NAME: KARMA D TARAP
CSCI E-89c Deep Reinforcement Learning
Part I of Final

Suppose each state $s \in \mathcal{S}$ of the Markov Decision Process can be represented by a vector of 3 real-valued features: $\mathbf{x}(s) = (x_1(s), x_2(s), x_3(s))^T$.

Given some policy π, suppose we model the state value function $v_\pi(s)$ with a *fully connected feedforward neural network* (please see the table below) which has three inputs ($x_1(s)$, $x_2(s)$, and $x_3(s)$), one hidden layer that consists of two neurons (u_1 and u_2) with Leaky Rectified Linear Unit (Leaky ReLU) activation functions, and one output ($\hat{v}(s, \mathbf{w})$) with the Leaky ReLU activation function.

The explicit representation of this network is

input layer	hidden layer	output layer
x_1 x_2 x_3	$u_1 = f(w_{01}^{(1)} + w_{11}^{(1)} x_1 + w_{21}^{(1)} x_2 + w_{31}^{(1)} x_3)$ $u_2 = f(w_{02}^{(1)} + w_{12}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{32}^{(1)} x_3)$	$\hat{v} = f(w_0^{(2)} + w_1^{(2)} u_1 + w_2^{(2)} u_2)$

Here, $f(x)$ denotes the following Leaky ReLU:

$$f(x) = \begin{cases} x, & \text{if } x \geq 0, \\ 0.1x, & \text{if } x < 0. \end{cases}$$

Assume that the weights,

$$\mathbf{w} = \big(\underbrace{w_{01}^{(1)}, w_{11}^{(1)}, w_{21}^{(1)}, w_{31}^{(1)}, w_{02}^{(1)}, w_{12}^{(1)}, w_{22}^{(1)}, w_{32}^{(1)}}_{\text{hidden layer}}, \underbrace{w_0^{(2)}, w_1^{(2)}, w_2^{(2)}}_{\text{output layer}} \big)^T,$$

are currently estimated as follows:

hidden layer	output layer
$w_{01}^{(1)} = -0.8, w_{11}^{(1)} = 0.2, w_{21}^{(1)} = 0.3, w_{31}^{(1)} = 0.9$ $w_{02}^{(1)} = 0.3, w_{12}^{(1)} = -0.5, w_{22}^{(1)} = -0.2, w_{32}^{(1)} = -0.4$	$w_0^{(2)} = 0.1, w_1^{(2)} = -0.3, w_2^{(2)} = 1.4$

Assume the agent minimizes the mean squared error loss function,

$$L \doteq \frac{1}{2} \left(\hat{v}(S_t, \mathbf{w}) - v_\pi(S_t) \right)^2,$$

using Stochastic Gradient Descent (SGD), i.e. the Neural Network is trained in mini-batches of size 1.

If for current state S_t, the features are $x_1(S_t) = 1.2$, $x_2(S_t) = 0.4$, and $x_3(S_t) = 0.3$; and the agent "observes" $v_\pi(S_t)$ (this, of course, means the agent uses MC return,

1-step TD return, etc. as a "measurement" of $v_\pi(S_t)$) to be 3.2, please find the next SGD update of the weights using $\alpha = 0.1$:

$$\mathbf{w} - \alpha \nabla L,$$

where $\nabla L \doteq \Big(\underbrace{\frac{\partial L}{\partial w_{01}^{(1)}}, \frac{\partial L}{\partial w_{11}^{(1)}}, \frac{\partial L}{\partial w_{21}^{(1)}}, \frac{\partial L}{\partial w_{31}^{(1)}}, \frac{\partial L}{\partial w_{02}^{(1)}}, \frac{\partial L}{\partial w_{12}^{(1)}}, \frac{\partial L}{\partial w_{22}^{(1)}}, \frac{\partial L}{\partial w_{32}^{(1)}}}_{\text{hidden layer}}, \underbrace{\frac{\partial L}{\partial w_{0}^{(2)}}, \frac{\partial L}{\partial w_{1}^{(2)}}, \frac{\partial L}{\partial w_{2}^{(2)}}}_{\text{output layer}} \Big)^T.$

Please notice that the "measurement" of the state-value $v_\pi(S_t)$ here is considered to be independent of $\mathbf{w}$ (please see, for example, the Semi-gradient 1-step Temporal-Difference (TD) prediction).

SOLUTION:

Forward pass

$u_1 = f(w_{01}^{(1)} + w_{11}^{(1)} x_1 + w_{21}^{(1)} x_2 + w_{31}^{(1)} x_3) = f(-0.17) = -0.017$

$u_2 = f(w_{02}^{(1)} + w_{12}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{32}^{(1)} x_3) = f(-0.5) = -0.05$

$\hat{v} = f(w_0^{(2)} + w_1^{(2)} u_1 + w_2^{(2)} u_2) = f(0.0351) = 0.0351$

(a) Error associated with the output layer:

$$\varepsilon^{(2)} \doteq \frac{\partial L}{\partial \hat{v}} = \hat{v} - v_\pi(S_t) = 0.0351 - 3.2 = -3.1649$$

(b) Errors associated with the hidden layer:

$$\varepsilon_h^{(1)} \doteq \frac{\partial L}{\partial u_h}, \quad h = 1, 2.$$

$$\varepsilon_1^{(1)} \doteq \frac{\partial L}{\partial u_1} = -3.1649 \cdot f'(-0.5) \cdot -0.3 = 0.094947$$

$$\varepsilon_2^{(1)} \doteq \frac{\partial L}{\partial u_2} = -3.1649 \cdot f'(-0.5) \cdot 1.4 = -0.443086$$

(c) Partial derivatives of the loss function with respect to weights in the output layer:

$$\frac{\partial L}{\partial w^{(2)}} = (-3.1649, 0.0538033, 0.158245)^T$$

(d) Partial derivatives of the loss function with respect to weights in the hidden layer:

$$\frac{\partial L}{\partial w_1^{(1)}} = (0.0094947, 0.01139364, 0.00379788, 0.00284841)^T$$

$$\frac{\partial L}{\partial w_2^{(1)}} = (-0.0443086, -0.05317032, -0.01772344, -0.01329258)^T$$

(e) The next SGD update of the weights using $\alpha = 0.1$:

$$\mathbf{w} - \alpha \nabla L,$$

where $\nabla L \doteq \Big(\underbrace{\frac{\partial L}{\partial w_{01}^{(1)}}, \frac{\partial L}{\partial w_{11}^{(1)}}, \frac{\partial L}{\partial w_{21}^{(1)}}, \frac{\partial L}{\partial w_{31}^{(1)}}, \frac{\partial L}{\partial w_{02}^{(1)}}, \frac{\partial L}{\partial w_{12}^{(1)}}, \frac{\partial L}{\partial w_{22}^{(1)}}, \frac{\partial L}{\partial w_{32}^{(1)}}}_{\text{hidden layer}}, \underbrace{\frac{\partial L}{\partial w_{0}^{(2)}}, \frac{\partial L}{\partial w_{1}^{(2)}}, \frac{\partial L}{\partial w_{2}^{(2)}}}_{\text{output layer}} \Big)^T.$

$$\mathbf{w} - \alpha \nabla L$$

$$= (-0.8, 0.2, 0.3, 0.9, 0.3, -0.5, -0.2, -0.4, 0.1, -0.3, 1.4)^T$$

$$-\alpha(0.009, 0.011, 0.00379, 0.0028, -0.044, -0.053, -0.0177, -0.013, -3.1649, 0.0538, 0.158)^T$$

$$= (-0.8009, 0.19886, 0.2996, 0.8997, 0.304, -0.49468, -0.198, -0.39867, 0.416, -0.305, 1.384)^T$$