RL

Homework-1

Ans 2. The oscillations and spikes in the early part of the curve for the optimistic method represents that the agent explores an ~~new~~ action and then gets disappointed by the reward for that action and explore a new action for the next time step. So, after trying out all the actions, the percentage of selecting an optimal action rises. (around 35 – 40 %). The spikes in the graph are caused because many bandit agents selected an optimal action at ~~a~~ that particular timestep.

Ans 4 ~~Ex 2.7~~ As can be seen from the graph of Average Rewards in stationary case, the slope of graph of UCB is more ~~increasing~~ than the graphs of ∈-greedy & optimal initial values. for time steps greater than 1000. ~~Hence~~, So, the optimal action taken by UCB is more than the rest.

For non-stationary case, the percentage of optimal action is reduced to 8 - 10 % but UCB is still the best method among the 3. ~~because~~ ~~For the~~ Since UCB explores more, it has higher optimal action percentage.

Ans 3 Ex. 2.7

$$Q_n = Q_{n-1} + B_n \left[R_{n-1} - Q_{n-1}\right]$$

$$Q_n = Q_{n-1}(1-B_n) + B_n R_{n-1}$$

$$Q_{n-1} = Q_{n-2}(1-B_{n-1}) + B_{n-1} R_{n-2}$$

$$Q_n = (1-B_n)(1-B_{n-1})Q_{n-2} + (1-B_n)B_{n-1}R_{n-2} + B_n R_{n-1}$$

$$\vdots$$

$$Q_n = \prod_{i=1}^{n}(1-B_i)\,Q_1 + \sum_{i=1}^{n-1} B_{i+1} R_i \prod_{j=i+1}^{n}(1-B_j)$$

Given: $B_n = \dfrac{\alpha}{\bar{O}_n}$

$$1-B_i = 1-\frac{\alpha}{\bar{O}_i} = \frac{\bar{O}_i - \alpha}{\bar{O}_i} = \frac{\bar{O}_{i-1}(1-\alpha) + \alpha - \alpha}{\bar{O}_i} = (1-\alpha)\frac{\bar{O}_{i-1}}{\bar{O}_i}$$

So, $$\prod_{i=1}^{n}(1-B_i) = \prod_{i=2}^{n}(1-B_i)\left[(1-\alpha)\frac{\bar{O}_0}{\bar{O}_1}\right]$$

& given that $\bar{O}_0 = 0$ (zero)

So, $$\prod_{i=1}^{n}(1-B_i) = 0$$

So, $Q_n = \sum_{i=1}^{n-1} B_{i+1} R_i \prod_{j=i+1}^{n} (1 - B_j)$

$$= \sum_{i=1}^{n-1} B_{i+1} R_i \prod_{j=i+1}^{n} (1-\alpha)\left(\frac{\bar{O}_{j-1}}{\bar{O}_j}\right)$$

$$= \sum_{i=1}^{n-1} (1-\alpha) B_{i+1} R_i \left[\frac{\bar{O}_{\cancel{j-1}}}{\cancel{\bar{O}_j}} \times \frac{\cancel{\bar{O}_j}}{\bar{O}_{j+1}} \times \cdots \times \frac{\cancel{\bar{O}_{n-1}}}{\bar{O}_n}\right]_{j=i+1}$$

$$= \sum_{j=1}^{n-1} (1-\alpha) B_{i+1} R_i \frac{\bar{O}_i}{\bar{O}_n}$$

$$= \frac{1-\alpha}{\bar{O}_n} \sum_{i=1}^{n-1} \frac{\bar{O}_i}{} B_{i+1} R_i$$

$$= \frac{(1-\alpha)\alpha}{\bar{O}_n} \sum_{i=1}^{n-1} \frac{\bar{O}_i}{\bar{O}_{i+1}} R_i$$

$$= \frac{(1-\alpha)\alpha}{\bar{O}_n} \sum_{i=1}^{n-1}$$