Question 1. (5 points)

Recall the learning rate parameter α of the temporal difference learning.

1. (1 point) Provide range for the α parameter.

   Answer:

   It must always hold that $\alpha \in (0, 1)$.

2. (1 point) Explain the meaning of the α parameter.

   Answer:

   The learning rate α is a meta-parameter of the learning algorithm. It regulates the learning speed by balancing the newly seen sample and previous samples stored in the weights of the model.

3. (4 points) What must hold for α so that the temporal difference learning converges?

   Answer:

   The following must be true:

   $$\sum_{t=0}^{\infty} \alpha_t = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t^2 < \infty,$$

   i.e., the first sum must diverge, the second one must converge.

   The previous formula is a more general concept, named Robbins-Munro theorem, and holds outside TD-learning as well. See the next question.

4. (1 point) Relate the temporal difference update rule

   $$\widehat{U}(s) := \widehat{U}(s) + \alpha \left(r(s) + \gamma \cdot \widehat{U}(s') - \widehat{U}(s) \right)$$

   to another well-known algorithm used in mathematical optimization.

   Answer:

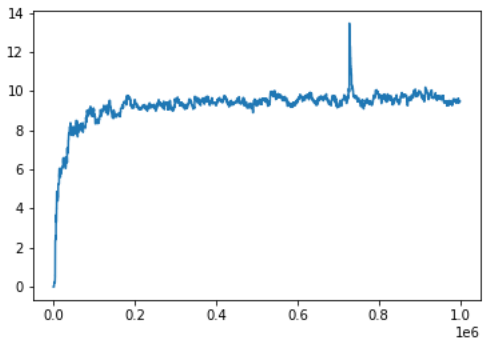
   The temporal difference learning rule is an instance of gradient descent.

   This TD-update is a special case of the Widrow-Hoff rule.

Question 2. (13 points)

In this problem, we will study the influence of learning rate α on the value estimates $\widehat{U}$. All figures show learning of U using the temporal-difference method for the same state over one million episodes. The learning rate was selected so that the conditions for convergence were met.

1. (2 points) Explain what causes the spike that you see around episode 700000.



Answer:

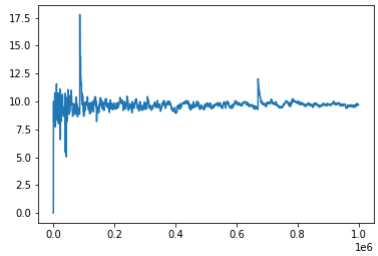
Learning is a stochastic process in a stochastic environment. Therefore, a rare event leading to a high reward might cause such a peak in learning.

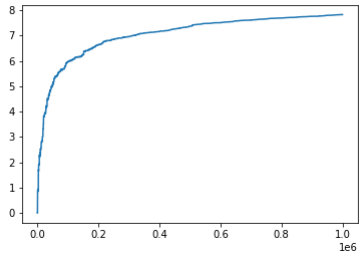
2. (2 points) Why do we need a different learning rate value for each state.

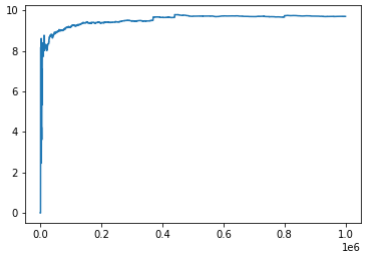
Answer:

The learning rate is a function of the number of visits of a state. For example, $\alpha(n_s) = \frac{1}{n_s}$. However, not all states are visited with the same frequency. Therefore, to speed up the learning process, we set different learning rates for different states. States that are visited frequently can converge faster than states that are visited rarely.

3. (6 points) Consider the following three scenarios of learning the value of a single state under a different learning rate. Explain which situation you consider optimal and identify when the learning rate was too small or too big. Propose a solution for the suboptimal cases.







Answer:

The first figure is an example of learning with a too high learning rate. This is explained by the shattering and high peaks in the beginning. As a result, the value estimates are oscillating around $U(s)$ even with one million epochs. The solution is either a lower learning rate or a higher number of learning episodes.

The second figure is an example of a too low learning rate. The values did not converge and are still rising to the $U(s)$. The solution is either to use a higher number of learning episodes, or to start with higher learning rate values.

The third figure does not suffer any of the problems stated above and should be considered a good example for learning of the state presented.

4. (3 points) The learning rate is a function of the number of visits of a state, i.e., $\alpha(n_s)$. Consider the following three functions

$$\alpha_1(n_s) = \frac{1}{10 + n_s}, \qquad \alpha_2(n_s) = \frac{3}{2 + n_s}, \qquad \alpha_3(n_s) = \frac{100}{99 + n_s}.$$

The figures in the question 3 were generated using those three learning rate functions. Match those functions to the figures and explain your decision.

Answer:

Learning rate α_1 is the smallest, α_3 is the highest. Therefore the figures were generated using α_3, α_1, and α_2 respectively.