Please submit your solution in the next class

(1) In many instances, immediate rewards are worth more than those in the future. To take this observation into account, the value \( V(s_t) \) of a particular state \( s_t \) is not the sum of all future rewards, but rather the sum of all future, discounted rewards,

\[
V(s_t) = r(s_t) + \gamma r(s_{t+1}) + \gamma^2 r(s_{t+2}) + \ldots = \sum_{\tau=0}^{\infty} \gamma^\tau r(s_{t+\tau})
\]  

where \( 0 < \gamma < 1 \). Here \( s_t \) is the state at time \( t \), i.e., the state in which the agent is right now, \( s_{t+1} \) the state that the agent will move to next and so on. Following the derivation in the lecture, show that the temporal-difference-learning rule in this case is given by

\[
V(s_t) \rightarrow V(s_t) + \varepsilon (r(s_t) + \gamma V(s_{t+1}) - V(s_t))
\]  

(2) In the lecture, we talked about the necessity to introduce models for the value of a state, so that one could properly generalize to new, unseen situations. One very simple model is given by the value function \( V(\cdot) = \cdot \cdot \cdot \) where \( \cdot \cdot \cdot \) is a vector of stimuli that could either be present (1) or absent (0).

(a) Take the example of two stimuli, \( \cdot = (u_1, u_2) \). Let us assume that the subject (agent) has already learned the value of a state in which the first stimulus is present, and the value of a state in which the second stimulus is present. The learned values are given by

\[
V(\cdot = (1, 0)) = \alpha \\
V(\cdot = (0, 1)) = \beta
\]

What are the values of the parameters \( \cdot = (w_1, w_2) \) that the agent has learnt? Now we assume that the agent, for the very first time, runs into a state in which both stimuli are present. What is the value of this state? What if we now add some uncertainty. What would be the value of a state where the first stimulus has 50% chance of being present and the second stimulus has 10%? In what situation do you think this sort of a more generalized model that you just came up with would not make much sense?

(b) Derive the temporal-difference learning rule for the parameters \( \cdot \) that need to be learned if the value function is \( V(\cdot) = \cdot \cdot \cdot \). Hint: Start from a loss function \( \cdot \) - what would be a suitable choice? Can you also derive a learning rule if the value function were given by \( V(\cdot) = \mathcal{F}(\cdot) \) with \( \mathcal{F}(\cdot) \) being a known (non-linear) function?