1.1) Taking in consideration that the reward for every arm is uniformly distributed, we can assume the following:

Given the 7 arms (a_1, …, a_1), each one with a uniform distribution [x_i, y_i) the expected reward for each one would be:

// $R(a_i) = (x_i + y_i) / 2$ Expected Reward for arm 'i'
$R(a_1) = (x_1 + y_1) / 2 = (-2 + 3) / 2 = ½ = 0.5$
$R(a_2) = (x_2 + y_2) / 2 = (1 + 4) / 2 = 5/2 = 2.5$
$R(a_3) = (x_3 + y_3) / 2 = (2 + 3) / 2 = 5/2 = 2.5$
$R(a_4) = (x_4 + y_4) / 2 = (-1 + 5) / 2 = 4 / 2 = 2$
$R(a_5) = (x_5 + y_5) / 2 = (0 + 4) / 2 = 4 / 2 = 2$
$R(a_6) = (x_6 + y_6) / 2 = (1 + 4) / 2 = 5 / 2 = 2.5$
$R(a_7) = (x_7 + y_7) / 2 = (3 + 7) / 2 = 10 / 2 = 5$

Then, the expected if every arm is selected uniformly we can expect the following reward:

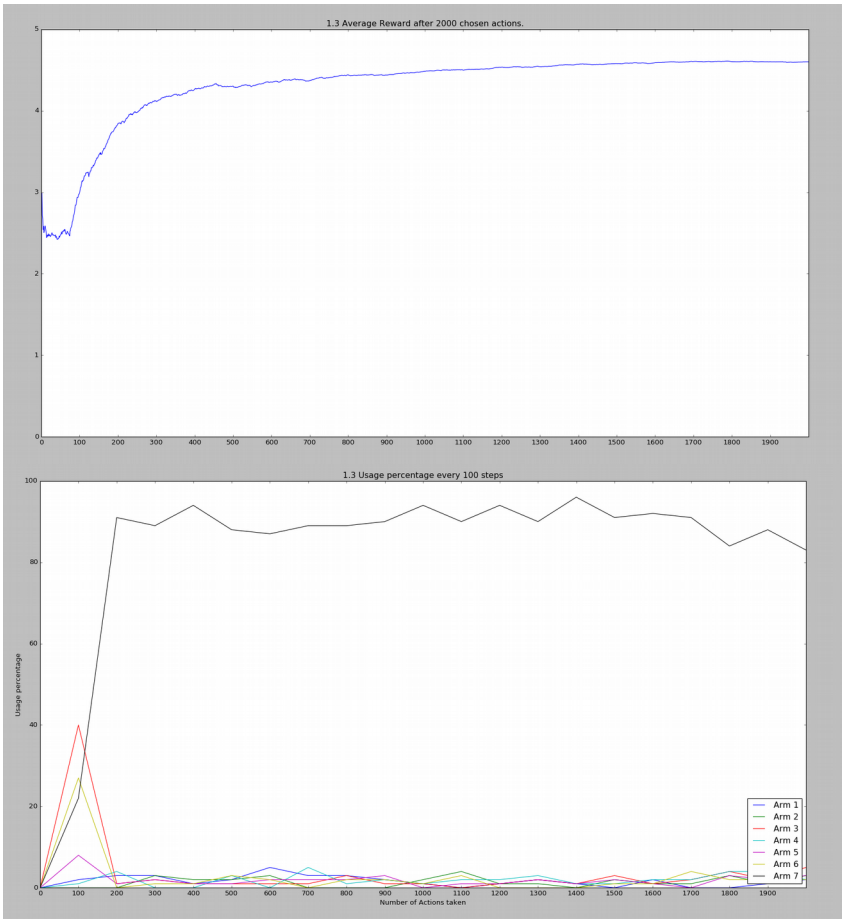
$(R(a_1) + R(a_2) + R(a_3) + R(a_4) + R(a_5) + R(a_6) + R(a_7)) / 7$

$= (0.5 + 2.5 + 2.5 + 2 + 2 + 2.5 + 5) / 7 = 17 / 7 = 2,43$ approx.

1.2) The sample average reward obtained after 20 uniformly chosen actions: 2.5837

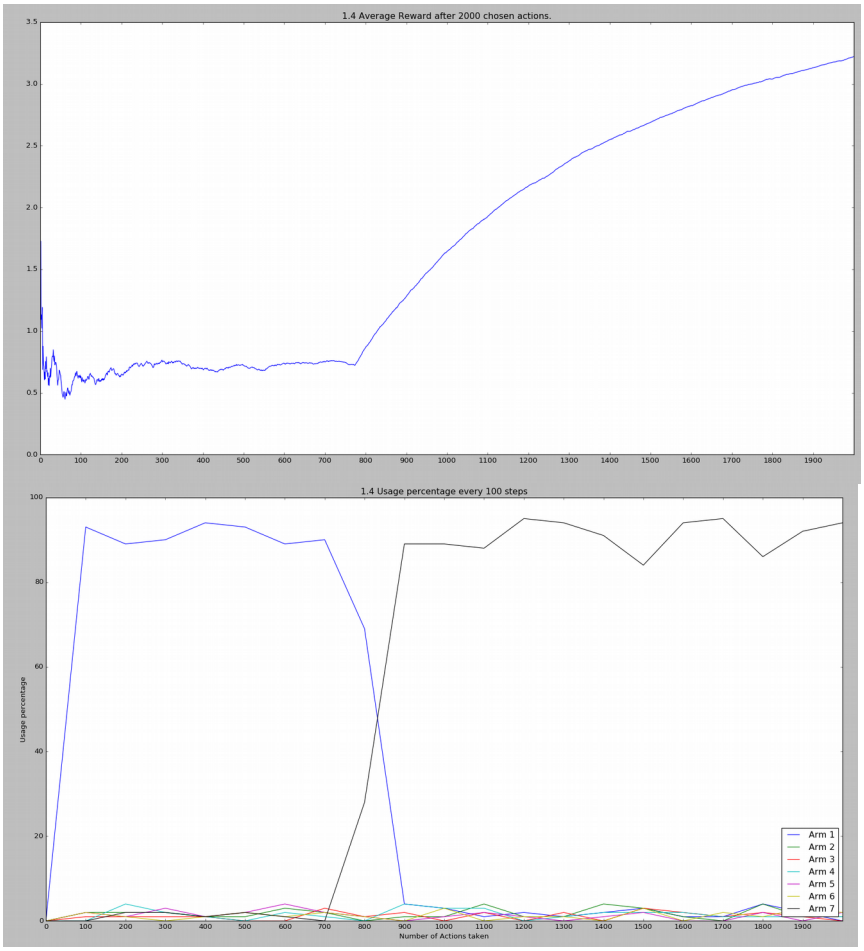
which is not very different result from the expected reward we got in 1.1) but in any case, very probably, with more and more chosen actions this value would be more and more similar.

1.3)



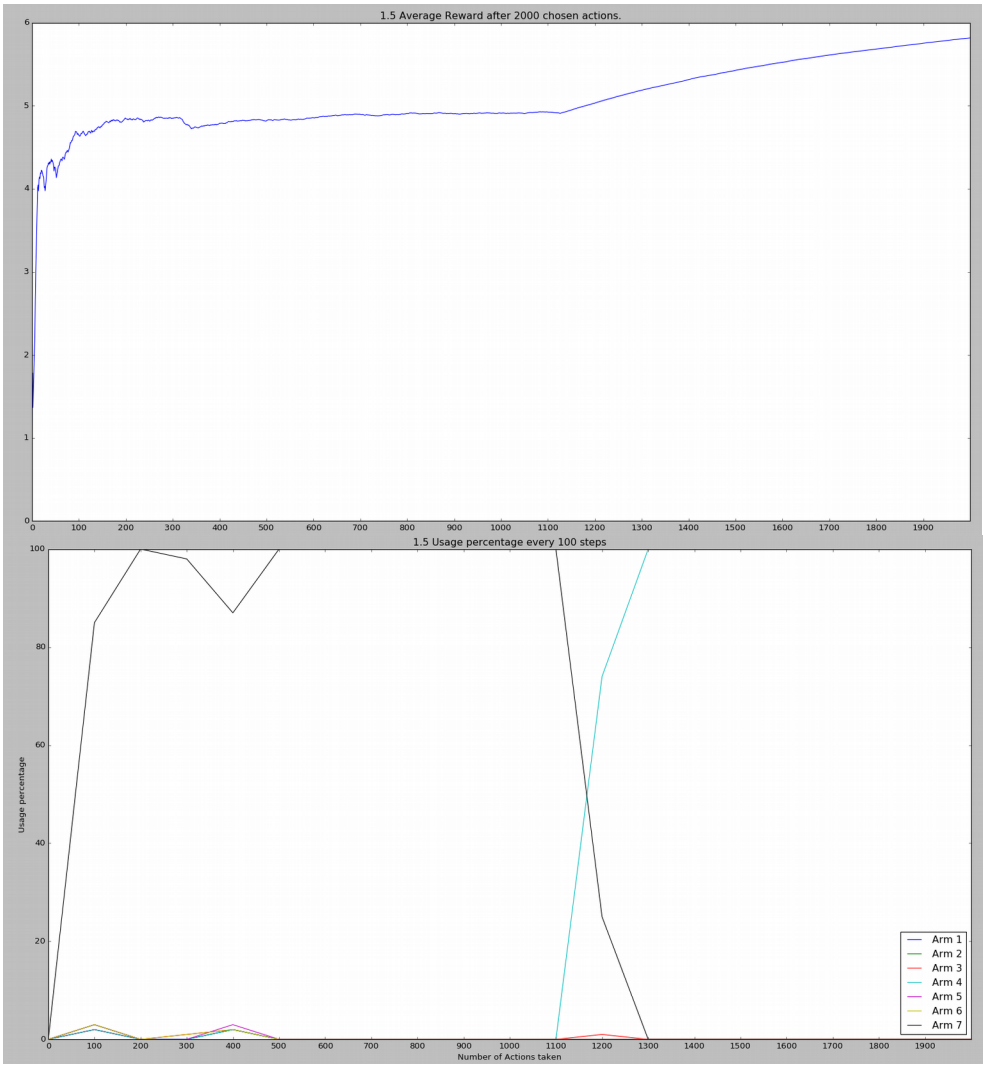
As shown in the plots, the Average reward was oscilating around 2.5 during the first 100 chosen actions, this is because the arm 3 and 6 where selected 40% and 33% of the times during this period, but after some iterations the arm 7 was detected as the best candidate which in fact it was and raised the average considerably above 4, after the algorithm learned that this option was the best, the other arms where selected apparently only when the exploration was needed.

1.4)



In this case, since we modified the learning rate, the algorithm had very hard times to know that the best option was the arm7 at the beginning, it took around 800 iterations to find this out, which affected the average reward hardly, and very probably, because of this learning rate change, the algorithm never positioned the arm4 very well, even when its average reward was growing, it was too slow for the arm4.

1.5)



In this case, we can see with a full greedy selection, the algorithm selected quickly arm7 as the best option and started using it, very probably because the expected value was high at the beginning and any other arm was discarded as best candidate with just a few iterations, but after around 100 hundred iterations after the distribution of arm4 was altered, the algorithm learned that this was a much better option to take,