COMP-597: Reinforcement Learning - Assignment 1

Posted Friday January 14, 2022
Due Thursday, January 27, 2022

The assignment can be carried out individually or in teams of two. Further instructions about how to submit will be provided by the TAs.

Bandit algorithms [100 points]

For this assignment, you will carry out some experimentation with bandit algorithms, in order to help you understand what we discussed in class, and to get used to the way in which we will run experiments for other assignments as well. You should submit a notebook with your code, results and explanations.

1. [5 points] Write a small simulator for a Bernoulli bandit with k arms. The probability of success p_i for each arm $i \in \{1, ...k\}$ should be provided as an input. The bandit should have a function called "sample" which takes as input the index of an action and provides a reward sample. Recall that a Bernoulli bandit outputs either 1 or 0, drawn from a binomial distribution of parameter p_k.

2. [5 points] Test your code with 3 arms of parameters $q_* = [0.7, , 0.7 - \delta, 0.7 - 2\delta$, with $\delta = 0.2$. Generate and save a set of 100 samples for each action. For the test, plot one graph for each action, containing the reward values obtained over the 100 draws, the empirical mean of the values, and the true q_* for each arm. Each graph will have an x-axis that goes to 100, two horizontal lines (true value and estimated value) and a set of points of value 0 and 1.

3. [10 points] Code the rule for estimating action values discussed in lecture 2, with a fixed learning rate α, in a function called "update", and using the incremental computation of the mean presented in lecture 2, in a function called `updateAvg` Using the previous data, plot for each action a graph showing the estimated q value as a function of the number of samples, using averaging as well as $\alpha = 0.01$ and $\alpha = 0.1$, and the true value. Each graph should have two curves and a horizontal line.

4. [10 points] Repeat the above experiment 5 times, starting with action value estimates of 0. Each run will still contain 100 samples for each action. Plot the same graph as above, but where the curves have the average and standard error over the 5 runs. Explain in 1-2 sentences what you observe. Which of the α values is better? How do they compare to averaging? If you wanted to optimize further, in what range of α would you look for better values?

5. [20 points] Code the ϵ-greedy algorithm discussed in class, with averaging updates, with ϵ provided as an input. We will run 5 independent runs. In each run, we will have 1000 time steps. We are interested in the following graphs:

   (a) The reward received over time, averaged at each time step over the 5 independent runs (with no smoothing over the time steps), and the standard error over the 5 runs

(b) The fraction of runs (out of 5) in which action 1 (which truly is best) is also estimated best based on the q values

(c) The instantaneous regret l_t (as discussed in lecture 3) (averaged over the 5 runs)

(d) The total regret L_t up to time step t (as discussed in lecture 3) (averaged over the 5 runs)

Generate this set of graphs, for the following values of ϵ: 0, 1/8, 1/4, 1/2, 1. Explain what you observe in the graphs. What is the best value of ϵ to use?

6. [5 points] Set $\delta = 0.02$ and run the experiment for $\epsilon = 1/4$ in this case. Compare the result with the one you obtained for the same ϵ in the previous problem. What do you observe and why?

7. [10 points] Write a function that decays ϵ, but in a simpler way than discussed in lecture 3. This algorithm will take as input ϵ and a decay factor $\lambda \in (0, 1)$ and will multiply ϵ by λ at every step: $\epsilon \leftarrow \epsilon\lambda$. Run the same experiment as above, with $\delta = 0.2$ but with a fixed starting value $\epsilon = 1/2$ and values of $\lambda = 0.999$ and $\lambda = 0.99$. Explain what you observe, and compare to the fixed ϵ case.

8. [15 points] Let us now consider a non-stationary problem. Let $\delta = 0.2$ and imagine that after 500 time steps, the parameter of actions 2 and 3 become $0.7 + \delta$ and $0.7 + 2\delta$ respectively. We will compare 4 conditions: fixed α vs averaging and fixed ϵ vs decaying. Based on your previous experimentations, pick values of α, ϵ and λ that you want to use, and explain why you picked these. Using these values, run 5 runs and plot the reward graph averaged over these runs. The graph should have 4 lines (with standard error bars). Explain what you see in the graph. Based on these results, which condition is best suited to cope with non-stationarity?

9. [20 points] Write a function that uses softmax (aka Boltzmann) exploration, as discussed in lecture 4. This function will use averaging for the value estimation, and fixed values of the temperature of 1000, 10, 1 and 0.1. Run the same experiment as above with $\delta = 0.2$ and explain what you see. Discuss the similarities and differences of softmax and ϵ-greedy based on these results.