Quiz and Explanations on Multi-Armed Bandit Algorithms

By Taha Majlesi

July 17, 2025

Contents

1 Multiple Choice Questions (80 Questions)

1. What is the core dilemma that the Multi-Armed Bandit (MAB) problem is designed to study?

   (a) The trade-off between model complexity and accuracy.

   (b) The trade-off between exploration and exploitation.

   (c) The trade-off between computational cost and memory usage.

   (d) The trade-off between supervised and unsupervised learning.

   Correct Answer: (b)

2. In the k-armed bandit problem, what does 'k' represent?

   (a) The number of time steps in an episode.

   (b) The number of agents.

   (c) The number of available actions or "arms".

   (d) The maximum possible reward.

   Correct Answer: (c)

3. The MAB problem is formally equivalent to a:

   (a) Multi-state Markov Decision Process (MDP).

   (b) One-state Markov Decision Process (MDP).

   (c) Hidden Markov Model (HMM).

   (d) Partially Observable Markov Decision Process (POMDP).

   Correct Answer: (b)

4. What is the "true action-value," $q_*(a)$?

   (a) The most recent reward received for action a.

   (b) The agent's current estimate of the value of action a.

   (c) The total reward accumulated so far.

   (d) The expected or mean reward for selecting action a.

   Correct Answer: (d)

5. The agent's primary objective in the MAB problem is to:

   (a) Try every action an equal number of times.

   (b) Maximize the total cumulative reward over time.

   (c) Minimize the number of exploratory actions.

   (d) Discover the reward distribution of every arm.

   Correct Answer: (b)

6. "Exploitation" in the context of MAB refers to:

   (a) Choosing an action randomly to gather information.
   (b) Choosing the action with the highest estimated value.
   (c) Updating the value estimate of an action.
   (d) Choosing an action that has never been tried before.

   Correct Answer: (b)

7. "Exploration" in the context of MAB refers to:

   (a) Choosing the action with the highest estimated value.
   (b) Choosing an action to improve the accuracy of value estimates.
   (c) Always sticking with the current best action.
   (d) Ignoring new information.

   Correct Answer: (b)

8. Which of the following is a common application of MABs?

   (a) Image classification.
   (b) Natural language translation.
   (c) Online advertising and recommender systems.
   (d) Audio signal processing.

   Correct Answer: (c)

9. The sample-average method estimates an action's value by:

   (a) Using only the most recent reward.
   (b) Averaging all rewards received for that action.
   (c) Using a complex probabilistic model.
   (d) Setting it to an optimistic initial value.

   Correct Answer: (b)

10. According to the law of large numbers, the sample-average estimate $Q_t(a)$ is guaranteed to converge to the true value $q_*(a)$ as:

    (a) The number of time steps t approaches infinity.
    (b) The number of times action a is chosen, $N_t(a)$, approaches infinity.
    (c) The learning rate approaches zero.
    (d) The number of arms k increases.

    Correct Answer: (b)

11. What is a major limitation of the non-incremental sample-average method?

    (a) It is statistically biased.

(b) It cannot be used for more than two arms.

(c) It requires storing all past rewards, which is memory-intensive.

(d) It never converges to the true value.

Correct Answer: (c)

12. The incremental update rule is given by $Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$. What does the term $(R_n - Q_n)$ represent?

(a) The new estimate.

(b) The step size.

(c) The prediction error.

(d) The cumulative reward.

Correct Answer: (c)

13. In the general learning rule, 'NewEstimate ¡- OldEstimate + StepSize(Target - OldEstimate)', what is the 'Target' for the basic bandit problem?

(a) The previous estimate, Q_n.

(b) The most recent reward, R_n.

(c) The step-size parameter, α.

(d) The true action-value, $q_*(a)$.

Correct Answer: (b)

14. A problem where the true action-values change over time is called:

(a) A stationary problem.

(b) A non-stationary problem.

(c) A contextual problem.

(d) A Bernoulli problem.

Correct Answer: (b)

15. Why is the sample-average method with a $\frac{1}{n}$ step size inadequate for non-stationary problems?

(a) The step size becomes too large over time.

(b) The step size decreases, preventing the agent from adapting to new changes.

(c) It only works for binary rewards.

(d) It requires knowledge of the true action-values.

Correct Answer: (b)

16. To handle non-stationary environments, the step-size parameter is often changed from $\frac{1}{n}$ to:

(a) A larger value, like n.

(b) A constant value, α.

(c) A value of zero.

(d) A randomly chosen value.

Correct Answer: (b)

17. Using a constant step-size α results in an estimate that is a(n):

(a) Simple average of all past rewards.

(b) Exponential recency-weighted average of past rewards.

(c) Median of all past rewards.

(d) Maximum of all past rewards.

Correct Answer: (b)

18. In the ϵ-greedy strategy, what happens with probability ϵ?

(a) The agent exploits.

(b) The agent explores by choosing a random action.

(c) The agent updates its value estimates.

(d) The agent terminates the process.

Correct Answer: (b)

19. A purely greedy agent corresponds to an ϵ-greedy strategy with:

(a) $\epsilon = 1$.

(b) $\epsilon = 0.5$.

(c) $\epsilon = 0$.

(d) $\epsilon = 0.1$.

Correct Answer: (c)

20. What is the main disadvantage of the ϵ-greedy strategy's exploration?

(a) It is too aggressive.

(b) It stops exploring after a fixed number of steps.

(c) It is undirected, meaning it might pick a known bad arm.

(d) It is computationally very expensive.

Correct Answer: (c)

21. The "optimistic initial values" method encourages exploration by:

(a) Initializing all action-value estimates to a very high value.

(b) Initializing all action-value estimates to zero.

(c) Using a very high value for ϵ.

(d) Randomly initializing the value estimates.

Correct Answer: (a)

22. How does an agent with optimistic initial values and a greedy policy behave initially?

    (a) It sticks to the first arm it chooses.

    (b) It is forced to try every arm at least once.

    (c) It behaves identically to an ϵ-greedy agent.

    (d) It chooses arms completely at random.

    Correct Answer: (b)

23. Optimistic initial values are primarily effective for:

    (a) Sustained exploration in non-stationary problems.

    (b) Initial exploration in stationary problems.

    (c) Problems with a very large number of arms.

    (d) Contextual bandit problems.

    Correct Answer: (b)

24. The Upper Confidence Bound (UCB) algorithm is based on the principle of:

    (a) Pessimism in the face of uncertainty.

    (b) Optimism in the face of uncertainty.

    (c) Randomness in the face of uncertainty.

    (d) Greediness in the face of uncertainty.

    Correct Answer: (b)

25. In the UCB formula $A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$, what does the term $c\sqrt{\frac{\ln t}{N_t(a)}}$ represent?

    (a) The estimated value.

    (b) The uncertainty bonus or exploration term.

    (c) The learning rate.

    (d) The prediction error.

    Correct Answer: (b)

26. In the UCB uncertainty bonus, what is the role of $N_t(a)$?

    (a) A larger $N_t(a)$ increases the bonus, encouraging exploration.

    (b) A smaller $N_t(a)$ increases the bonus, encouraging exploration of less-tried arms.

    (c) $N_t(a)$ has no effect on the bonus.

    (d) A smaller $N_t(a)$ decreases the bonus.

Correct Answer: (b)

27. What is the key difference between UCB's exploration and ϵ-greedy's exploration?

    (a) UCB's exploration is random, while ϵ-greedy's is directed.

    (b) UCB's exploration is directed, while ϵ-greedy's is random/undirected.

    (c) UCB does not explore.

    (d) ϵ-greedy does not explore.

    Correct Answer: (b)

28. What distinguishes a contextual bandit from a simple MAB?

    (a) Contextual bandits have more arms.

    (b) Contextual bandits receive side information (a "context") before choosing an action.

    (c) Contextual bandits have deterministic rewards.

    (d) Contextual bandits do not involve an exploration-exploitation trade-off.

    Correct Answer: (b)

29. In a contextual bandit, the goal is to learn a policy that maps:

    (a) Actions to a single best reward.

    (b) Time steps to actions.

    (c) Contexts to the best actions.

    (d) Agents to contexts.

    Correct Answer: (c)

30. The LinUCB algorithm assumes that the expected reward is a(n) _____ function of the context features.

    (a) Exponential.

    (b) Logarithmic.

    (c) Quadratic.

    (d) Linear.

    Correct Answer: (d)

31. In LinUCB, what method is used to estimate the unknown coefficient vector θ_a?

    (a) Gradient Descent.

    (b) K-Means Clustering.

    (c) Ridge Regression.

    (d) Principal Component Analysis (PCA).

    Correct Answer: (c)

32. What is the purpose of the identity matrix I_d in the LinUCB estimation formula $\hat{\theta}_a = (D_a^T D_a + I_d)^{-1} D_a^T b_a$?

    (a) To increase the learning speed.

    (b) To act as a regularization term to prevent overfitting.

    (c) To represent the context features.

    (d) To ensure the rewards are positive.

    Correct Answer: (b)

33. In LinUCB's selection rule, exploration is driven by:

    (a) The number of times an arm has been pulled, $N_t(a)$.

    (b) A fixed probability ϵ.

    (c) The uncertainty of the reward prediction for the current context.

    (d) The magnitude of the last reward.

    Correct Answer: (c)

34. How does Thompson Sampling represent its belief about an arm's quality?

    (a) As a single point estimate, $Q_t(a)$.

    (b) As a full probability distribution (a posterior).

    (c) As the highest reward seen so far.

    (d) As a constant value.

    Correct Answer: (b)

35. How does Thompson Sampling select an action?

    (a) It chooses the arm with the highest mean in its posterior distribution.

    (b) It draws a random sample from each arm's posterior and selects the arm with the highest sample.

    (c) It chooses an arm randomly with probability ϵ.

    (d) It chooses the arm with the widest posterior distribution.

    Correct Answer: (b)

36. In the context of Thompson Sampling, what is a conjugate prior?

    (a) A prior that is uniform over all possibilities.

    (b) A prior distribution such that the posterior distribution is in the same family.

    (c) A prior that is guaranteed to be incorrect.

    (d) A prior that is based on a linear model.

    Correct Answer: (b)

37. For a Bernoulli bandit problem (binary rewards), what is the conjugate prior for the success probability parameter?

    (a) The Gaussian distribution.

    (b) The Poisson distribution.

    (c) The Beta distribution.

    (d) The Uniform distribution.

    Correct Answer: (c)

38. In a Beta-Bernoulli model for Thompson Sampling, if the prior is Beta(α, β) and a success (reward=1) is observed, the posterior becomes:

    (a) Beta($\alpha, \beta + 1$).

    (b) Beta($\alpha + 1, \beta + 1$).

    (c) Beta(α, β).

    (d) Beta($\alpha + 1, \beta$).

    Correct Answer: (d)

39. The parameters α and β of a Beta distribution can be intuitively interpreted as:

    (a) The mean and variance.

    (b) The number of prior successes and failures.

    (c) The learning rate and exploration constant.

    (d) The context vector and reward vector.

    Correct Answer: (b)

40. What is a key advantage of Thompson Sampling over UCB, often observed empirically?

    (a) It is simpler to implement.

    (b) It often demonstrates superior performance and lower regret.

    (c) It does not require any memory.

    (d) It has stronger theoretical guarantees.

    Correct Answer: (b)

41. A purely greedy agent ($\epsilon = 0$) often performs poorly because:

    (a) It explores too much.

    (b) It can get stuck on a suboptimal action discovered early.

    (c) It is computationally unstable.

    (d) It requires too much memory.

    Correct Answer: (b)

42. Which algorithm surgically directs exploration toward actions where the value estimate is least reliable?

    (a) ϵ-greedy.

    (b) A purely greedy agent.

    (c) UCB.

    (d) A random agent.

    Correct Answer: (c)

43. LinUCB will likely outperform a non-contextual algorithm like UCB only if:

    (a) The number of arms is small.

    (b) The rewards are stationary.

    (c) The provided context is genuinely predictive of rewards.

    (d) The rewards are binary.

    Correct Answer: (c)

44. The update rule $Q_{n+1} = Q_n + \alpha(R_n - Q_n)$ is most suitable for:

    (a) Stationary problems where convergence to the true mean is desired.

    (b) Non-stationary problems where adaptability is key.

    (c) Problems where rewards are always positive.

    (d) Contextual bandit problems.

    Correct Answer: (b)

45. In the LinUCB update step, which quantities are updated after an arm is chosen?

    (a) The matrices A_a and vectors b_a for all arms.

    (b) The matrices A_a and vectors b_a for only the chosen arm.

    (c) Only the exploration parameter α.

    (d) Only the context vector $x_{t,a}$.

    Correct Answer: (b)

46. The initial prior Beta(1, 1) in Thompson Sampling corresponds to what distribution?

    (a) A Gaussian distribution.

    (b) A distribution heavily skewed towards 0.

    (c) A distribution heavily skewed towards 1.

    (d) A uniform distribution.

    Correct Answer: (d)

47. The term "regret" in bandit literature refers to:

(a) The difference between the optimal reward and the reward obtained by the agent.

(b) The number of times the agent explored.

(c) The computational time of the algorithm.

(d) The memory used by the algorithm.

Correct Answer: (a)

48. Which algorithm does not require an explicit exploration parameter like ϵ or c?

(a) ϵ-greedy.

(b) UCB.

(c) Thompson Sampling.

(d) LinUCB.

Correct Answer: (c)

49. The complexity of LinUCB is sensitive to:

(a) The number of time steps t.

(b) The dimensionality of the context feature vector, d.

(c) The maximum possible reward.

(d) The value of ϵ.

Correct Answer: (b)

50. What is the "Target" in the incremental update rule for a non-stationary problem?

(a) The previous estimate Q_n.

(b) The step-size α.

(c) The recent reward R_n.

(d) The initial value Q_1.

Correct Answer: (c)

51. The main advantage of ϵ-greedy is its:

(a) High performance.

(b) Simplicity.

(c) Directed exploration.

(d) Suitability for contextual problems.

Correct Answer: (b)

52. In the UCB formula, what is the purpose of the $\ln t$ term in the numerator?

(a) It decreases the exploration bonus over time.

(b) It ensures that the bonus term eventually goes to zero.

(c) It ensures that exploration continues, as it increases with time, guaranteeing all arms are eventually revisited.

(d) It is a regularization term.

Correct Answer: (c)

53. If a conjugate prior is not available for a Thompson Sampling problem, what is the consequence?

(a) The algorithm cannot be used.

(b) The posterior update requires more complex approximation techniques like MCMC.

(c) The algorithm becomes equivalent to UCB.

(d) The rewards must be binary.

Correct Answer: (b)

54. In LinUCB, the matrix $A_a = D_a^T D_a + I_d$ is proportional to the:

(a) Inverse of the covariance matrix of the parameter estimate $\hat{\theta}_a$.

(b) The reward vector b_a.

(c) The context vector $x_{t,a}$.

(d) The exploration parameter α.

Correct Answer: (a)

55. A wider posterior distribution in Thompson Sampling indicates:

(a) High certainty about the arm's value.

(b) High uncertainty about the arm's value.

(c) That the arm is optimal.

(d) That the arm is suboptimal.

Correct Answer: (b)

56. Which of these is NOT part of the formal definition of a k-armed bandit problem?

(a) An agent.

(b) A set of k actions.

(c) A set of reward distributions.

(d) A set of states with transition probabilities.

Correct Answer: (d)

57. The "optimism in the face of uncertainty" principle suggests:

(a) If we are uncertain about an action's value, we should assume it is low.

(b) If we are uncertain about an action's value, we should assume it is high to encourage exploration.

(c) We should always be optimistic about the total reward.

(d) We should ignore uncertainty.

Correct Answer: (b)

58. In a Beta-Bernoulli model, if the prior is Beta(1, 1) and a failure (reward=0) is observed, the posterior becomes:

(a) Beta(2, 1).

(b) Beta(1, 2).

(c) Beta(2, 2).

(d) Beta(1, 1).

Correct Answer: (b)

59. The choice between a decreasing step size ($\frac{1}{n}$) and a constant step size (α) reflects an assumption about:

(a) The number of arms.

(b) The stability (stationarity) of the problem environment.

(c) The availability of context.

(d) The type of reward (binary or continuous).

Correct Answer: (b)

60. In the LinUCB selection rule, the term $x_{t,a}^T \hat{\theta}_a$ represents:

(a) The uncertainty of the prediction.

(b) The exploitation component (the predicted reward).

(c) The regularization term.

(d) The context vector for a different arm.

Correct Answer: (b)

61. Thompson Sampling naturally balances exploration and exploitation through:

(a) An explicit exploration parameter ϵ.

(b) An explicit uncertainty bonus term.

(c) The process of sampling from posterior distributions.

(d) Initializing values optimistically.

Correct Answer: (c)

62. What is a primary disadvantage of UCB?

(a) It is too simple.

(b) It can be overly conservative and its performance is sensitive to the choice of 'c'.

(c) It cannot handle non-stationary problems.

(d) It does not explore enough.

Correct Answer: (b)

63. The incremental update rule is an instance of a general learning principle also found in:

(a) K-Means clustering.

(b) Temporal-Difference (TD) learning and stochastic gradient descent.

(c) Decision tree construction.

(d) Support Vector Machines.

Correct Answer: (b)

64. In the context of clinical trials as a bandit problem, what does an "arm" represent?

(a) A patient.

(b) A hospital.

(c) A medical treatment.

(d) A research scientist.

Correct Answer: (c)

65. If an arm in LinUCB has been pulled many times, but for contexts very different from the current one, the uncertainty term will be:

(a) Large, encouraging exploration.

(b) Small, discouraging exploration.

(c) Zero.

(d) Unchanged.

Correct Answer: (a)

66. The "Bayesian Heuristic" refers to:

(a) Always choosing the action with the highest prior probability.

(b) Representing beliefs as probability distributions and updating them with data.

(c) Using linear models for all problems.

(d) A method for setting the ϵ parameter.

Correct Answer: (b)

67. Which algorithm's effectiveness is highly sensitive to the quality of feature engineering?

(a) ϵ-greedy.

(b) UCB.

(c) LinUCB.

(d) Thompson Sampling (non-contextual).

Correct Answer: (c)

68. The optimal action a_* is the action with the:

(a) Highest estimated value $Q_t(a)$.

(b) Highest true action-value $q_*(a)$.

(c) Lowest variance in rewards.

(d) Most uncertainty.

Correct Answer: (b)

69. What is the primary reason for using an incremental update rule?

(a) It is more accurate.

(b) It is more computationally and memory efficient.

(c) It works better for contextual bandits.

(d) It guarantees faster convergence.

Correct Answer: (b)

70. In the ϵ-greedy algorithm, when exploring, how is an action chosen?

(a) From the set of unexplored arms.

(b) From the set of all k actions with equal probability.

(c) Based on the lowest estimated value.

(d) Based on the UCB formula.

Correct Answer: (b)

71. The LinUCB algorithm is a combination of the UCB principle and:

(a) A Bayesian model.

(b) A linear model.

(c) A deep neural network.

(d) A random forest.

Correct Answer: (b)

72. In Thompson Sampling, a narrow posterior distribution for an arm implies:

(a) The agent is uncertain about the arm's value.

(b) The agent is certain about the arm's value.

(c) The arm has never been chosen.

(d) The arm is the optimal one.

Correct Answer: (b)

73. Which algorithm is considered a good, simple baseline to compare against?

    (a) LinUCB.

    (b) Thompson Sampling.

    (c) UCB.

    (d) ϵ-greedy.

    Correct Answer: (d)

74. The term A_a^{-1} in the LinUCB formula is proportional to the:

    (a) Mean of the parameter estimate.

    (b) Covariance matrix of the parameter estimate $\hat{\theta}_a$.

    (c) Reward vector.

    (d) Learning rate.

    Correct Answer: (b)

75. The main challenge in the MAB problem stems from the fact that:

    (a) The number of arms is too large.

    (b) The rewards are always zero.

    (c) The true action-values $q_*(a)$ are initially unknown.

    (d) The agent has a limited number of time steps.

    Correct Answer: (c)

76. Which algorithm explicitly adds a bonus for uncertainty to the estimated value?

    (a) ϵ-greedy.

    (b) Thompson Sampling.

    (c) UCB.

    (d) Greedy algorithm.

    Correct Answer: (c)

77. The use of a constant step-size parameter α is particularly useful for giving more weight to:

    (a) Initial rewards.

    (b) All rewards equally.

    (c) Recent rewards.

    (d) Rewards from the best arm.

    Correct Answer: (c)

2 Explainable Questions (30 Questions)

1. **Question:** Explain the exploration-exploitation dilemma in your own words, using a real-world example other than slot machines (e.g., choosing a restaurant).

2. **Question:** What is the difference between the true action-value $q_*(a)$ and the estimated action-value $Q_t(a)$? Why is this distinction fundamental to the bandit problem?

3. **Question:** Derive the incremental update rule $Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$ starting from the basic definition of the sample-average method.

4. **Question:** Explain why a constant step-size parameter α is more suitable for non-stationary problems than a decreasing step-size like $\frac{1}{n}$. What does the resulting estimate represent?

5. **Question:** Describe the ϵ-greedy algorithm. What is its main advantage and its primary disadvantage?

6. **Question:** How does the "optimistic initial values" method work? Explain the mechanism that encourages exploration.

7. **Question:** Explain the Upper Confidence Bound (UCB) action selection rule. Break down its formula and describe the role of both the exploitation and exploration terms.

8. **Question:** Compare and contrast the exploration strategies of ϵ-greedy and UCB. Why is UCB's exploration considered more "directed"?

9. **Question:** What is a contextual bandit? How does it differ from a standard multi-armed bandit, and what new capability does it introduce?

10. **Question:** What is the core assumption of the LinUCB algorithm regarding the relationship between context and rewards?

11. **Question:** In the LinUCB algorithm, explain the role of the matrices A_a and the vectors b_a. How are they updated?

12. **Question:** Explain the exploration term in the LinUCB selection rule. How does it use the current context to guide exploration more efficiently than the standard UCB?

13. **Question:** Describe the core idea behind Thompson Sampling. How does its Bayesian approach to action selection differ from the frequentist approach of UCB?

14. **Question:** What is a conjugate prior, and why is it so useful for implementing Thompson Sampling? Provide the example of the Beta-Bernoulli model.

15. **Question:** Walk through one full cycle (one time step) of the Thompson Sampling algorithm for a Bernoulli bandit problem.

16. **Question:** Why is a purely greedy strategy (i.e., $\epsilon = 0$) often a poor choice for a multi-armed bandit problem?

17. **Question:** What is the general learning rule 'NewEstimate ¡- OldEstimate + StepSize(Target - OldEstimate)'? Identify each component for the standard incremental update rule for bandit problems.

18. **Question:** If you were designing a news article recommender system for a website, would you choose a standard MAB or a contextual bandit? Justify your choice.

19. **Question:** Explain the concept of an "exponential recency-weighted average" and why it is desirable in non-stationary environments.

20. **Question:** What is the role of the parameter c in the UCB algorithm? What would happen if c were very large or very small (close to zero)?

21. **Question:** What is the role of the regularization term (I_d) in the ridge regression formula used by LinUCB?

22. **Question:** Can Thompson Sampling be used if a conjugate prior is not available for the reward model? If so, what is the challenge?

23. **Question:** Compare the practical advantages and disadvantages of UCB and Thompson Sampling. When might you prefer one over the other?

24. **Question:** Explain how the principle of "optimism in the face of uncertainty" is implemented in both the Optimistic Initial Values method and the UCB algorithm.

25. **Question:** In the LinUCB algorithm, what does the term $x_{t,a}^T A_a^{-1} x_{t,a}$ represent intuitively?

26. **Question:** How does Thompson Sampling's mechanism naturally balance exploration and exploitation without an explicit parameter for it?

27. **Question:** Describe two different real-world applications of multi-armed bandits, specifying what the "arms" and "rewards" would be in each case.

28. **Question:** What is the key difference in the update step of LinUCB compared to a non-contextual algorithm like UCB?

29. **Question:** Why is the MAB problem considered a simplified, yet powerful, reinforcement learning model? What key RL element is simplified or removed?

30. **Question:** You are given a choice between ϵ-greedy, UCB, and Thompson Sampling for a new stationary bandit problem. Which would you likely choose for the best performance and why?

3 Answers to Explainable Questions

1. **Answer:** The exploration-exploitation dilemma is the challenge of choosing between what you know works and trying something new that might be better. **Example:** When choosing a restaurant for dinner, **exploitation** is going to your favorite Italian place that you know is delicious. You are guaranteed a good meal. **Exploration** is trying the new Thai restaurant that just opened down the street. It's a risk—it could be amazing and become your new favorite, or it could be terrible, and you'll have wasted your money and evening. The dilemma is balancing the safety of a known good option against the potential long-term benefit of discovering an even better one.

2. **Answer:**

   - $q_*(a)$ **(True Action-Value):** This is the theoretical, true mean reward of an action a. It's a fixed (in stationary problems) but unknown property of the environment. It's what the agent is trying to discover.
   - $Q_t(a)$ **(Estimated Action-Value):** This is the agent's current guess about the true value of action a at time step t. It is calculated based on the rewards the agent has observed so far.

   This distinction is fundamental because the entire problem exists in the gap between them. The agent only has access to $Q_t(a)$ to make decisions, but its goal is to find the action with the highest $q_*(a)$. The challenge is to make $Q_t(a)$ a good enough approximation of $q_*(a)$ to make the right choice.

3. **Answer:**

   (a) The sample-average estimate after n rewards for an action is $Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i$.

   (b) We can separate the most recent reward, R_n: $Q_{n+1} = \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right)$.

   (c) The previous estimate, Q_n, was the average of the first $n-1$ rewards: $Q_n = \frac{1}{n-1}\sum_{i=1}^{n-1} R_i$. This means $\sum_{i=1}^{n-1} R_i = (n-1)Q_n$.

   (d) Substitute this back into the equation for Q_{n+1}: $Q_{n+1} = \frac{1}{n}\left(R_n + (n-1)Q_n\right)$.

   (e) Distribute the $\frac{1}{n}$: $Q_{n+1} = \frac{1}{n}R_n + \frac{n-1}{n}Q_n = \frac{1}{n}R_n + \left(1 - \frac{1}{n}\right)Q_n$.

   (f) Rearrange the terms: $Q_{n+1} = Q_n - \frac{1}{n}Q_n + \frac{1}{n}R_n$.

   (g) Factor out $\frac{1}{n}$: $Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$.

4. **Answer:** In a non-stationary problem, the true action-values $q_*(a)$ change over time.

   - A decreasing step-size $\frac{1}{n}$ becomes very small as time goes on. This means new rewards have very little impact on the estimate, effectively "freezing" the learning process. The agent gives equal weight to very old rewards and recent rewards, so it cannot adapt if the underlying reward distribution changes.
   - A constant step-size α ensures that the agent always gives a significant, fixed weight to recent rewards. This allows the estimate to "forget" old information and continuously track changes in the environment.

The resulting estimate is an **exponential recency-weighted average**, where the influence of past rewards decays exponentially the older they are.

5. **Answer:** The ϵ-greedy algorithm is a simple action-selection policy. With a high probability $(1 - \epsilon)$, it acts greedily by choosing the arm with the highest current estimated value (exploitation). With a small probability (ϵ), it ignores the estimates and chooses an arm completely at random from all available arms (exploration).

   - **Main Advantage:** It is extremely simple to understand and implement, making it a great baseline.
   - **Primary Disadvantage:** Its exploration is "undirected" or "blind." When it explores, it is just as likely to pick a known bad arm as it is to pick a promising but uncertain arm. This is an inefficient way to gather information.

6. **Answer:** The optimistic initial values method works by setting the initial value estimates for all actions, $Q_1(a)$, to a value that is known to be higher than any possible reward (e.g., if rewards are between 0 and 1, initialize $Q_1(a) = 5$ for all a). The agent then acts greedily. The mechanism is as follows: The agent picks an arm. The reward it receives will be lower than the optimistic initial value. The arm's value estimate is then updated to a more realistic, lower number. Now, all other unexplored arms still have the high optimistic value, so the greedy agent is forced to choose one of them next. This process repeats, encouraging the agent to try every single arm at least once before it starts to settle on the truly best one.

7. **Answer:** The Upper Confidence Bound (UCB) action selection rule chooses the action that maximizes an upper confidence bound on its true value. The formula is: $A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$.

   - **Exploitation Term ($Q_t(a)$):** This is the current estimated value of the action. It favors arms that have performed well in the past.
   - **Exploration Term ($c\sqrt{\frac{\ln t}{N_t(a)}}$):** This is the uncertainty bonus. It is large for arms that have been tried infrequently ($N_t(a)$ is small) or when a lot of time has passed ($\ln t$ is large). This term encourages the agent to try arms whose values it is uncertain about. The parameter c controls the weight of this exploration bonus.

8. **Answer:**

   - **ϵ-greedy exploration:** Is **undirected**. With probability ϵ, it picks an action uniformly at random from all actions, regardless of their history or potential.
   - **UCB exploration:** Is **directed**. It doesn't explore randomly. Instead, it systematically favors actions for which its value estimate is most uncertain (i.e., $N_t(a)$ is small relative to t). It surgically directs exploration to where it is most needed to reduce uncertainty, making it far more efficient at finding the optimal arm.

9. **Answer:** A contextual bandit is an extension of the MAB problem where, before making a choice, the agent receives some side information called a "context" (represented as a feature vector x_t).

- **Difference:** In a standard MAB, the best action is assumed to be the same in all situations. In a contextual bandit, the best action can change depending on the context.
- **New Capability:** This introduces the ability for **personalization**. The agent learns a policy that maps specific contexts to the best actions, rather than learning a single best action for all situations.

10. **Answer:** The core assumption of the LinUCB algorithm is the **linear payoff assumption**. It assumes that the expected reward of an arm a is a linear function of its context feature vector $x_{t,a}$. This is expressed as $\mathbb{E}[r_{t,a}|x_{t,a}] = x_{t,a}^T \theta_{a,*}$, where $\theta_{a,*}$ is an unknown weight vector that the algorithm needs to learn for each arm.

11. **Answer:** For each arm a:

    - A_a**:** Is a $d \times d$ matrix (where d is the feature dimension) that accumulates information about the contexts seen for arm a. It is initialized to the identity matrix I_d.
    - b_a**:** Is a d-dimensional vector that accumulates the rewards for arm a, weighted by the context in which they were received. It is initialized to a zero vector.

    Update Rule: When an arm a_t is chosen and reward r_t is observed in context x_{t,a_t}, only the matrix and vector for that specific arm are updated:

    - $A_{a_t} \leftarrow A_{a_t} + x_{t,a_t} x_{t,a_t}^T$
    - $b_{a_t} \leftarrow b_{a_t} + r_t x_{t,a_t}$

12. **Answer:** The exploration term in LinUCB is $\alpha\sqrt{x_{t,a}^T A_a^{-1} x_{t,a}}$. This term represents the standard deviation (a measure of uncertainty) of the reward prediction for the *current context* $x_{t,a}$. Unlike standard UCB, which only considers how many times an arm has been pulled ($N_t(a)$), LinUCB considers how similar the current context is to past contexts where the arm was pulled. If an arm has been pulled many times but in very different contexts, the uncertainty for this new context will still be high, encouraging exploration. This makes exploration far more targeted and efficient in a high-dimensional feature space.

13. **Answer:**

    - **Core Idea:** Thompson Sampling (TS) is a Bayesian algorithm that maintains a full probability distribution (a posterior) representing its belief about the value of each arm, instead of just a single point estimate.
    - **Difference:** UCB takes a frequentist approach, calculating a point estimate ($Q_t(a)$) and adding a deterministic confidence bound. TS takes a Bayesian approach. For action selection, it embraces the uncertainty fully by sampling a plausible value from each arm's belief distribution and acting greedily with respect to those samples. This probabilistic selection naturally handles the exploration-exploitation trade-off.

14. **Answer:** A **conjugate prior** is a choice of prior distribution for a parameter such that when it is updated with new data (via the likelihood), the resulting

posterior distribution belongs to the same family of distributions. **Usefulness:** This is a huge computational convenience. It means the Bayesian update step can be performed with a simple algebraic update to the distribution's parameters, rather than requiring complex and slow numerical integration (like MCMC). **Example:** For a Bernoulli likelihood (success/failure), the **Beta distribution** is the conjugate prior. If your prior belief about the success probability is Beta(α, β), and you observe a success, your new posterior belief is simply Beta$(\alpha + 1, \beta)$.

15. **Answer:** Let's assume we have two arms.

    (a) **Sample:** At time t, the agent has posterior distributions for each arm, say Beta$(S_1 + 1, F_1 + 1)$ for arm 1 and Beta$(S_2 + 1, F_2 + 1)$ for arm 2. It draws one random sample from each: $\theta_1 \sim$ Beta$(S_1 + 1, F_1 + 1)$ and $\theta_2 \sim$ Beta$(S_2 + 1, F_2 + 1)$.

    (b) **Select:** The agent compares the samples. Let's say $\theta_2 > \theta_1$. It selects arm 2 to play.

    (c) **Update:** The agent plays arm 2 and observes a reward, say a success ($r_t = 1$). It then updates the parameters for arm 2 only: $S_2 \leftarrow S_2 + 1$. The parameters for arm 1 remain unchanged for this time step. The new posterior for arm 2 is now Beta$(S_2 + 2, F_2 + 1)$.

16. **Answer:** A purely greedy strategy only ever exploits its current knowledge. At the beginning of the process, its knowledge is based on very few samples. If, by chance, a suboptimal arm gives a high reward on the first pull, the greedy agent might decide it's the best arm. It will then continue to pull that arm indefinitely, never exploring to find out that another arm is actually much better. It gets stuck in a "local optimum" and fails to maximize its long-term cumulative reward.

17. **Answer:** The general learning rule is: 'NewEstimate ¡- OldEstimate + StepSize(Target - OldEstimate)'. For the standard incremental update rule, $Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$:

    - **NewEstimate:** Q_{n+1}, the updated value estimate.
    - **OldEstimate:** Q_n, the value estimate before the update.
    - **StepSize:** $\frac{1}{n}$, the learning rate that decreases over time.
    - **Target:** R_n, the most recently observed reward, which serves as a noisy but unbiased sample of the true value.

18. **Answer:** I would choose a **contextual bandit**. **Justification:** The best article to recommend is almost certainly not the same for every user. The choice depends heavily on context. This context could include user features (demographics, location, past articles they've read) and information about the current session (time of day, device being used). A standard MAB would only find the single article that is most popular on average, while a contextual bandit (like LinUCB) could learn a personalized policy, such as "show sports articles to users who have previously read sports articles" or "show breaking news to users browsing in the morning." This personalization is crucial for a good user experience and high engagement.

19. **Answer:** An exponential recency-weighted average is a type of moving average where the weight assigned to each past data point (in this case, a reward) decreases exponentially the further back in time it was received. This means the most recent rewards have the largest influence on the current estimate, while very old rewards have very little influence. This is desirable in non-stationary environments because it allows the agent to "forget" outdated information from when the reward distributions might have been different, and instead adapt its estimates to reflect the current state of the environment.

20. **Answer:** The parameter c in the UCB formula controls the degree of exploration. It adjusts the "confidence level" or the size of the uncertainty bonus.

    - **If c is very large:** The exploration term dominates. The agent will be highly exploratory, frequently choosing arms with high uncertainty even if their estimated value is low. It will be slow to exploit the best arm.
    - **If c is close to zero:** The exploration term becomes negligible. The UCB algorithm will behave almost identically to a purely greedy agent, choosing the arm with the highest $Q_t(a)$ and exploring very little.

21. **Answer:** The term I_d (an identity matrix, often scaled by a parameter λ) is an L2 regularization (or Tikhonov regularization) term. Its primary purpose is to prevent overfitting by penalizing large values in the coefficient vector $\hat{\theta}_a$. A crucial secondary benefit is mathematical stability: it ensures that the matrix $(D_a^T D_a + I_d)$ is always invertible, even if $D_a^T D_a$ is not (which can happen early on when few data points have been collected).

22. **Answer:** Yes, Thompson Sampling can still be used. The concept of posterior sampling is general. However, the lack of a conjugate prior presents a significant computational challenge. Without conjugacy, the posterior distribution does not have a simple closed-form solution. Therefore, one must resort to more complex and computationally expensive approximation techniques, such as Markov Chain Monte Carlo (MCMC) methods (like Gibbs sampling) or Variational Inference, to draw samples from the posterior distribution at each step.

23. **Answer:**

    - **UCB:**
      - *Advantages:* Has strong theoretical regret guarantees. Its deterministic nature can make it easier to debug and analyze.
      - *Disadvantages:* Can be overly conservative, as it may continue to explore arms that are clearly suboptimal to satisfy its mathematical bounds. Performance is sensitive to the choice of the exploration parameter c.
    - **Thompson Sampling:**
      - *Advantages:* Often shows superior empirical performance (lower regret). It naturally and elegantly balances the trade-off without tuning parameters like c. Its probabilistic nature can make it quicker to abandon bad arms.
      - *Disadvantages:* Can be computationally intensive without a conjugate prior. The Bayesian concepts can be less intuitive for beginners.

- **Preference:** For best empirical performance, Thompson Sampling is often a strong choice. If theoretical guarantees and deterministic behavior are critical for a specific application, UCB might be preferred.

24. **Answer:** Both methods implement "optimism in the face of uncertainty" to drive exploration.

    - **Optimistic Initial Values:** This is a simple, one-shot implementation of the principle. By setting initial estimates high, the agent is "optimistic" about all unexplored arms. It is forced to explore them because the reality of the first reward received from any chosen arm will be less than the initial optimistic guess, making other unexplored arms look more attractive.
    - **UCB Algorithm:** This is a more formal and sustained implementation. At every step, it calculates an optimistic upper confidence bound for each arm's value. The uncertainty term $c\sqrt{\frac{\ln t}{N_t(a)}}$ is an explicit bonus for uncertainty. The agent then greedily chooses the arm with the highest optimistic bound, ensuring that uncertain arms are given a chance to be explored.

25. **Answer:** Intuitively, the term $x_{t,a}^T A_a^{-1} x_{t,a}$ represents the **variance of the reward prediction** at the specific context point $x_{t,a}$. The matrix A_a accumulates information about the directions in the feature space where the arm has been tried. Its inverse, A_a^{-1}, is proportional to the covariance of the parameter estimate. Therefore, this quadratic form measures how uncertain the model's prediction is for the current context vector. A large value means high uncertainty in this region of the feature space, which boosts the exploration score.

26. **Answer:** Thompson Sampling balances the trade-off through the properties of the posterior distributions.

    - **Exploration:** An arm with high uncertainty (e.g., it has been pulled only a few times) will have a wide posterior distribution. When sampling from this wide distribution, there is a non-trivial chance of drawing a very high value, which would lead to the arm being selected. This encourages exploration of uncertain arms.
    - **Exploitation:** An arm that has been pulled many times and found to be good will have a narrow posterior distribution centered around a high value. It will consistently produce high samples and thus be chosen frequently (exploited). Conversely, a well-understood bad arm will have a narrow posterior centered at a low value and will rarely be chosen.

27. **Answer:**

    (a) **Application: Online Advertising**

        - **Arms:** The different advertisements that can be displayed in a slot on a webpage.
        - **Reward:** A binary value: 1 if the user clicks the ad, 0 if they do not. The goal is to maximize the click-through rate.

    (b) **Application: Dynamic Pricing**

- **Arms:** A set of discrete price points for a product (e.g., \$9.99, \$12.99, \$14.99).
- **Reward:** The revenue generated from a user being shown a certain price. This could be the price itself if they purchase, and 0 if they don't. The goal is to find the price that maximizes total revenue.

28. **Answer:** The key difference is in what gets updated.

    - In a **non-contextual algorithm like UCB**, when an arm is chosen, you only update the statistics for that single arm (e.g., its count $N_t(a)$ and its value estimate $Q_t(a)$). The information learned is isolated to that specific arm.
    - In **LinUCB**, when an arm is chosen, you update the parameters of its linear model (θ_a) using the context vector $x_{t,a}$. This update not only improves the prediction for that specific context but also helps generalize to other, similar contexts. The learning is not just about the arm, but about the relationship between contexts and rewards for that arm.

29. **Answer:** The MAB problem is a powerful reinforcement learning model because it isolates and focuses on the core challenge of the **exploration-exploitation trade-off** in its purest form. The key RL element that is simplified or removed is the concept of **state transitions**. In a full MDP, an agent's action in a state can influence which state it ends up in next. In the MAB problem, there is effectively only one state; the choice of an arm does not affect the reward distributions or availability of any other arms in the future. This simplification allows for a focused study of learning under uncertainty without the added complexity of long-term planning across different states.

30. **Answer:** For a new stationary bandit problem where performance is the main goal, I would most likely choose **Thompson Sampling**. **Justification:**

    - **Performance:** Empirical evidence consistently shows that Thompson Sampling often outperforms both ϵ-greedy and UCB in terms of achieving lower regret and higher cumulative reward.
    - **Principled Exploration:** Unlike ϵ-greedy's random exploration, TS's exploration is directed by uncertainty in a very natural and effective way.
    - **No Parameter Tuning:** Unlike UCB, which requires careful tuning of the exploration parameter c, Thompson Sampling's core algorithm does not have an equivalent hyperparameter, making it easier to deploy "out of the box" (assuming a standard non-informative prior like Beta(1,1)). Its elegant mechanism of sampling from posteriors is a robust way to manage the trade-off.