Question 1. (5 points)

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

1. (1 point) Provide range for the α parameter.
2. (1 point) Explain the meaning of the α parameter.
3. (4 points) What must hold for α so that the temporal difference learning converges?
4. (1 point) Relate the temporal difference update rule

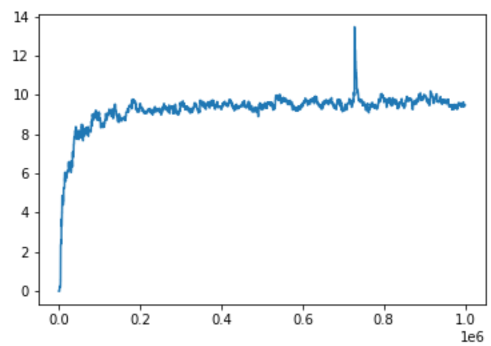
$$\widehat{U}(x) = \widehat{U}(x) + \alpha \left(r(x) + \gamma \cdot \widehat{U}(x') - \widehat{U}(obs) \right)$$

   to another well-known algorithm used in mathematical optimization.

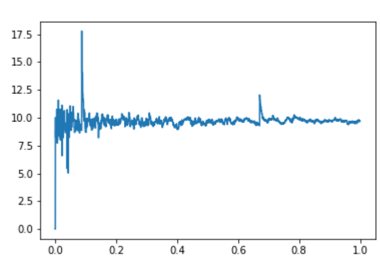
Question 2. (13 points)

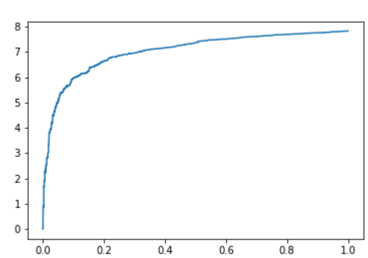
In this problem, we will study the influence of learning rate α on the value estimates $\widehat{U}$. All figures show learning of U using the temporal-difference method for the same state over one million episodes. The learning rate was selected so that the conditions for convergence were met.

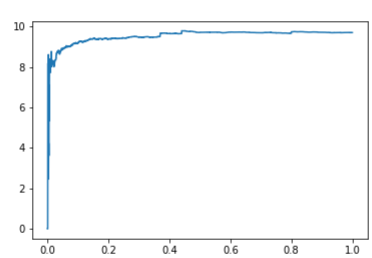
1. (2 points) Explain what causes the spike that you see around episode 700000.



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







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

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

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