-1} \gamma^{k-t}[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)]$$

Let us expand the summation in the above equation:

$$\begin{aligned}
G_{t:t+n} = {} & Q_{t-1}(S_t, A_t) + \gamma^0[R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t)] \\
& + \gamma^1[R_{t+2} + \gamma Q_{t+1}(S_{t+2}, A_{t+2}) - Q_t(S_{t+1}, A_{t+1})] \\
& + \gamma^2[R_{t+3} + \gamma Q_{t+2}(S_{t+3}, A_{t+3}) - Q_{t+1}(S_{t+2}, A_{t+2})] \\
& + \gamma^3[R_{t+4} + \gamma Q_{t+3}(S_{t+4}, A_{t+4}) - Q_{t+2}(S_{t+3}, A_{t+3})] \\
& + \ldots \\
& + \gamma^{n-x-1}[R_{t+n-x} + \gamma Q_{t+n-x-1}(S_{t+n-x}, A_{t+n-x}) - Q_{t+n-x-2}(S_{t+n-x-1}, A_{t+n-x-1})] \\
& = Q_{t-1}(S_t, A_t) \\
& + R_{t+1} + \gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t) \\
& + \gamma R_{t+2} + \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) - \gamma Q_t(S_{t+1}, A_{t+1}) \\
& + \gamma^2 R_{t+3} + \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) - \gamma^2 Q_{t+1}(S_{t+2}, A_{t+2}) \\
& + \gamma^3 R_{t+4} + \gamma^4 Q_{t+3}(S_{t+4}, A_{t+4}) - \gamma^3 Q_{t+2}(S_{t+3}, A_{t+3}) \\
& + \ldots \\
& + \gamma^{n-x-1} R_{t+n-x} + \gamma^{n-x} Q_{t+n-x-1}(S_{t+n-x}, A_{t+n-x}) - \gamma^{n-x-1} Q_{t+n-x-2}(S_{t+n-x-1}, A_{t+n-x-1})
\end{aligned}$$

However, we can see that the terms in the above equation cancel out in the same way as in Case 1.

$$\begin{aligned}
G_{t:t+n} &= R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-x-1} R_{t+n-x} + \gamma^{n-x} Q_{t+n-x-1}(S_{t+n-x}, A_{t+n-x}) \\
&= \sum_{i=1}^{n-x} \gamma^{i-1} R_{t+i} + \gamma^{n-x} Q_{t+n-x-1}(S_{t+n-x}, A_{t+n-x})
\end{aligned}$$

And as we assumed, x is the number of steps by which $t + n$ overshoots T. Therefore, $n - x$ is the 'reduced horizon' of the n-step return of Sarsa (7.4). So, in general, as the horizon shrinks, we can keep taking $n - x = n$ steps, and thus the equation above can simply be written generally as:

Problem 4 continued on next page...

$$G_{t:t+n} = \sum_{i=1}^{n} \gamma^{i-1} R_{t+i} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

Which is the same as the rewritten form of the n-step return of Sarsa (7.4), as shown in equation 1 of this Problem. Therefore, this case is also proven.