National University of Singapore
School of Computing

Semester 1, AY2023-24

CS4246/CS5446 AI Planning and Decision Making

Tutorial Week 9: Real world RL

Guidelines

You may discuss the content of the questions with your classmates. But everyone should work on and be ready to present ALL the solutions.

Problem 1: Approximating TD Learning

[RN 3e 21.4] Write out the parameter update equations for temporal difference (TD) learning with

$$\hat{U}(x, y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{(x - x_g) + (y - y_g)}.$$

Solution:

$$\begin{aligned}
\theta_0 &\leftarrow \theta_0 + \alpha(R(s) + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s)), \\
\theta_1 &\leftarrow \theta_1 + \alpha(R(s) + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s))x, \\
\theta_2 &\leftarrow \theta_2 + \alpha(R(s) + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s))y, \\
\theta_3 &\leftarrow \theta_3 + \alpha(R(s) + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s))\sqrt{(x - x_g) + (y - y_g)}.
\end{aligned}$$

Problem 2: Approximating Q-Learning

Consider a system with a single state variable x that can take value 0 or 1 and actions a_1 and a_2. An agent can observe the value of the state variable as well as the reward in the observed state. Assume a discount factor $\gamma = 0.9$.

(a) Perform two steps of Q-learning with the observed transitions shown below in (i) and (ii) using a table representation of the Q-function. Use a learning rate of $\alpha = 0.5$ starting from a table with all entries initialized to 0. Show the Q-function after each step.

(i) First observed transition: initial value of $x = 0$, observed reward $r = 10$, action a_1, next state $x = 1$.

Solution:

Applying $Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a))$, we get $Q(0,a1) \leftarrow 0 + 0.5(10 + 0.9\max(0,0) - 0) = 5$

Q	$x=0$	$x=1$
a_1	5	0
a_2	0	0

(ii) Second observed transition: from $x = 1$, observed reward $r = -5$, action a_2, next state $x = 0$.

Solution:

$Q(1,a_2) \leftarrow 0 + 0.5(-5 + 0.9\max(5,0) - 0) = -0.25$

Q	$x=0$	$x=1$
a_1	5	0
a_2	0	-0.25

(b) Now perform Q-learning with function approximation using $Q(x,a_1) = \beta_1 x$ and $Q(x,a_2) = \beta_2 x$. Use a learning rate $\alpha = 0.5$ starting from parameters $\beta_1 = 0$ and $\beta_2 = 0$. Show the parameter values after each step.

(i) First observed transition: initial value of $x = 1$, observed reward $r = 10$, action a_1, next state $x = 1$.

Solution:

Applying

$$\beta_i \leftarrow \beta_i + \alpha(R(x) + \gamma \max_{a'} Q(x',a') - Q(x,a))\frac{\partial Q(x,a)}{\partial \beta_i},$$

we get

$$\beta_1 \leftarrow 0 + 0.5(10 + 0.9\max(0,0) - 0)1 = 5$$
$$\beta_2 \leftarrow 0 + 0.5(10 + 0.9\max(0,0) - 0)0 = 0$$

(ii) Second observed transition: from $x = 1$, observed reward $r = -5$, action a_2, next state $x = 0$.

Solution:

$$\beta_1 \leftarrow 5 + 0.5(-5 + 0.9\max(0,0) - 0)0 = 5$$
$$\beta_2 \leftarrow 0 + 0.5(-5 + 0.9\max(0,0) - 0)1 = -2.5$$

(c) After enough data is observed, which method would give better performance, the tabular method in (a) or the function approximation method in (b)? Why? Suggest how the poorer performing method can be improved.

Solution:
The tabular method will work better as it will converge to the optimal solution whereas the function approximation with $Q(x, a1) = \beta_1 x$ and $Q(x, a2) = \beta_2 x$ is unable to represent any function where the value for $x = 0$ is non-zero. One way to improve would be to use a better function approximator. In this case adding a bias to each function $Q(x, a_1) = \beta_1 x + \delta_1$ and $Q(x, a2) = \beta_2 x + \delta_2$ is sufficient as it will be able to represent any function that the table can represent.

Problem 3: Q-Learning with continuous state

Consider a system with a single continuous variable x and actions a_1 and a_2. An agent can observe the value of the state variable as well as the reward in the observed state. Assume $\gamma = 0.9$.

(a) Assume that function approximation is used with $Q(x, a_1) = w_{0,1} + w_{1,1}x + w_{2,1}x^2$ and $Q(x, a_2) = w_{0,2} + w_{1,2}x + w_{2,2}x^2$. Give the Q-learning update equations.

Solution:
If action a_i is taken, at state x and next state is x':

$$\begin{aligned} w_{0,i} &\leftarrow w_{0,i} + \alpha \left(R(x) + \gamma \max_{a'} Q(x', a') - Q(x, a_i) \right) \\ w_{1,i} &\leftarrow w_{1,i} + \alpha \left(R(x) + \gamma \max_{a'} Q(x', a') - Q(x, a_i) \right) x \\ w_{2,i} &\leftarrow w_{2,i} + \alpha \left(R(x) + \gamma \max_{a'} Q(x', a') - Q(x, a_i) \right) x^2 \end{aligned}$$

while $w_{k,j}$ is unchanged for $j \neq i$.

(b) Assume that $w_{i,j} = 1$ for all i, j. The following transition is observed: $x = 0.5$, observed reward $r = 10$, action a_1, next state $x' = 1$. What are the updated values of the parameters assuming $\alpha = 0.5$?

Solution:

$$\begin{aligned} w_{0,1} &\leftarrow 1 + 0.5\,(10 + 0.9 \max(3, 3) - 1.75) = 6.475 \\ w_{1,1} &\leftarrow 1 + 0.5\,(10 + 0.9 \max(3, 3) - 1.75)\,0.5 = 3.7375 \\ w_{2,1} &\leftarrow 1 + 0.5\,(10 + 0.9 \max(3, 3) - 1.75)\,0.25 = 2.36875 \end{aligned}$$

The other parameters are unchanged.