having mean of each arms are respectively $\hat{\mu}_1 = 0.3$, $\hat{\mu}_2 = 0.5$, $\hat{\mu}_3 = 0.4$. Let say that we got these sample mean after 100 times of sampling. where the first arm $N_1 = 25$ times second one $N_2 = 55$ times, third one $N_3 = 20$ times selected. Find the $\hat{\mu}_1, \hat{\mu}_2$ & $\hat{\mu}_3$ at 101, 102 and 103 times. The rewards are given by 8, 27 and 8.

epoch	arm 1	arm 2	arm 3
101	5	3	1
102	3	1	2
103	4	5	3

when $t = 101$,

$$= \left(\hat{\mu}_1 + \sqrt{\frac{2\log(t)}{N_1}}, \hat{\mu}_2 + \sqrt{\frac{2\log(t)}{N_2}}, \hat{\mu}_3 + \sqrt{\frac{2\log(t)}{N_3}} \right)$$

$$= \left(\underbrace{0.3 + \sqrt{\frac{2\log(101)}{25}}}_{\hat{\mu}_1}, \underbrace{0.5 + \sqrt{\frac{2\log(101)}{55}}}_{\hat{\mu}_2}, \underbrace{0.4 + \sqrt{\frac{2\log(101)}{20}}}_{\hat{\mu}_3} \right)$$

$$= (0.7, 0.76, 0.84).$$

Now $\max(0.7, 0.76, 0.84) = 0.84$.

Thus, $t = 101$ the $\max = 0.84 \Rightarrow$ arm 3 the algorithm will choose arm 3.

Now updating the mean and reward of arm 3.

$N_3 = 21$, reward, $r = 8 + 1 = 9$.

$$\text{mean}, \hat{\mu}_3 = \frac{9}{21} = 0.428.$$

Now, new → $\hat{\mu}_1 = 0.3$, $\hat{\mu}_2 = 0.5$, $\hat{\mu}_3 = 0.428$.

For $t = 102$;

$$\left(\hat{\mu}_1 + \sqrt{\frac{2\log(t)}{N_1}}, \hat{\mu}_2 + \sqrt{\frac{2\log(t)}{N_2}}, \hat{\mu}_3 + \sqrt{\frac{2\log(t)}{N_3}} \right)$$

$$= 0.3 + \sqrt{\frac{2\log(102)}{25}}, 0.5 + \sqrt{\frac{2\log(102)}{55}}, 0.43 + \sqrt{\frac{2\log(102)}{21}}$$

$$= 0.70, 0.77, 0.86$$

Now $\max(0.70, 0.77, 0.86) = 0.86$

For $t = 101$ the max = 0.86 ⇒ arm 3 the algorithm will choose arm 3.

Now updating the mean and reward of arm 3

$N_3 = 22$, reward, $r = 8 + 2 = 10$

$$\text{mean } \hat{\mu}_3 = \frac{10}{22} = 0.45.$$

Now, new → $\hat{\mu}_1 = 0.3$, $\hat{\mu}_2 = 0.5$, $\hat{\mu}_3 = 0.45$

For $t = 103$;

$$\left(\hat{\mu}_1 + \sqrt{\frac{2\log(t)}{N_1}}, \hat{\mu}_2 + \sqrt{\frac{2\log(t)}{N_2}}, \hat{\mu}_3 + \sqrt{\frac{2\log(t)}{N_3}} \right)$$

$$= 0.3 + \sqrt{\frac{2\log(103)}{25}}, 0.5 + \sqrt{\frac{2\log(103)}{55}}, 0.45 + \sqrt{\frac{2\log(103)}{22}}$$

$$= 0.701, 0.77, 0.88$$

Now $\max(0.70, 0.77, 0.88) = 0.88$

For $t = 101$ the max = 0.88 ⇒ arm 3 the algorithm will choose arm 3.

Now updating the mean and reward of arm 3

$N_3 = 23$, reward, $r = 8 + 3 = 11$

$$\text{mean } \hat{\mu}_3 = \frac{11}{23} = 0.48.$$

Now, new → $\hat{\mu}_1 = 0.3$, $\hat{\mu}_2 = 0.5$, $\hat{\mu}_3 = 0.48$