Assignment #3

E0503474 / Yu-Ting TSENG (with None Collaborator)

Oct 30, 2020

Problem 1: Q-Learning with Continuous State
Consider a system with a single continuous state 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 a discount factor $\gamma = 0.9$.

(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.

$Q(x, a) \leftarrow Q(x, a) + \alpha[R(x) + 0.9 \max_{a'} Q(x', a') - Q(x, a)]$,
where $a' \in a_1, a_2$, α indicates learning rate and R is the reward.

(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 a learning rate of 0.5?

$w_{i,1} \leftarrow w_{i,1} + \alpha[r + 0.9 \max_{a' \in a_1, a_2} Q(x', a') - Q(x, a_1)]\frac{\partial Q(x,a_1)}{\partial w_{i,1}}$, where $i = 0, 1, 2$;
$w_{i,1} \leftarrow 1+0.5[10+0.9(1+1+1)-(1+0.5+0.25)]\frac{\partial w_{0,1}+0.5w_{1,1}+0.25w_{2,1}}{\partial w_{i,1}}$, where $i = 0, 1, 2$;
$w_{0,1} \leftarrow 1 + 5.475 * 1 = 6.475$;
$w_{1,1} \leftarrow 1 + 5.475 * 0.5 = 3.7375$;
$w_{2,1} \leftarrow 1 + 5.475 * 0.25 = 2.36875$;

Problem 2: Policy Gradient with Continuous State
Assume that Q-function with function approximation is used together with the softmax function to form a policy $\pi_\theta(s, a) = e^{Q_\theta(s,a)}/\sum_{a'} e^{Q_\theta(s,a')}$. Assume that there are two actions with $Q(x, a_1) = w_{0,1} + w_{1,1}x + w_{2,1}x^2$ and $Q(x, a_2) = w_{0,2} + w_{1,2}x + w_{2,2}x^2$ for a real valued variable x.

(a) Give the update equations for the *REINFORCE* algorithm. Assume that the the return at the current step is G and the action taken is a_1.

$w_{i,j} \leftarrow w_{i,j} + \alpha * G\nabla_{w_{i,j}} \ln \pi_{w_{i,j}}(x, a_1)$,
where $i = 0, 1, 2$, $j = 1, 2$, and α indicates learning rate.

(b) Assume that $w_{i,j} = 1$ for all i, j and return $G = 5$ is received. What are the updated values of the parameters assuming $x = 0.5$ and a learning rate of 0.5?

$w_{i,j} \leftarrow w_{i,j} + \alpha * G\nabla_{w_{i,j}} \ln \pi_{w_{i,j}}(x, a_1)$, where $i = 0, 1, 2$ and $j = 1, 2$;
$w_{0,1} \leftarrow 1 + 0.5 * 5\nabla_{w_{0,1}} \ln \frac{e^{w_{0,1}+0.5+0.25}}{e^{w_{0,1}+0.5+0.25}+e^{1+0.5+0.25}} = 1 + 2.5 * \frac{e}{e+e} = 1 + 1.25 = 2.25$
$w_{1,1} \leftarrow 1+0.5*5\nabla_{w_{1,1}} \ln \frac{e^{1+0.5w_{1,1}+0.25}}{e^{1+0.5w_{1,1}+0.25}+e^{1+0.5+0.25}} = 1+2.5*\frac{0.5e^{0.5}}{e^{0.5}+e^{0.5}} = 1+0.625 = 1.625$

$$w_{2,1} \leftarrow 1 + 0.5 * 5\nabla_{w_{2,1}} \ln \frac{e^{1+0.5+0.25w_{2,1}}}{e^{1+0.5+0.25w_{2,1}}+e^{1+0.5+0.25}} = 1 + 2.5 * \frac{0.25e^{0.25}}{e^{0.25}+e^{0.25}} = 1.3125$$
$$w_{0,2} \leftarrow 1 + 0.5 * 5\nabla_{w_{0,2}} \ln \frac{e^{1+0.5+0.25}}{e^{1+0.5+0.25}+e^{w_{0,2}+0.5+0.25}} = 1 + 2.5 * \frac{-e}{e+e} = 1 - 1.25 = 0.25$$
$$w_{1,2} \leftarrow 1 + 0.5 * 5\nabla_{w_{1,2}} \ln \frac{e^{1+0.5+0.25}}{e^{1+0.5+0.25}+e^{1+0.5w_{1,2}+0.25}} = 1 + 2.5 * \frac{-0.5e^{0.5}}{e^{0.5}+e^{0.5}} = 1 - 0.625 = 0.325$$
$$w_{2,2} \leftarrow 1 + 0.5 * 5\nabla_{w_{2,2}} \ln \frac{e^{1+0.5+0.25}}{e^{1+0.5+0.25}+e^{1+0.5+0.25w_{2,2}}} = 1 + 2.5 * \frac{-0.25e^{0.25}}{e^{0.25}+e^{0.25}} = 0.6875$$