NAME: . . . .
CSCI S-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: