Reinforcement Learning Homework 1

Lavanya Verma (lavanya18155@iiitd.ac.in)

September 8, 2021

1. Problem 3

Assuming the bandits are stationary and ϵ-greedy has run for long enough for it to figure out optimal sequence of actions. Then ϵ-greedy algorithm will choose the optimal action for (1-ϵ)*100% of the time, and for the rest ϵ factor it will choose sub-optimal action.

Average reward,

$$\begin{aligned} E[R] &= \sum_{i=1}^{n} \frac{\epsilon}{n} q_i + (1-\epsilon) q^{opt} \\ &= \epsilon \sum_{i=1}^{n} \frac{q_i}{n} + (1-\epsilon) q^{opt} \\ &= \epsilon \bar{q} + (1-\epsilon) q_{opt} \end{aligned}$$

Epsilon	Average Reward	Exploration	Exploration in 2000 steps
0.01	0.01 + 0.99*q_{opt}	99%	10 steps
0	1*q_{opt}	100%	0 steps

But in case $\epsilon = 0$, assumption may even take longer to find optimal switch since it only rely on the optimal action, discourages exploration. Using OIV with greedy algorithm, allows it to exploit the best optimal action, by strategically forcing exploration in the initial phase.

2. Problem 4

Equation 2.5

$$\begin{aligned} Q_{n+1} &= Q_n + \alpha_n (R_n - Q_n) \\ &= \alpha_n R_n + (1-\alpha_n) Q_n \\ &= \alpha_n R_n + (1-\alpha_n) \left[\alpha_{n-1} R_{n-1} + (1-\alpha_{n-1}) Q_{n-1} \right] \\ &= \alpha_n R_n + (1-\alpha_n) R_{n-1} \alpha_{n-1} + (1-\alpha_n)(1-\alpha_{n-1}) Q_{n-1} \\ \Rightarrow &= Q_1 \prod_{i=1}^{n} (1-\alpha_i) + \sum_{i=1}^{n-1} (\alpha_i R_i \prod_{j=i+1}^{n} (1-\alpha_j)) + \alpha_n R_n \end{aligned}$$

For sample average,

$$\alpha_i = \frac{1}{i}; \forall i \geq 1$$

So the contribution of the term

$$Q_1 \prod_{i=1}^{n} (1 - \alpha_i)$$

in sample average goes to zero(i = 1).
If $\alpha_i = \alpha$, where α is a constant.

$$Q_{n+1} = (1 - \alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1 - \alpha)^{n-i} R_i$$

As $\alpha \in (0, 1]$, for higher α,

$$(1 - \alpha)^n \approx 0$$

Contribution of Q_1 would be higher for smaller alpha because of the exponent being smaller than 1.

3. Problem 2.8

Log term in UCB gives unbounded confidence to each action during the initial phase, thereby promoting exploration. In the initial ten steps algorithm chooses each action one by one, at eleventh step it chooses the action which highest reward since start, choosing most optimal action at eleventh step. Spike resulted out due to constructed inference of all 2000 games, however after the eleventh step its choice of action reward it dictated on the specific values(mean) of arms and stochasticity of system, therefore the pattern breaks and hence the demise.