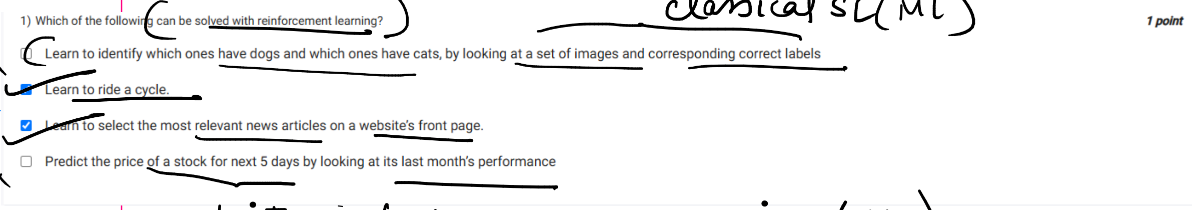
RL-Week-1 Practice Assignment - 1

classical SL(ML)



1) Which of the following can be solved with reinforcement learning? 1 point

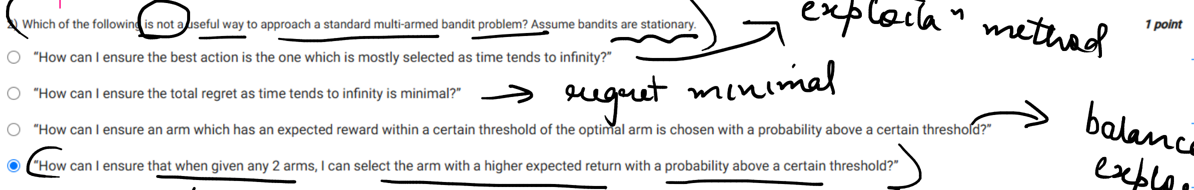
□ Learn to identify which ones have dogs and which ones have cats, by looking at a set of images and corresponding correct labels

☑ Learn to ride a cycle.

☑ Learn to select the most relevant news articles on a website's front page.

□ Predict the price of a stock for next 5 days by looking at its last month's performance

historical data + regression (ML)



2) Which of the following is not a useful way to approach a standard multi-armed bandit problem? Assume bandits are stationary. 1 point

exploitation method

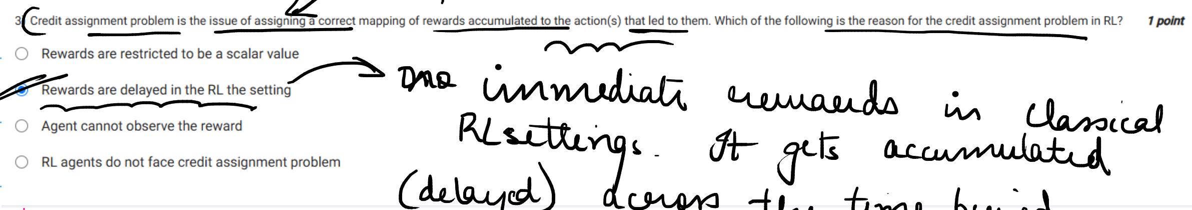
○ "How can I ensure the best action is the one which is mostly selected as time tends to infinity?"

○ "How can I ensure the total regret as time tends to infinity is minimal?" → regret minimal

○ "How can I ensure an arm which has an expected reward within a certain threshold of the optimal arm is chosen with a probability above a certain threshold?" → balances exploration & exploitation

● "How can I ensure that when given any 2 arms, I can select the arm with a higher expected return with a probability above a certain threshold?"

→ not a correct approach to deal with optimality



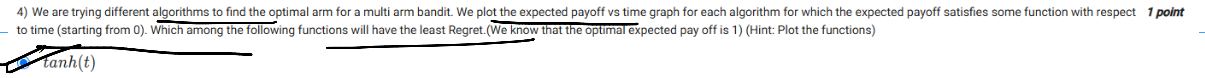
3) Credit assignment problem is the issue of assigning a correct mapping of rewards accumulated to the action(s) that led to them. Which of the following is the reason for the credit assignment problem in RL? 1 point

○ Rewards are restricted to be a scalar value

● Rewards are delayed in the RL the setting → No immediate rewards in classical RL settings. It gets accumulated (delayed) across the time period.

○ Agent cannot observe the reward

○ RL agents do not face credit assignment problem



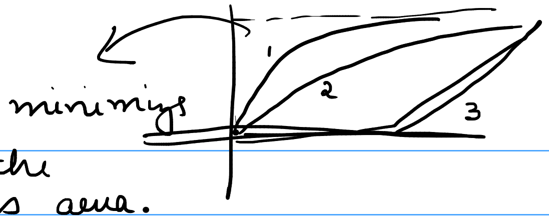
4) We are trying different algorithms to find the optimal arm for a multi arm bandit. We plot the expected payoff vs time graph for each algorithm for which the expected payoff satisfies some function with respect to time (starting from 0). Which among the following functions will have the least Regret.(We know that the optimal expected pay off is 1) (Hint: Plot the functions) 1 point

● $tanh(t)$

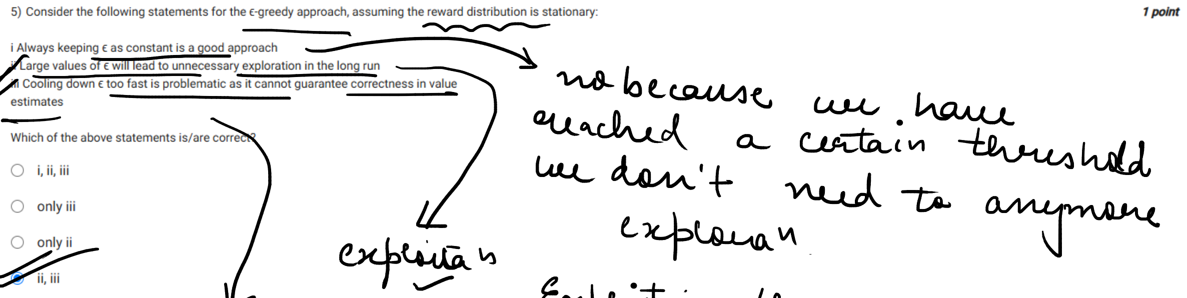
○ $1 - 2^{-t}$

○ $x/20$ if $x < 20$ and 1 after that

○ Same regret for all the above functions



minimizes the loss area.



5) Consider the following statements for the ϵ-greedy approach, assuming the reward distribution is stationary: 1 point

i Always keeping ϵ as constant is a good approach
ii Large values of ϵ will lead to unnecessary exploration in the long run
iii Cooling down ϵ too fast is problematic as it cannot guarantee correctness in value estimates

Which of the above statements is/are correct?

○ i, ii, iii

○ only iii

○ only ii

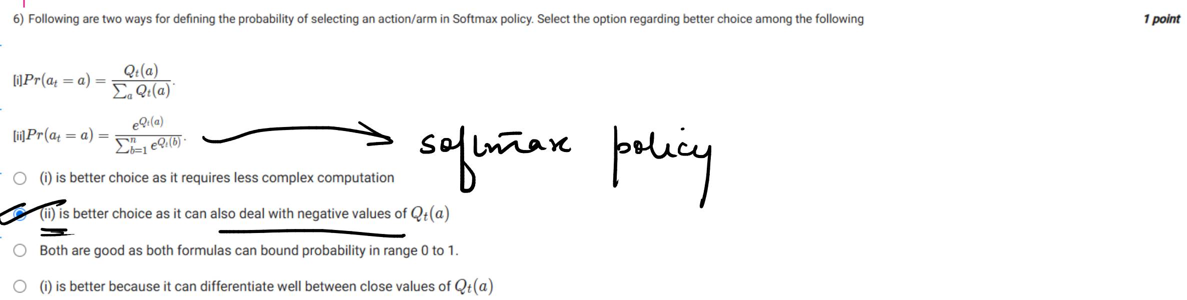
● ii, iii

no because we have reached a certain threshold we don't need to anymore exploraⁿ

exploitaⁿ

Exploiting the ε would be a good move.

you can not start exploitaⁿ too early.



6) Following are two ways for defining the probability of selecting an action/arm in Softmax policy. Select the option regarding better choice among the following 1 point

[i]$Pr(a_t = a) = \frac{Q_t(a)}{\sum_a Q_t(a)}$.

[ii]$Pr(a_t = a) = \frac{e^{Q_t(a)}}{\sum_{b=1}^{n} e^{Q_t(b)}}$. → softmax policy

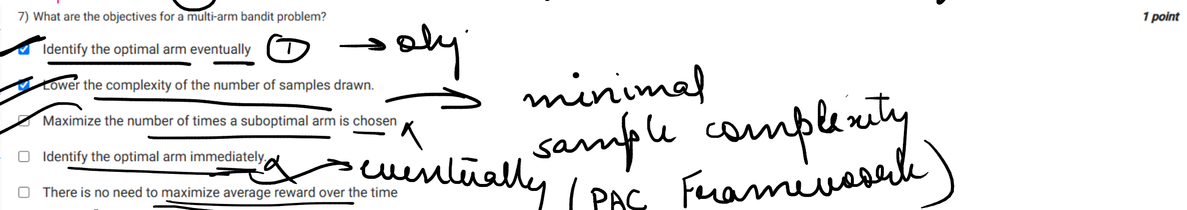
○ (i) is better choice as it requires less complex computation

● (ii) is better choice as it can also deal with negative values of $Q_t(a)$

○ Both are good as both formulas can bound probability in range 0 to 1.

○ (i) is better because it can differentiate well between close values of $Q_t(a)$

① → bottleneck of most of the experiments
② → only b can handle -ve value.
③ → a might return final answer -ve.
④ → a does not provide exploraⁿ if the probabilities are too close.



7) What are the objectives for a multi-arm bandit problem? 1 point

☑ Identify the optimal arm eventually ① → obj.

☑ Lower the complexity of the number of samples drawn. → minimal sample complexity

☐ Maximize the number of times a suboptimal arm is chosen

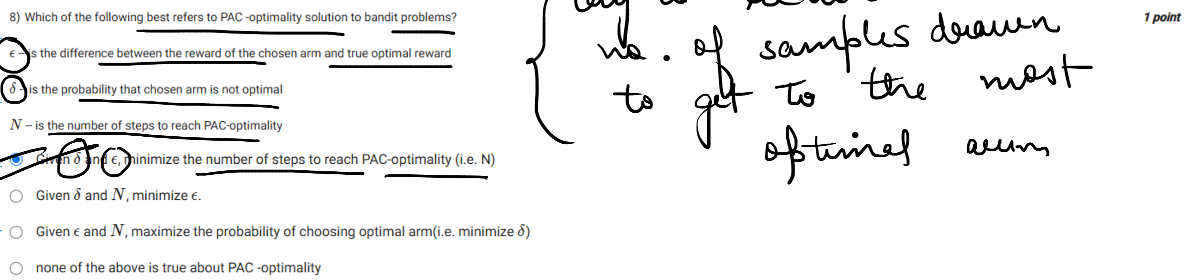
☐ Identify the optimal arm immediately → eventually (PAC Framework)

☐ There is no need to maximize average reward over the time

min

(4 →

min. sample complexity



8) Which of the following best refers to PAC -optimality solution to bandit problems? 1 point

ϵ -- is the difference between the reward of the chosen arm and true optimal reward

δ -- is the probability that chosen arm is not optimal

N -- is the number of steps to reach PAC-optimality

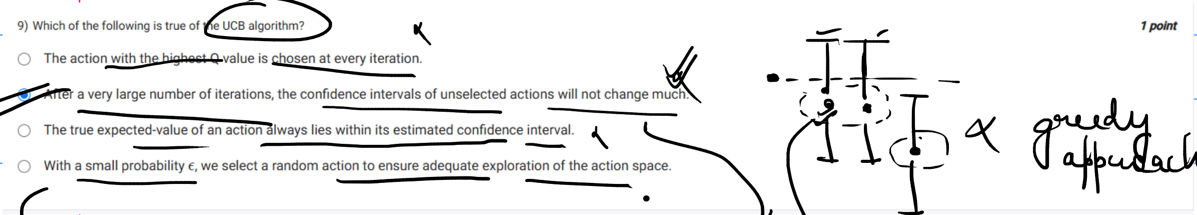
● Given δ and ϵ, minimize the number of steps to reach PAC-optimality (i.e. N)

○ Given δ and N, minimize ϵ.

○ Given ϵ and N, maximize the probability of choosing optimal arm(i.e. minimize δ)

○ none of the above is true about PAC -optimality

try to reduce the no. of samples drawn to get to the most optimal arm regret optimality , confidence intervals



9) Which of the following is true of the UCB algorithm?

1 point

○ The action with the highest Q-value is chosen at every iteration.

● After a very large number of iterations, the confidence intervals of unselected actions will not change much.

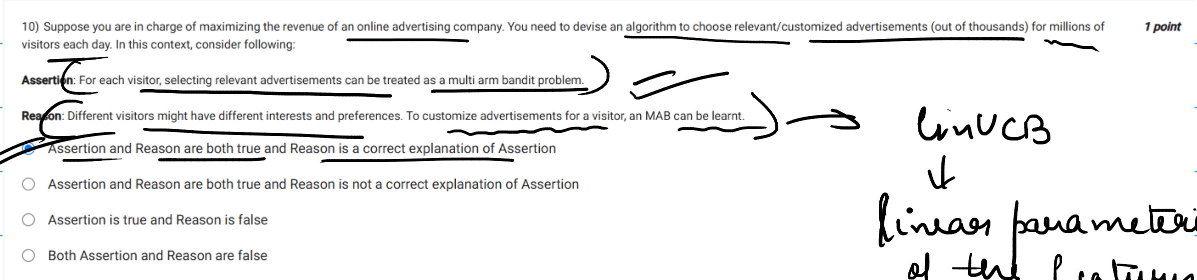
○ The true expected-value of an action always lies within its estimated confidence interval.

○ With a small probability ϵ, we select a random action to ensure adequate exploration of the action space.

× greedy approach

UCB is a deterministic algo.

this arm gets chosen many time interval will shrink.



10) Suppose you are in charge of maximizing the revenue of an online advertising company. You need to devise an algorithm to choose relevant/customized advertisements (out of thousands) for millions of visitors each day. In this context, consider following:

1 point

Assertion: For each visitor, selecting relevant advertisements can be treated as a multi arm bandit problem.

Reason: Different visitors might have different interests and preferences. To customize advertisements for a visitor, an MAB can be learnt.

● Assertion and Reason are both true and Reason is a correct explanation of Assertion

○ Assertion and Reason are both true and Reason is not a correct explanation of Assertion

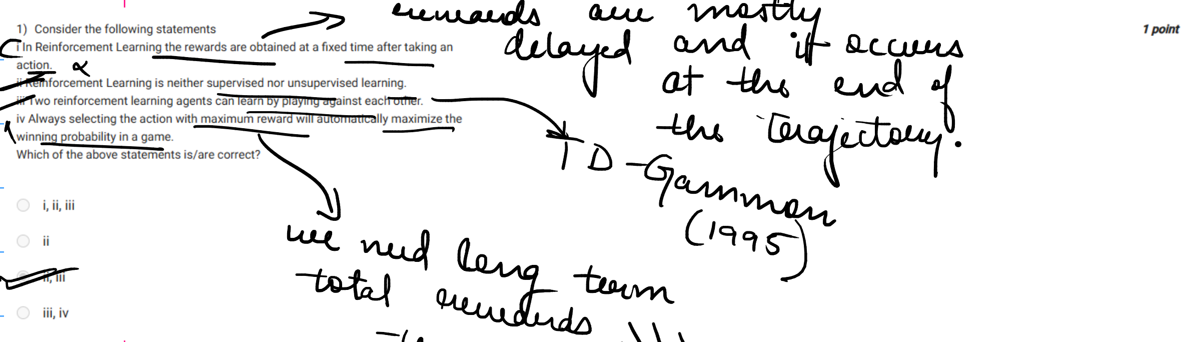
○ Assertion is true and Reason is false

○ Both Assertion and Reason are false

linUCB
↓
linear parameterization of the features

peoples → features → group ⟷ AB train

Graded Assignment - 1



1) Consider the following statements

1 point

i In Reinforcement Learning the rewards are obtained at a fixed time after taking an action. ×

ii Reinforcement Learning is neither supervised nor unsupervised learning.

iii Two reinforcement learning agents can learn by playing against each other.

iv Always selecting the action with maximum reward will automatically maximize the winning probability in a game.

Which of the above statements is/are correct?

○ i, ii, iii

○ ii

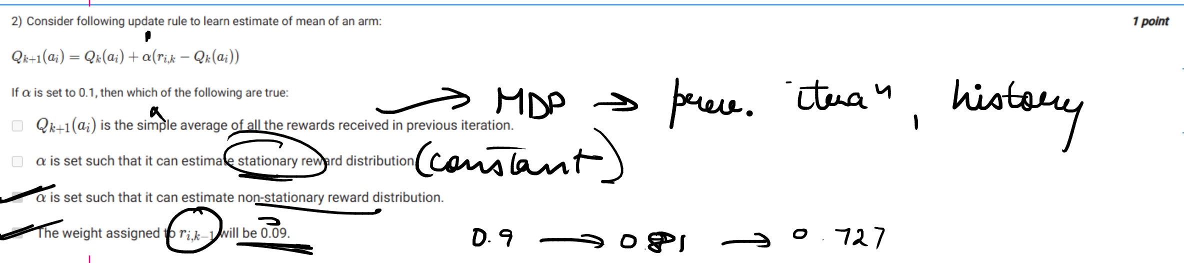
● ii, iii

○ iii, iv

rewards are mostly delayed and it occurs at the end of the trajectory.

TD-Gammon (1995)

we need long term total rewards >>> than the immediate rewards. impt.



2) Consider following update rule to learn estimate of mean of an arm:

1 point

$$Q_{k+1}(a_i) = Q_k(a_i) + \alpha(r_{i,k} - Q_k(a_i))$$

If α is set to 0.1, then which of the following are true:

□ $Q_{k+1}(a_i)$ is the simple average of all the rewards received in previous iteration.

□ α is set such that it can estimate stationary reward distribution. (constant)

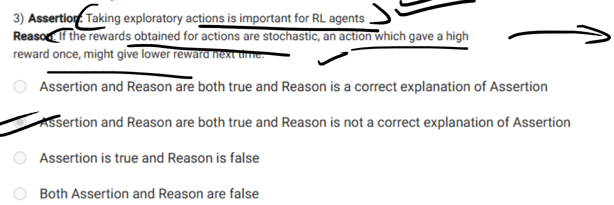
☑ α is set such that it can estimate non-stationary reward distribution.

☑ The weight assigned to $r_{i,k-1}$ will be 0.09.

→ MDP → prev. itera^n , history

0.9 → 0.81 → 0.727

needs to evaluate the estimate of each ac^n correctly



3) **Assertion:** Taking exploratory actions is important for RL agents
Reason: If the rewards obtained for actions are stochastic, an action which gave a high reward once, might give lower reward next time.

○ Assertion and Reason are both true and Reason is a correct explanation of Assertion

● Assertion and Reason are both true and Reason is not a correct explanation of Assertion

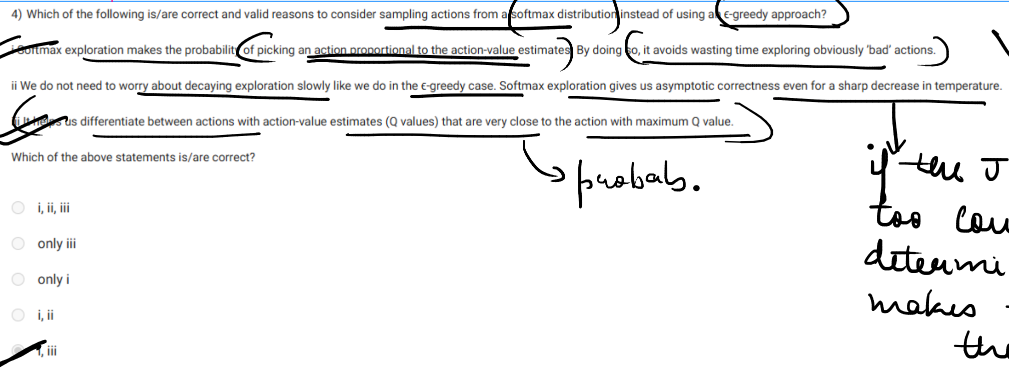
○ Assertion is true and Reason is false

○ Both Assertion and Reason are false

1 point

deterministic

exploraⁿ should continue to search for more advantageous ac^{n}s.



4) Which of the following is/are correct and valid reasons to consider sampling actions from a softmax distribution instead of using an ϵ-greedy approach?

1 point

i Softmax exploration makes the probability of picking an action proportional to the action-value estimates. By doing so, it avoids wasting time exploring obviously 'bad' actions.

ii We do not need to worry about decaying exploration slowly like we do in the ϵ-greedy case. Softmax exploration gives us asymptotic correctness even for a sharp decrease in temperature.

iii It helps us differentiate between actions with action-value estimates (Q values) that are very close to the action with maximum Q value.

Which of the above statements is/are correct?

○ i, ii, iii

○ only iii

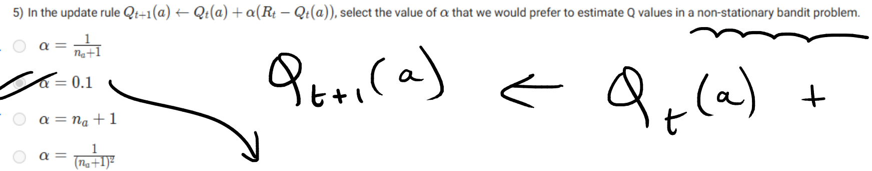
○ only i

○ i, ii

● i, iii

probab.

if the τ becomes too low, the deterministic approach makes the process the greedy.



5) In the update rule $Q_{t+1}(a) \leftarrow Q_t(a) + \alpha(R_t - Q_t(a))$, select the value of α that we would prefer to estimate Q values in a non-stationary bandit problem.

1 point

○ $\alpha = \frac{1}{n_a+1}$

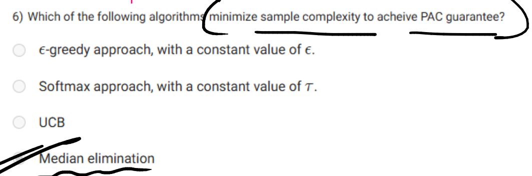
● $\alpha = 0.1$

○ $\alpha = n_a + 1$

○ $\alpha = \frac{1}{(n_a+1)^2}$

$$Q_{t+1}(a) \leftarrow Q_t(a) + \alpha\left(R_t - Q_t(a)\right)$$

constant value of α helps us to, past samples' importance its decreases exponentially, thus, averaging the total rewards.



6) Which of the following algorithms minimize sample complexity to acheive PAC guarantee?

1 point

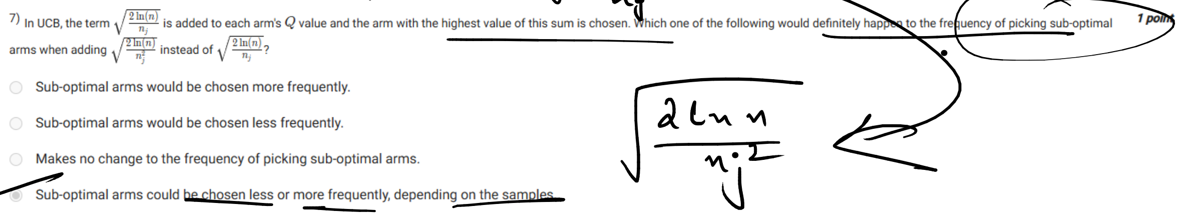
○ ϵ-greedy approach, with a constant value of ϵ.

○ Softmax approach, with a constant value of τ.

○ UCB

● Median elimination

$\max(Q(a_j)) + \sqrt{\frac{2\ln n}{n_j}}$



7) In UCB, the term $\sqrt{\frac{2\ln(n)}{n_j}}$ is added to each arm's Q value and the arm with the highest value of this sum is chosen. Which one of the following would definitely happen to the frequency of picking sub-optimal arms when adding $\sqrt{\frac{2\ln(n)}{n_j^2}}$ instead of $\sqrt{\frac{2\ln(n)}{n_j}}$? *1 point*

○ Sub-optimal arms would be chosen more frequently.

○ Sub-optimal arms would be chosen less frequently.

○ Makes no change to the frequency of picking sub-optimal arms.

● Sub-optimal arms could be chosen less or more frequently, depending on the samples. ✓

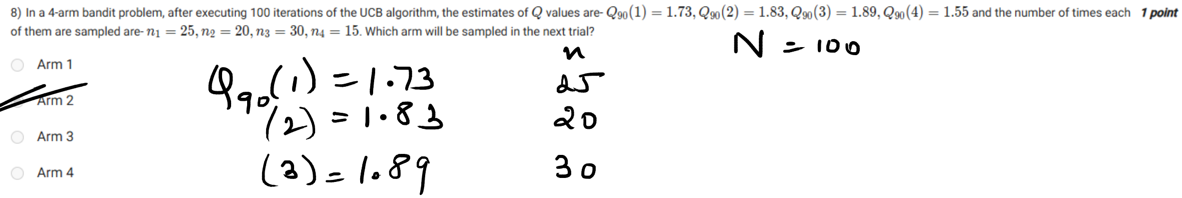
$\sqrt{\frac{2\ln n}{n_j^2}}$

it provides noisy reward distributions

$\frac{n}{n_j} \Rightarrow$ frequency

$\frac{n}{n_j^2}$ ↓ ↑ grow depending on the no of samples provided

shrink



8) In a 4-arm bandit problem, after executing 100 iterations of the UCB algorithm, the estimates of Q values are- $Q_{90}(1) = 1.73, Q_{90}(2) = 1.83, Q_{90}(3) = 1.89, Q_{90}(4) = 1.55$ and the number of times each of them are sampled are- $n_1 = 25, n_2 = 20, n_3 = 30, n_4 = 15$. Which arm will be sampled in the next trial? *1 point*

○ Arm 1

○ Arm 2 ✓

○ Arm 3

○ Arm 4

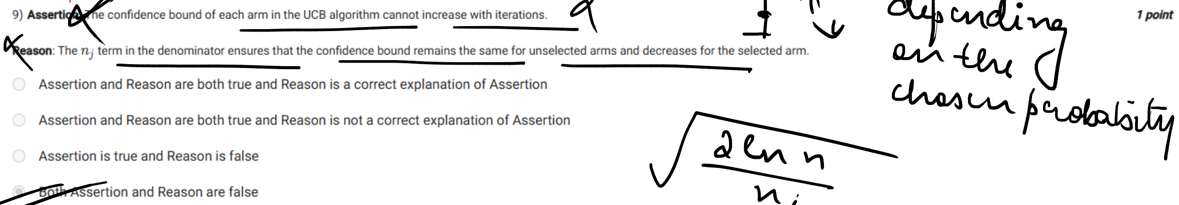
$N = 100$

	n
$Q_{90}(1) = 1.73$	25
$(2) = 1.83$	20
$(3) = 1.89$	30
$(4) = 1.55$	15

$Q_i + \sqrt{\frac{2\ln N}{n_i}}$

① $1.73 + \sqrt{\frac{2\ln 100}{25}} = 2.337$

② 2.509 ③ 2.444 ④ 2.334



9) Assertion: The confidence bound of each arm in the UCB algorithm cannot increase with iterations. *1 point*

Reason: The n_j term in the denominator ensures that the confidence bound remains the same for unselected arms and decreases for the selected arm.

○ Assertion and Reason are both true and Reason is a correct explanation of Assertion

○ Assertion and Reason are both true and Reason is not a correct explanation of Assertion

○ Assertion is true and Reason is false

● Both Assertion and Reason are false ✓

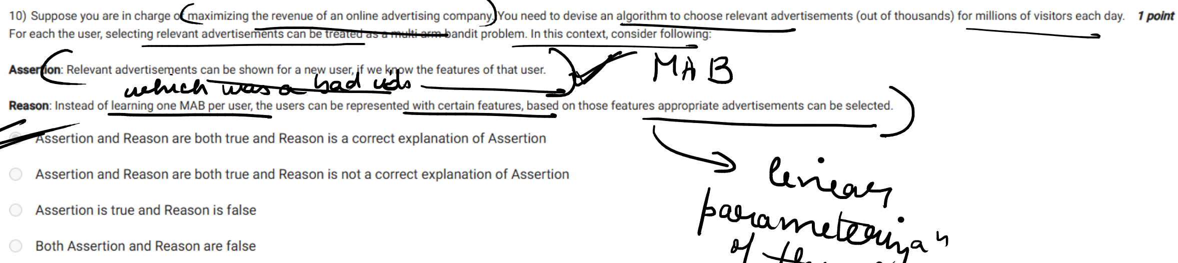
α

I ↑↓ depending on the chosen probability

$\sqrt{\frac{2\ln n}{n_j}}$

selected arm n_j ↑ ✓

unselected arm n_j ↓ ✓



10) Suppose you are in charge of maximizing the revenue of an online advertising company. You need to devise an algorithm to choose relevant advertisements (out of thousands) for millions of visitors each day. *1 point*
For each the user, selecting relevant advertisements can be treated as a multi-arm bandit problem. In this context, consider following:

Assertion: Relevant advertisements can be shown for a new user, if we know the features of that user. ✓ MAB

which was a bad ads

Reason: Instead of learning one MAB per user, the users can be represented with certain features, based on those features appropriate advertisements can be selected.

● Assertion and Reason are both true and Reason is a correct explanation of Assertion ✓

○ Assertion and Reason are both true and Reason is not a correct explanation of Assertion

○ Assertion is true and Reason is false

○ Both Assertion and Reason are false

→ linear parameterization of the features