Date
DELTA Pg No.

2. Exercise - 2.6

In Initial steps, Maximum agents will perform same action and so return rewards for all tests. are not good enough.

They will select optimal action with high probability depending on value of returned reward., once they tried all actions.

Spike in initial stage of graph is due to the most of agents selects the optimal action at that time stamp & they try to explore by selecting another actions. Due to this, we observed Spike in initial stage of graph.

Scanned with CamScanner

4.

When we follow greedy approach it look best action at present while non-greedy (like ε-greedy) by to explore in search of better reward while selecting action, it is best way to select action that action is how close to estimates are to be being maximal optimal & there are less uncertainties in estimation.

One way is to do it by setting Upper bound Condition. It is called Upper Confidence Bound.

$$A_t = \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right] \quad \text{--- (i)}$$

$Q_t(a)$ will be selected using non-greedy approach.
$N_t(a)$: No. of times action 'a' has to selected prior to time, t & c is constant for controlling exploration.

UCB doesn't support non-stationary problems because even we are updating action values is best way.

There is much diff. betn UCB & Optimist initial values, UCB choose deterministically favouring actions with inital q-values to encourage exploration in inital phase with no bias & fast convergence.
UCB is not applicable for non-stationary case.

Scanned with CamScanner