Start Initialize simulator
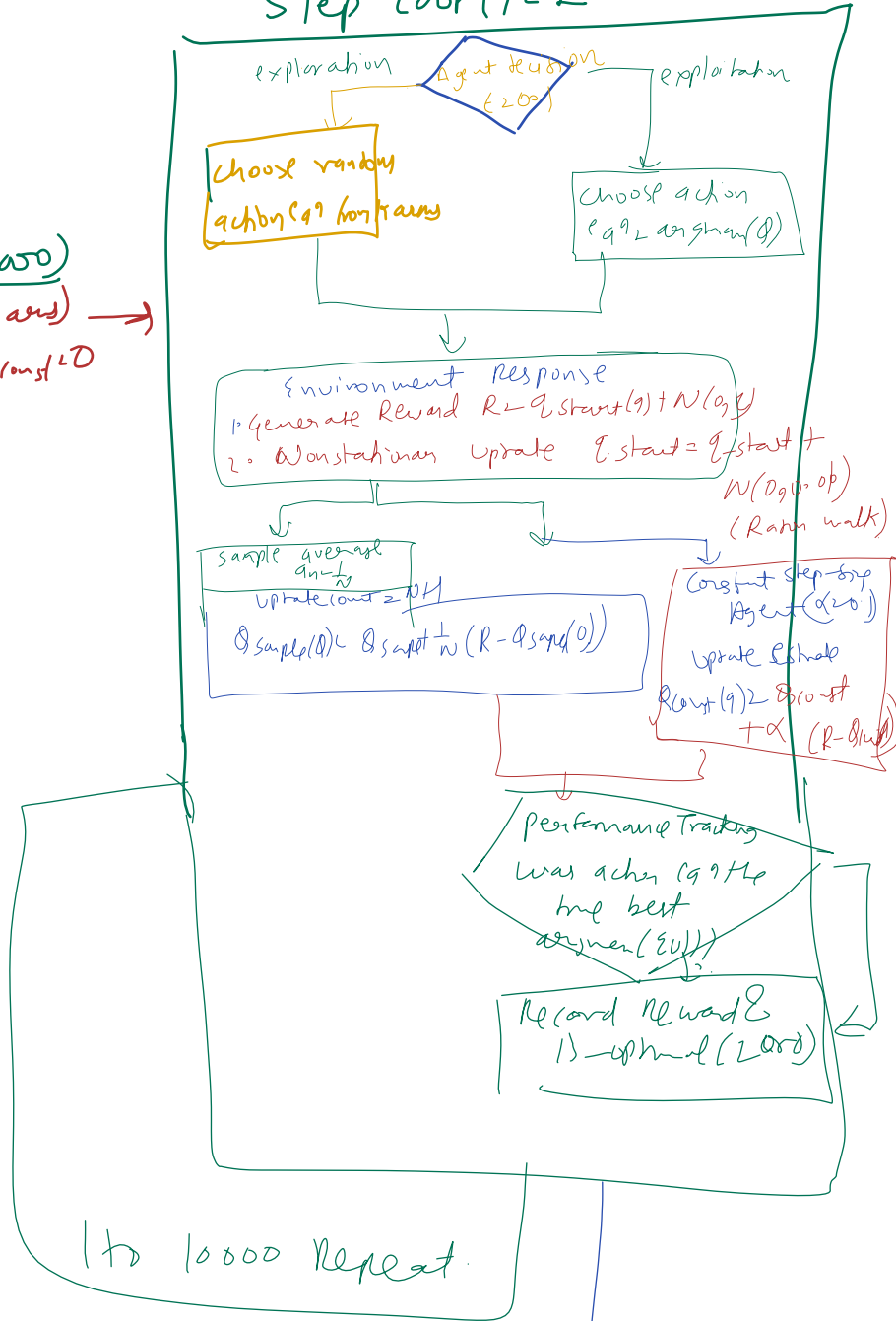
Initialization (Per Run, R = 2000)
- Initalize q-star = [0, 0, 0] (K = 10 arms)
- Initalize Q-estimates $Q_{sample} = 0$, $Q_{const} = 0$
- Initial Action count $N = 0$

Step Loop (T = 1 to 10000 steps)

exploration — Agent decision ($\xi = 20\%$) — exploitation

choose random action (q) for training

choose action e.g. $q = \arg\max(Q)$

Environment Response
1. Generate Reward $R = q_{star}(q) + N(0,1)$
2. Nonstationary Update $q_{star} = q_{star} + N(0, 0.01)$ (Random walk)

sample average $q_{n+1} = \frac{1}{N}$
update (count = 0+1)
$Q_{sample}(0) \leftarrow Q_{sample} + \frac{1}{N}(R - Q_{sample}(0))$

Constant step-size Agent ($\alpha = 0.1$)
update estimate
$Q_{const}(q) = Q_{const} + \alpha(R - Q_{const})$

Performance Tracking
was action (q) the true best $\arg\max(q_{star})$?

Record Reward & 1 = optimal (2000)

1 to 10000 Repeat.

if T = 10,000 end Run

Aggregation & Visualization

Plot 1 Average Reward vs steps

Plot 2 optimal Action vs steps

Analysis → plot and Const step tracks better