Last name: Li First name: Minghan SID#: 1561234

Collaborators:

CMPUT 366/609 Assignment 1: Step sizes & Bandits

Due: Tuesday Sept 19 by gradescope

Policy: Can be discussed in groups (acknowledge collaborators) but must be written up individually

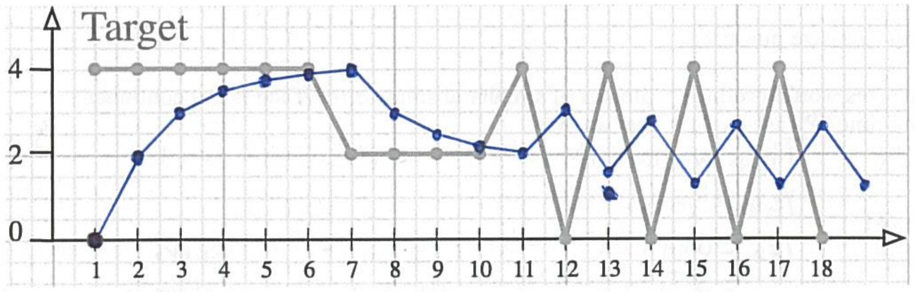
There are a total of 100 points on this assignment, plus 15 points available as extra credit!

Question 1 [50 points] Step-sizes. Plotting recency-weighted averages.

Equation 2.5 (from the SB textbook, 2nd edition) is a key update rule we will use throughout the course. This exercise will give you a better hands-on feel for how it works. This question has **five** parts.

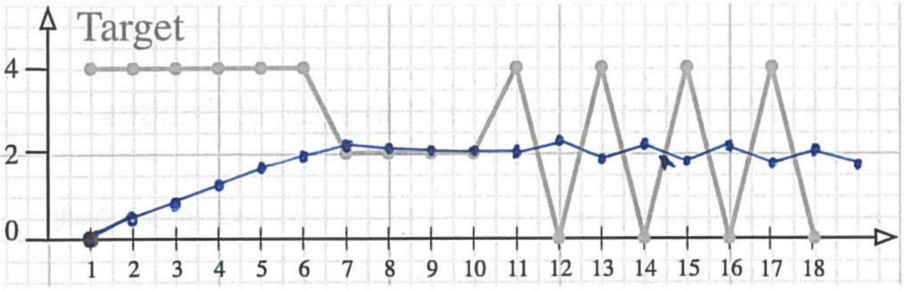
Do all the plots in this question by hand. To make it easier for you, I'll include some graphing area and a start on the first plot here, so you should just be able to print these pages out and draw on them.

Part 1. [15 pts.]
Suppose the target is 4.0 for six steps, then 2 for four steps, and then alternates between 4.0 and 0 for the remaining time steps, as shown by the grey line in the graph below. Suppose the initial estimate is 0 ($Q_1 = 0$), and that the step-size (in the equation) is 0.5. Your job is to apply Equation 2.5 iteratively to determine the estimates for time steps 1-19 (one time-step past step 18). Plot them on the graph below, using a blue pen, connecting the estimate points by a blue line. The first estimate Q_1 is already marked below:



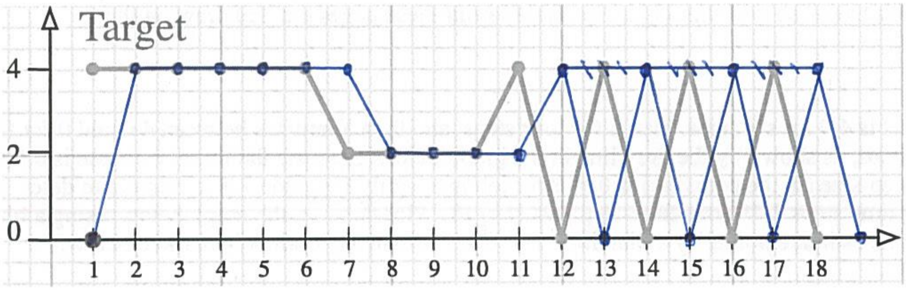
$\alpha = 1/2$

Part 2. [5 pts] Repeat the graphing/plotting portion of Part 1, this time with a step size of 1/8.



α = 1/8

Part 3. [5 pts.] Repeat with a step size of 1.0.



α = 1

Part 4. [10 pts.] Best step-size questions.

Which of these step sizes would produce estimates of smaller absolute error if the target continued alternating for a long time? Please explain your answer.

Answer : $\alpha = 1/8$

We set the target to be $T_n = 2 + (-1)^n \cdot 2$

$T_0 = 4$

Then we have the following recursive relationship:

$$Q_n = Q_{n-1} + \alpha (T_{n-1} - Q_{n-1})$$
$$= \alpha T_{n-1} + (1-\alpha) Q_{n-1}$$
$$= \cdots$$
$$= \alpha \sum_{i=0}^{n} (1-\alpha)^i T_{n-i-1} + (1-\alpha)^n Q_0$$

When $\alpha > 1$, the seriers diverges, so we can eliminate $\alpha = 1$. Then when $\alpha < 1$,

as $n \to \infty$, $(1-\alpha)^n Q_0 \to$

$\therefore \lim_{n \to \infty} Q_n = \alpha \sum_{i=0}^{n} (1-\alpha)^i T_{n-}$

we set error $= |T_n - Q_n$

Then we can get

$$\text{error} = \begin{cases} \frac{4\alpha}{1-(1-\alpha)^2} & n \text{ oc} \\ 4 - \frac{4\alpha(1-\alpha)}{1-(1-\alpha)^2} & n \text{ eve} \end{cases}$$

$\therefore \text{argmin}(\text{error}) = 1/8$

$\alpha = [1/2, 1/8]$

Which of these step sizes would produce estimates of smaller absolute error if the target remained constant for a long time? Please explain your answer.

Answer : $\alpha = 1$

We use the conclusion from the last question

~~error~~

We set $T_n = c$. Then

absolute error is

$$|T_n - Q_n| = \left| T_n - \alpha \sum_{i=0}^{n} (1-\alpha)^i T_{n-i-1} - (1-\alpha)^n Q_0 \right|$$
$$\xrightarrow{n \to \infty} \left| T_n - \alpha \sum_{i=0}^{n} (1-\alpha)^i T_{n-i-1} \right|$$
$$= \left| c - c \cdot \alpha \sum_{i=0}^{n} (1-\alpha)^i \right|$$

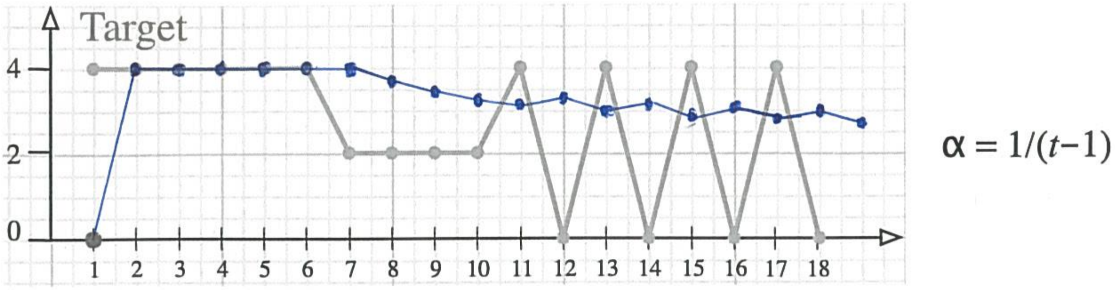
When $\alpha = 1$, $\alpha \cdot \sum_{i=1}^{n} (1-\alpha)^i = 1$

Then $|T_n - Q_n| = 0$

otherwise $|T_n - Q_n| > 0$

$\therefore \alpha = 1$

Part 5. [15 pts.] Repeat with a step size of 1/(*t*-1) (i.e., the first step size you will use is 1, the second is 1/2, the third is 1/3, etc.).



Based on all of these graphs, why is the 1/(t-1) step size appealing?

1. When $t=2$, $Q_n = T_{n-1}$, which is the unbiased estimation of the target, updates quickly.
2. When the target is constant, it can converge on it.; when the target alternates, it can also chase the target and get minimum absolute error as $n \to \infty$ (use the conclusion from the last question

Why is the 1/(t-1) step size not always the right choice?

Answer: Because when the target is slowly changing, or the target suddenly changes when t is alrealy very large, the update $1/(t-1)(T_{n-1}-Q_{n-1})$ becomes infinitesimally small, causing the Q value remain basically the same, therefore get the very wrong estimation of the target.

Question 2 [10 points] Bandit Example. Consider a multi-arm bandit problem with k = 5 actions, denoted 1, 2, 3, 4, and 5. Consider applying to this problem a bandit algorithm using ε-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$ for all a. Suppose the initial sequence of actions and rewards is $A_1 = 2$, $R_1 = -2$, $A_2 = 1$, $R_2 = 5$, $A_3 = 3$, $R_3 = 3$, $A_4 = 1$, $R_4 = 4$, $A_5 = 4$, $R_5 = 3$, $A_6 = 2$, $R_6 = -1$. On some of these time steps the ε case may have occurred causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

Action	Reward	Time Step	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$	$Q_t(5)$	random action occurs?
initialize		T_0	0	0	0	0	0	definitely
2	-2	T_1	0	-2	0	0	0	definitely ~~might~~
1	5	T_2	5	-2	0	0	0	might
3	3	T_3	5	-2	3	0	0	might
1	4	T_4	4.5	-2	3	0	0	might
4	3	T_5	4.5	-2	3	3	0	might
2	-1	T_6	4.5	-1.5	3	3	0	might

Answer: Random Actions will definitely occur on time step T_0, T_1 and might occur on the rest.

Bonus Question 1

Epsilon-Greedy
Advantages: easy to implement; can be easy to apply in complex settings
Disadvantages: sensitive to alpha, epsilon and initial Q estimates

UCB
Advantages: less sensitive to hyper-parameters compared to epsilon-greedy, learning rate can be pretty big and tuning the parameter c within some range won't hurt the performance too much.
Disadvantages: difficult to apply in more complex settings; uses much space because we have to store table of the number of selected actions.
*The plot of the UCB agent's performance is included in the zip file.

Bonus Question 2

$Q_n = \prod_{i=1}^{n}(1-\alpha_i)Q_0 + \sum_{i=1}^{n}\alpha_i\prod_{j=i+1}^{n}(1-\alpha_j)Ri$

Bonus Question 3
The spike will appear in almost every experiment, which is decided by the property of optimistic initialization. In the early steps because we can systematically explore, so the agent has a good chance to find the best action; however, because the actual reward of the optimal action is lower than the initial estimation, with increasingly choosing the optimal action, the Q value drops and the agent will choose other actions of higher Q value. Although the Q value of the optimal action is more accurate than the estimation of other acitons in this case, because we are using optimistic initialization, we are forcing the agent to explore more and therefore it will has a spike at the early stage.