Last name: Liu First name: Jiaqi SID#: jiaqi11 1514854

Collaborators: Zeyu Liu, Ricky Wang, Xiaohui Liu, Yuhang Xie

TAs from the lab

Online resources: github.com/JKCooper2. → for bonus question.

CMPUT 366Assignment 1: Step sizes & Bandits

Due: Tuesday Sept 18 by gradescope

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

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

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

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

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

... pages out and draw on them.

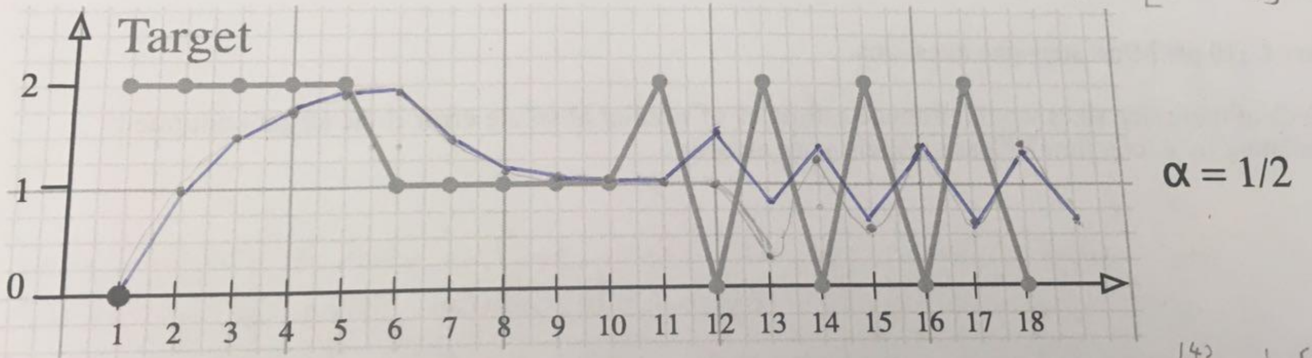
Part 1. [15 pts.]

Suppose the target is 2.0 for five steps, then 1 for five steps, and then alternates between 2.0 and 0 for 8 more steps, as shown by the grey line in the graph below. Suppose the initial estimate is 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. 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:

$Q_{15} = Q_{14} + \frac{1}{2} \times [0 - Q_{14}] = \frac{5135}{8192} \approx$

$Q_{18} = Q_{17} + \frac{1}{2}[2 - Q_{17}] \approx 1.328353882$ $\quad Q_{16} = Q_{15} + \frac{1}{2} \times [2 - Q_{15}] \approx 1.31341$

$Q_{19} = Q_{18} + \frac{1}{2}[0 - Q_{18}] \approx 0.664176944$ $\quad Q_{17} = Q_{16} + \frac{1}{2} \times [0 - Q_{16}] \approx 0.6567$



$Q_{n+1} = Q_n + \alpha[R_n - Q_n]$

$Q_2 = Q_1 + \frac{1}{2}[R_1 - Q_1] = 0 + \frac{1}{2}[2 - 0] = 1$

$Q_3 = Q_2 + \frac{1}{2}[R_2 - Q_2] = 1 + \frac{1}{2}[2 - 1] = \frac{3}{2} = 1.5$

$Q_4 = Q_3 + \frac{1}{2}[R_3 - Q_3] = \frac{3}{2} + \frac{1}{2}[2 - \frac{3}{2}] = \frac{7}{4} = 1.75$

$Q_5 = Q_4 + \frac{1}{2}[R_4 - Q_4] = \frac{7}{4} + \frac{1}{2}[2 - \frac{7}{4}] = \frac{15}{8} = 1.875$

$\frac{15}{}$ $\quad$ 31 $\quad$ 1.9375

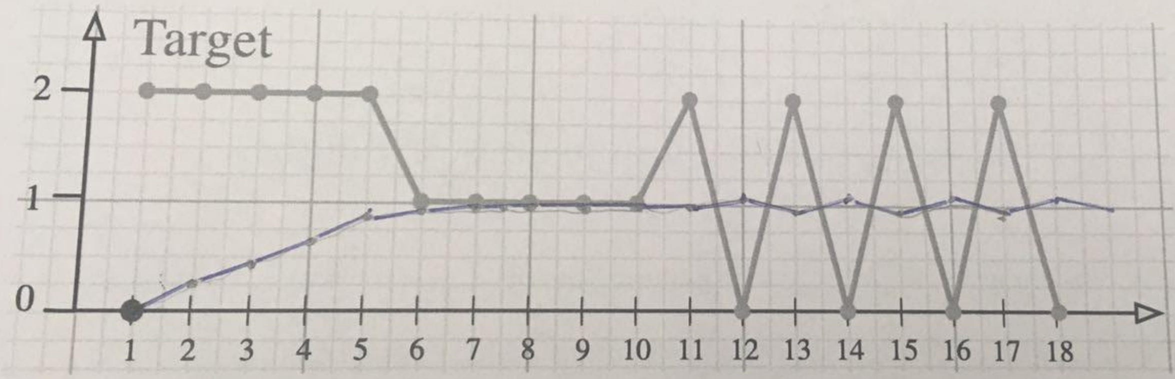
$Q_{10} = \frac{143}{128} + \frac{1}{2} \times [1 - \frac{143}{128}]$
$= 1.05859375$

$Q_{11} = \frac{271}{256} + \frac{1}{2} \times [1 -$
$= \frac{527}{512} = 1.02929$

$Q_{12} = \frac{527}{512} + \frac{1}{2} \times [1 -$
$= \frac{1039}{1024} = 1.01$

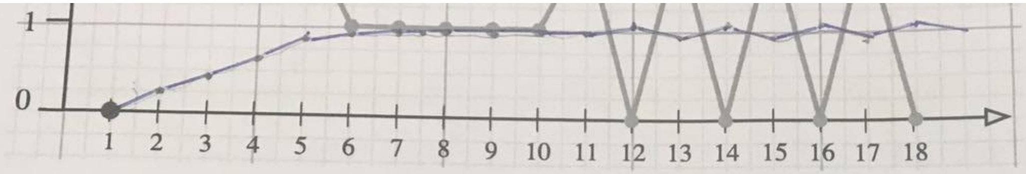
$Q_{13} = \frac{1031}{1024}$

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



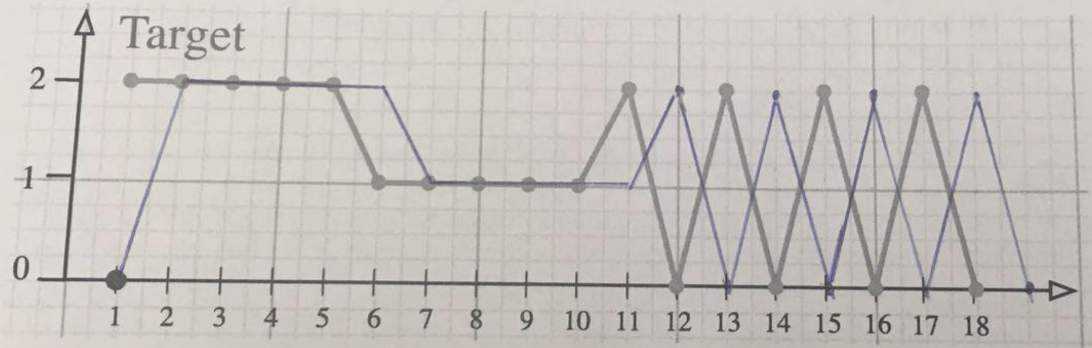
$\alpha = 1/8$

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



$\alpha = 1/8$

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



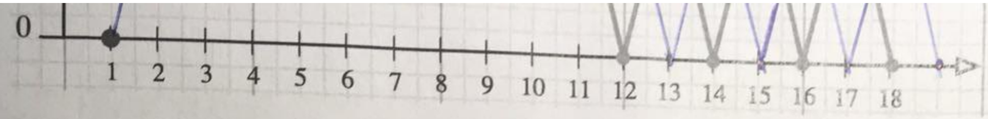
$\alpha = 1$

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.

$\alpha = \frac{1}{8}$

the absolute error for $\alpha = 1$ is 2×7 which is bigger than



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.

$\alpha = \frac{1}{8}$

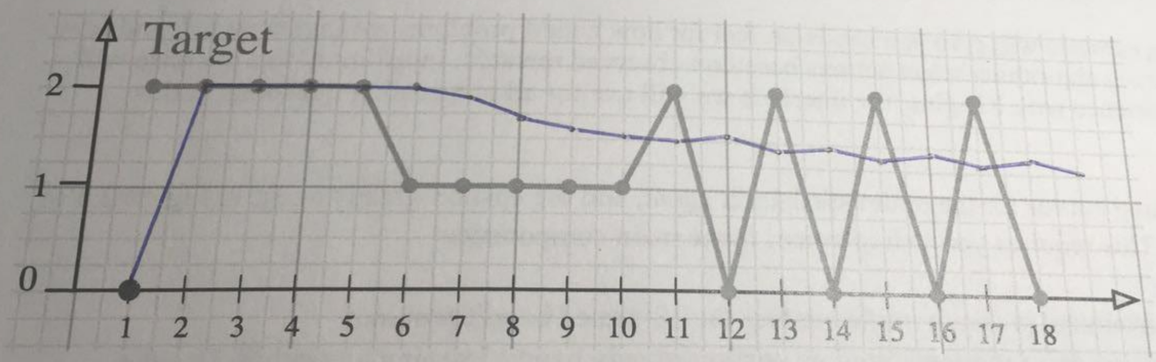
the absolute error for $\alpha=1$ is 2×7 which is bigger than $\alpha=\frac{1}{2}$, $\alpha=\frac{1}{2}$ has bigger fluctuation than $\alpha=\frac{1}{8}$, which cause a bigger absolute error. Based on these three graph, the bigger α is, the more fluctuation you will get on

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.

$\alpha=1$

the absolute error for $\alpha=1$ is $2+1=3$, which is smaller than $\alpha=\frac{1}{8}$ and $\alpha=\frac{1}{2}$ since it has better estimation ~~for~~ when target remained constant for a long time. Based on these three graph, ~~with sm~~, $\alpha=1$ would produce estimates of smaller absolute error if the target remained constant for a long time.

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



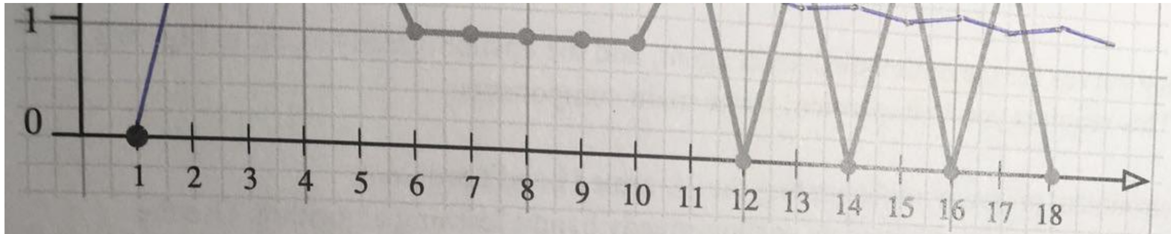
$\alpha = 1/(t-1)$

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

Because it provides better estimation than $\alpha = 1$ and $\alpha = \frac{1}{8}$, $\alpha = \frac{1}{2}$ no matter target is alternating or constent.

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

Because if the graph starts with the



$\alpha = 1/(t-1)$

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

Because it provides better estimation than $\alpha = 1$ and $\alpha = \frac{1}{8}$, $\alpha = \frac{1}{2}$ no matter target is alternating or constent.

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

Because if the graph starts with the alternating target and end with a constent target $\alpha = \frac{1}{(t-1)}$ will just generate a ~~smo~~ decreasing α, which ~~is~~ will give big absolute error

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?

① ② ③ ④ ⑤

0: 0 0 0 0 0

randomly ies exploration act

target $\alpha = \frac{}{(t-1)}$ will just generate a ~~sma~~ decreasing α, which ~~is~~ will give big absolute error constent

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?

		1	2	3	4	5
	0:	0	0	0	0	0
? not ε	1:	0	-2	0	0	0
? possible	2:	5	-2	0	0	0
✓ ε	3:	5	-2	3	0	0
✗ greedy	4:	>4	-2	3	0	0
✓ ε	5:	>4	-2	3	3	0
✓	6:	>4	[illegible]	3	3	0

By default, it choose randomly, its exploration action.

Since -2 is less than 0, it chose randomly again.

By greedy method it would choose 1, so its exploration again

This is greedy method, it is exploit.

1, 2, 4 ε possibly have occured

3, 5, 6 ε definitely have occured

by formula: $5 + \alpha(4-5)$

$= 5 - \alpha$

$\because \alpha \in (0, 1]$

$\therefore 5 - \alpha > 4$

2) all your code (including any graphing code used to generate your plot) [30 pts.]

Bonus Programing Question. [5 pts.]
Implement the UCB agent described in chapter two and evaluate it on the bandit environment from Question 3. Can you get the UCB agent to outperform the epsilon-greedy agent? Feel free to modify the parameters of the epsilon-greedy agent (alpha, epsilon, and the initial Q estimates) in order to better understand the relative strengths of both algorithms. Describe how we would go about determining and reporting on which agent is better for this task.

According to figure on Textbook Page 33, UCB could be the best option for this problem. If we choose larger parameters.

Bonus Question. [5 points extra credit]
Exercise 2.4 from Sutton and Barto (*Reward weighting for general step sizes*)

(The weight is produced in the sum denoted as. $\Pi(1-\alpha)^{n-i}$) α could be any value without specific action.

Bonus Question. [5 pts.]
Exercise 2.6 from Sutton and Barto (*Mysterious Spikes.* Use your implementation from Question 3 to better understand what is happening in Figure 2.3)

The oscillations in the early part of the curve may be due to the programe have worse or smaller Q values estimatrons for the poorly exploration. Because it use explore at these early stage with no clue whether the choice is good or bad. It has to try multiple times. before realize that they are bad. It takes time that these bad options drop below the best option.