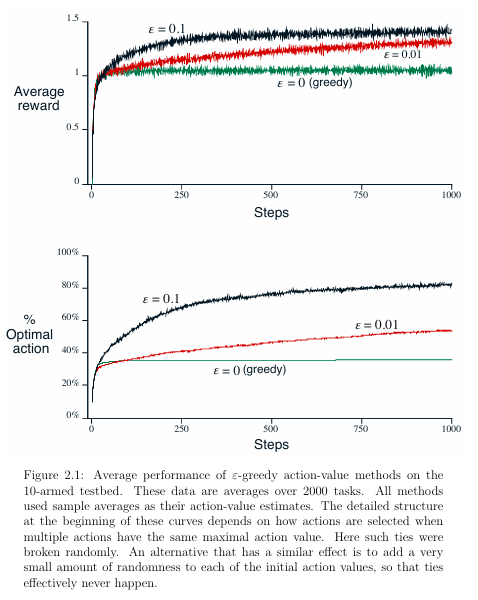
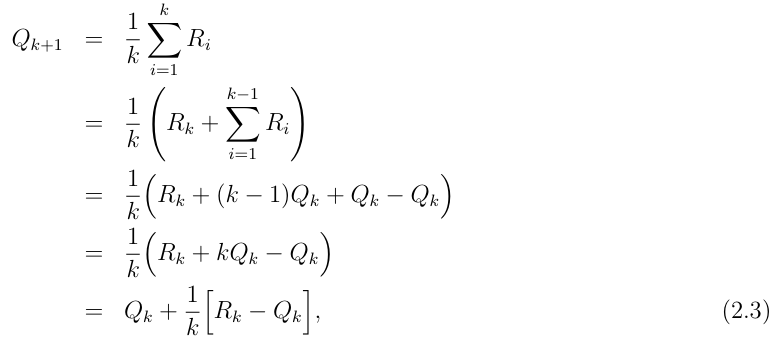
**Chapter 2 Multi-arm Bandits**

2.2 Action-Value Methods



2.3 Incremental Implementation (Tracking a Stationary Problem)

Given this average and a kth reward for the action, Rk, then the average of all k rewards can be computed by:



which holds even for k = 1, obtaining Q2 = R1 for arbitrary Q1. This simple mentation requires memory only for Qk and k, and only the small computation (2.3) for each new reward.

The update rule (2.3) is of a form that occurs frequently throughout this book. The general form is:



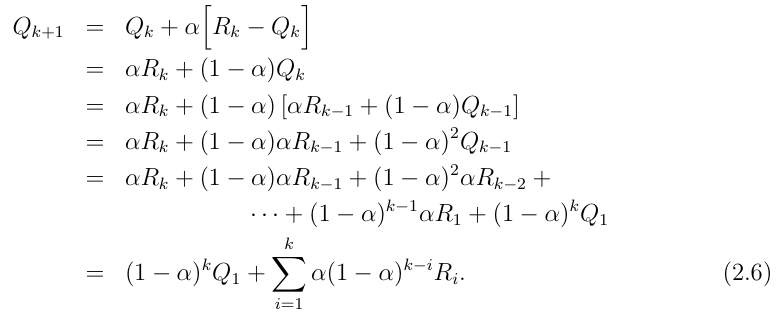
We sometimes use the informal shorthand = 1 k to refer to this case, leaving the dependence of k on the action implicit.

2.4 Tracking a Nonstationary Problem

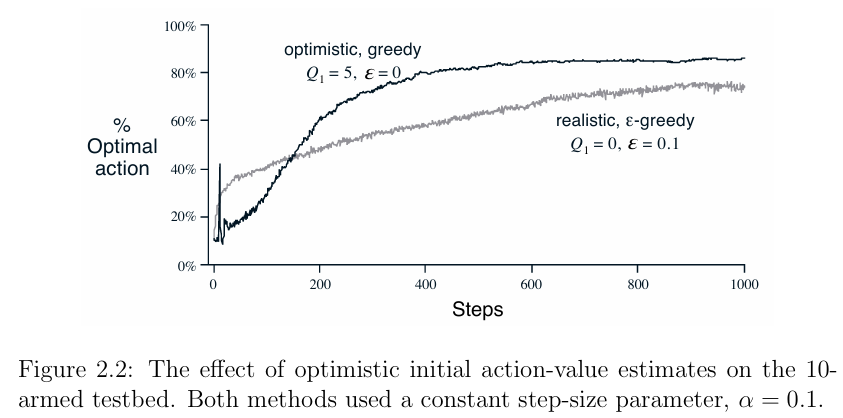
the incremental update rule (2.3) for updating an average Qk of the k 1 past rewards is modified to be:



where the step-size parameter is constant. This results in Qk+1 being a weighted average of past rewards and the initial estimate Q1:



2.5 Optimistic Initial Values



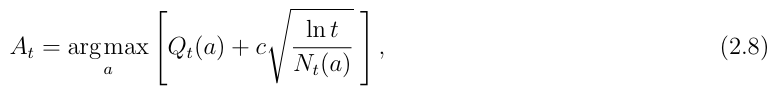
For the sample-average methods, the bias disappears once all actions have been selected at least once, but for methods with constant e , the bias is permanent, though decreasing over time as given by (2.6).

Figure 2.2 shows the performance on the 10-armed bandit testbed of a greedy method using Q1(a) = +5, for all a. For comparison, also shown is an-greedy method with Q1(a) = 0. Initially, the optimistic method performs worse because it explores more, but eventually it performs better because its exploration decreases with time. We call this technique for encouraging exploration optimistic initial values.

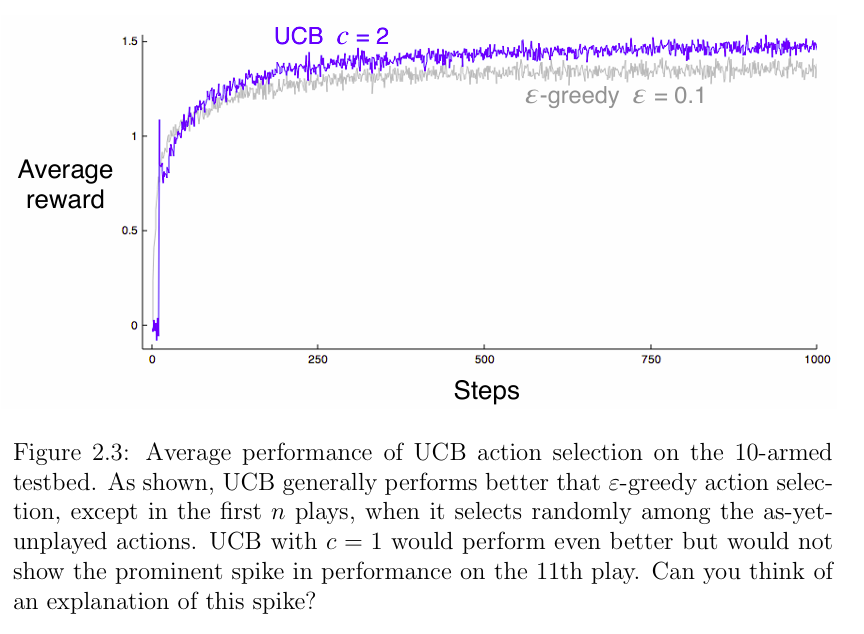
2.6 Upper-Confidence-Bound Action Selection

Exploration is needed because the estimates of the action values are uncertain. The greedy actions are those that look best at present, but some of the other actions may actually be better.-greedy action selection forces the non-greedy actions to be tried, but indiscriminately, with no preference for those that are nearly greedy or particularly uncertain.

It would be better to select among the non-greedy actions according to their potential for actually being optimal, taking into account both how close their estimates are to being maximal and the uncertainties in those estimates. One e ective way of doing this is to select actions as:

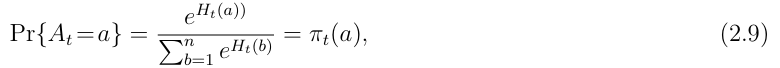


where ln(t) denotes the natural logarithm of t (the number that e 271828 would have to be raised to in order to equal t), and the number c > 0 controls the degree of exploration. If Nt(a) = 0, then a is considered to be a maximizing action.

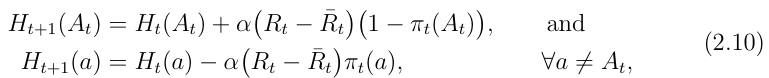


2.7 Gradient Bandits

we consider learning a numerical preference Ht(a) for each action a. The larger the preference, the more often that action is taken, but the preference has no interpretation in terms of reward.



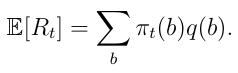
There is a natural learning algorithm for this setting based on the idea of stochastic gradient ascent. On each step, after selecting the action At and receiving the reward Rt, the preferences are updated by:

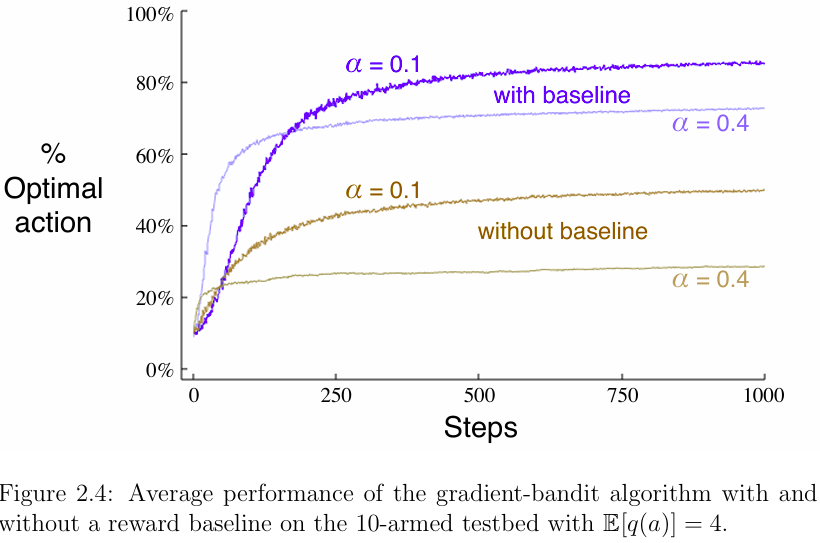


In exact gradient ascent, each preference Ht(a) would be incrementing proportional to the increments exact on performance:



where the measure of performance here is the expected reward:





2.9 Summary

The-greedy methods choose randomly a small fraction of the time, whereas UCB methods choose deterministically but achieve exploration by subtly favoring at each step the actions that have so far received fewer samples.

Gradient-bandit algorithms estimate not action values, but action preferences, and favor the more preferred actions in a graded, probabalistic manner using a soft-max distribution.

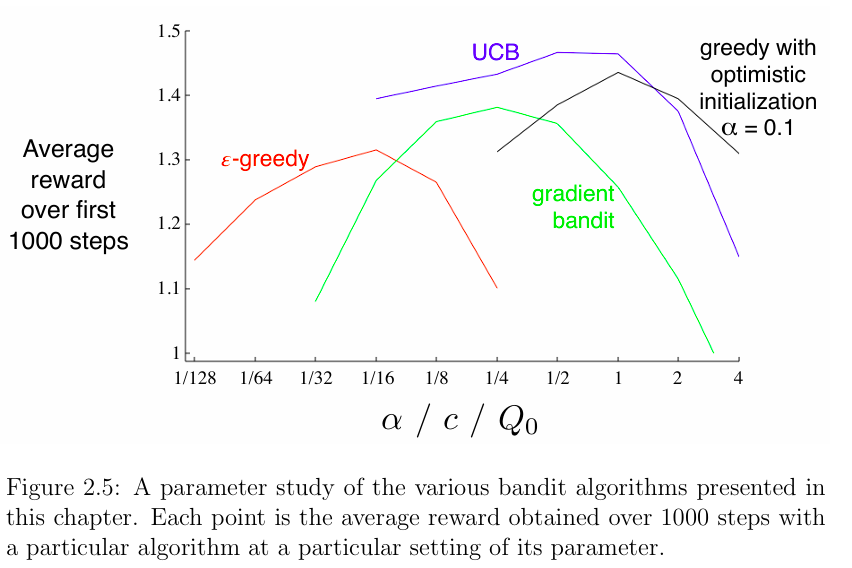
The simple expedient of initializing estimates optimistically causes even greedy methods to explore significantly.

It is natural to ask which of these methods is best. Although this is a di cult question to answer in general, we can certainly run them all on the 10-armed testbed that we have used.

A complication is that they all have a parameter; to get a meaningful comparison we will have to consider their performance as a function of their parameter. Our graphs so far have shown the course of learning over time for each algorithm and parameter setting, but it would be too visually confusing to show such a learning curve for each algorithm and parameter value.

Instead we summarize a complete learning curve by its average value over the 1000 steps; this value is proportional to the area under the learning curves we have shown up to now.

Figure 2.5 shows this measure for the various bandit algorithms from this chapter, each as a function of its own parameter shown on a single scale on the x-axis.



Note that the parameter values are varied by factors of two and presented on a log scale. Note also the characteristic inverted-U shapes of each algorithms performance; all the algorithms perform best at an intermediate value of their parameter, neither too large nor too big.

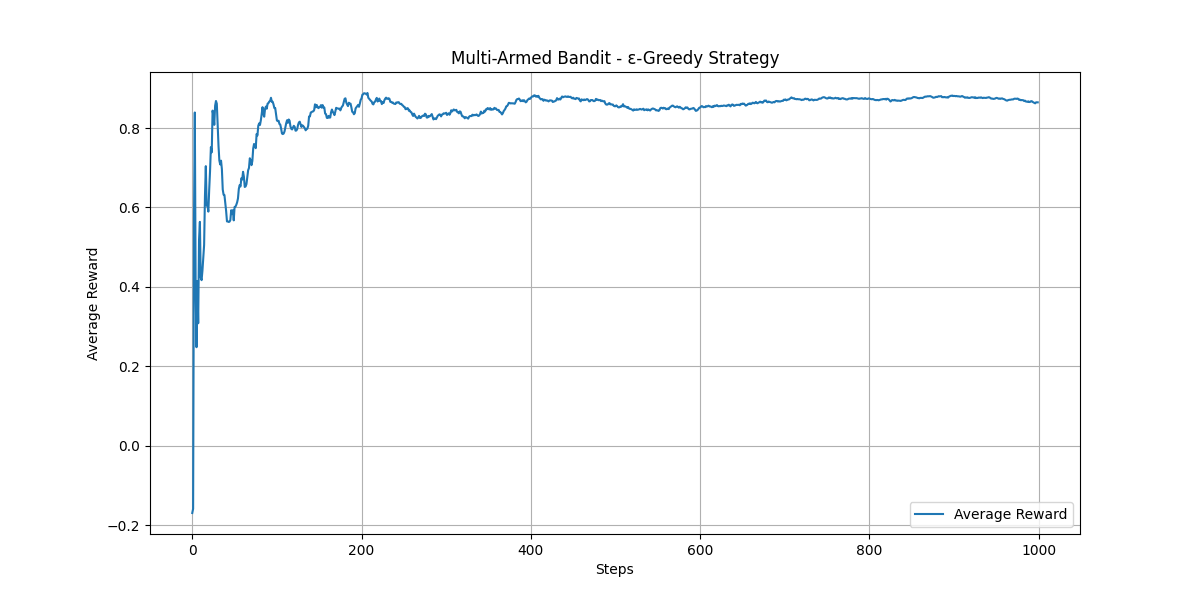
In assessing a method, we should attend not just to how well it does at its best parameter setting, but also to how sensitive it is to its parameter value. All of these algorithms are fairly insensitive, performing well over a range of parameter values varying by about an order of magnitude. Overall, on this problem, UCB seems to perform best.

**Python Programs:**

**Multi-Armed Bandit with ε-Greedy Strategy (Stationary)**

**Explanation:**

1. **Bandit Class**:
   * Initializes the number of arms and their true values.
   * Contains methods to pull an arm and update estimates.
2. **Epsilon Greedy Agent Class**:
   * Chooses an arm to pull based on the ε-greedy policy.
3. **Simulate Function**:
   * Runs the simulation for a specified number of steps, pulling arms and updating estimates.
4. **Parameters**:
   * You can adjust the number of arms, steps, and the epsilon value for exploration.



The multi-armed bandit program serves several purposes, particularly in the context of reinforcement learning and decision-making. Here are some key uses:

### 1. ****Exploration vs. Exploitation****

* The program illustrates the fundamental trade-off in reinforcement learning: balancing exploration (trying new options) and exploitation (choosing the best-known option).

### 2. ****Algorithm Testing****

* It provides a platform for testing different strategies (like ε-greedy) for action selection, which can be compared against other methods (like UCB, Thompson Sampling).

### 3. ****Performance Evaluation****

* By running simulations, you can evaluate the performance of the bandit algorithm, such as how quickly it converges to the optimal arm and how well it balances exploration and exploitation.

### 4. ****Application in Real-World Scenarios****

* The multi-armed bandit approach is applicable in various fields, such as:
  + **Online Advertising**: Selecting which ad to display to maximize click-through rates.
  + **A/B Testing**: Determining the best version of a webpage or product by comparing different variations.
  + **Recommendation Systems**: Choosing which items to recommend to users based on their preferences.

### 5. ****Educational Tool****

* It serves as an educational example for understanding the concepts of reinforcement learning, helping learners grasp the mechanics of bandit algorithms and their implementations.

### 6. ****Dynamic Decision-Making****

* The program can be adapted to solve dynamic decision-making problems where the environment changes over time, allowing the agent to continuously learn and adapt its strategy.

### Summary

Overall, the program is a practical demonstration of reinforcement learning principles, providing insights into decision-making processes across various domains.

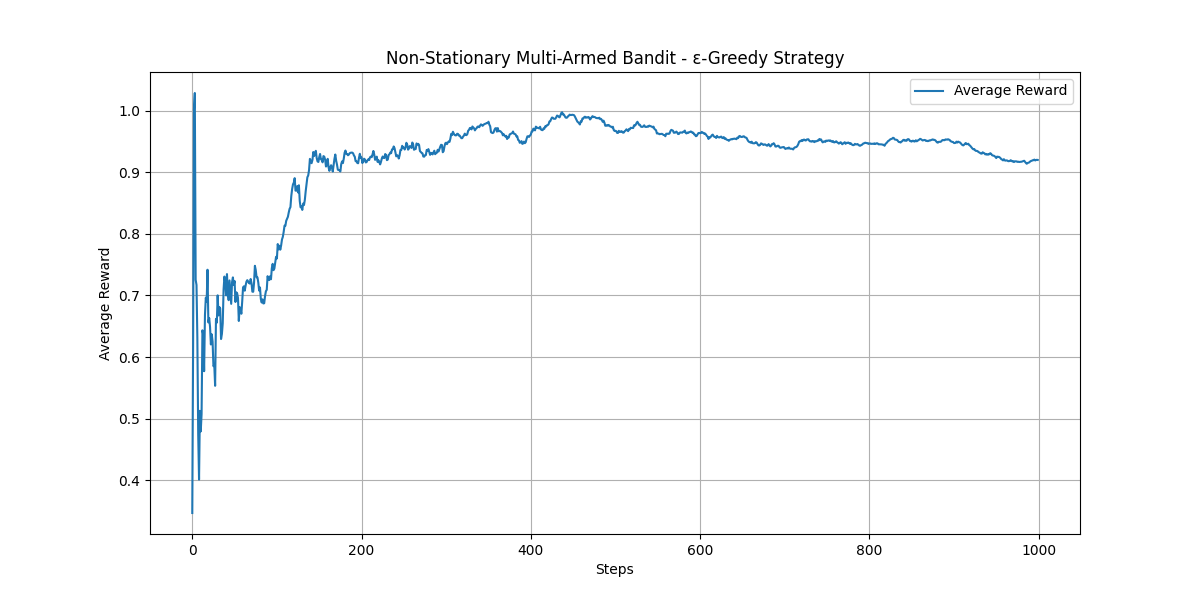
### Non-Stationary Multi-Armed Bandit Implementation

**Key Components of the Code**

1. **Bandit Class**:
   * The pull\_arm method simulates pulling an arm while incorporating a small random walk to the true values (q\_true) to reflect non-stationarity.
2. **EpsilonGreedyAgent Class**:
   * Implements the Epsilon-Greedy strategy for action selection.
3. **simulate Function**:
   * Runs the simulation for a specified number of steps, updating estimates based on the rewards received.

**Non-Stationary Dynamics**

* The true values of each arm (q\_true) are updated slightly at each step to simulate a changing environment.
* You can adjust the randomness of this change by modifying the parameters in the np.random.normal() function within the pull\_arm method.



The non-stationary multi-armed bandit problem has various applications across multiple fields, particularly in scenarios where the environment or the underlying reward distributions change over time. Here are some key uses:

**1. Online Advertising**

* **Dynamic Ad Selection**: In online marketing, the effectiveness of ads can change based on user behavior, time of day, or seasonal trends. Non-stationary bandit algorithms can optimize which ads to show to maximize click-through rates.

**2. Recommendation Systems**

* **Personalized Recommendations**: Platforms like Netflix or Amazon can use non-stationary bandit approaches to adapt recommendations based on changing user preferences over time.

**3. Clinical Trials**

* **Adaptive Treatment Allocation**: In medical research, non-stationary bandit algorithms can help allocate treatments to patients based on observed effectiveness, adapting to changing patient responses.

**4. A/B Testing**

* **Dynamic Experimentation**: Businesses can use non-stationary bandits to dynamically allocate traffic to different variations of a webpage, adapting to real-time performance data to optimize conversions.

**5. Finance**

* **Trading Strategies**: In algorithmic trading, strategies can be adapted based on changing market conditions, optimizing for returns as the financial landscape evolves.

**6. Game Development**

* **Player Behavior Adaptation**: In video games, non-stationary bandit algorithms can adjust difficulty levels or in-game rewards based on player performance and engagement.

**7. Robotics**

* **Adaptive Learning**: Robots can use non-stationary bandit strategies to learn and adapt their actions based on changing environments or tasks over time.

**8. Network Optimization**

* **Resource Allocation**: In telecommunications, non-stationary bandit approaches can optimize the allocation of bandwidth or resources based on fluctuating demand patterns.

**Summary**

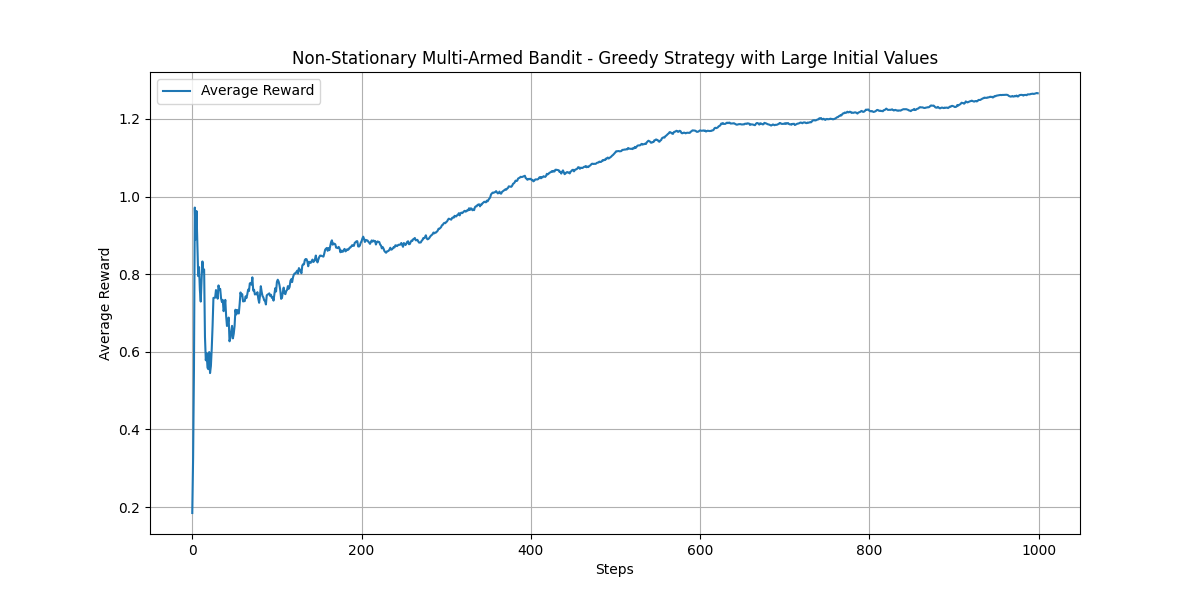
Non-stationary multi-armed bandit algorithms are essential for applications where the optimal choice may change over time, allowing for dynamic adaptation and improved decision-making in uncertain environments. This adaptability can lead to better performance and higher returns across various sectors.

**Optimistic Initial Values (Non-Stationary)**

**Key Changes:**

1. **Initial Values**: Set self.q\_estimates in the Bandit class to a large constant (e.g., 10.0).
2. **Agent Class**: Removed the exploration aspect by creating a GreedyAgent that always selects the arm with the highest estimated value.

This modification will allow the agent to exploit its initial overestimation of the arms while it learns about the true rewards.



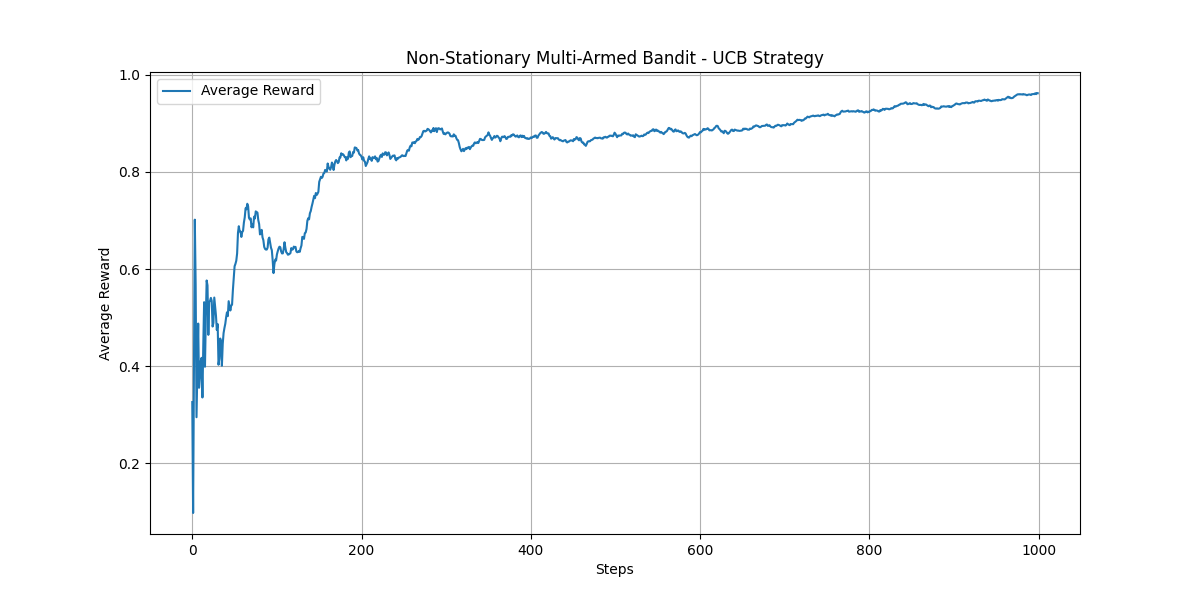
**UCB Algorithm for Non-Stationary Multi-Armed Bandit**

**Key Components of the Code**

1. **Bandit Class**:
   * Similar to the previous implementation, it simulates pulling an arm with a non-stationary reward.
2. **UCBAgent Class**:
   * Implements the UCB strategy for action selection:
     + It calculates UCB values for each arm based on the average reward and uncertainty (confidence interval).
     + The arm with the highest UCB value is selected.
3. **simulate Function**:
   * Runs the simulation for a specified number of steps, updating estimates based on the rewards received.

**Explanation of UCB Strategy**

* **Exploration vs. Exploitation**: The UCB algorithm explicitly balances exploration (trying out arms that have not been selected often) and exploitation (selecting arms that have shown good performance).
* **Confidence Interval**: The UCB value for each arm increases with uncertainty (i.e., the number of times that arm has been pulled). This encourages exploration of less-tried arms.
* **Non-Stationarity Handling**: Similar to the Epsilon-Greedy approach, the true values of each arm change over time, and the UCB agent adapts its strategy accordingly.



The Upper Confidence Bound (UCB) algorithm for non-stationary multi-armed bandit problems is useful in various contexts. Here are its key applications and advantages:

**1. Effective Exploration and Exploitation**

* UCB balances exploration (trying out less-known options) and exploitation (favoring known high-reward options) effectively, adapting as new data becomes available.

**2. Dynamic Environments**

* It is designed to handle non-stationary environments where the reward distributions change over time, making it suitable for real-world applications like online advertising and recommendation systems.

**3. Theoretical Guarantees**

* UCB has strong theoretical foundations, offering performance bounds that can assure users of its effectiveness compared to other strategies.

**4. Reduced Regret**

* The algorithm minimizes regret (the difference between the rewards obtained and the rewards that could have been obtained by always playing the best arm) over time, leading to better long-term performance.

**5. Adaptability**

* UCB can be adapted to incorporate prior knowledge about the arms, making it versatile for various applications where some information is known.

**6. Real-Time Decision Making**

* It is particularly useful in scenarios requiring real-time decision-making, such as dynamic pricing and adaptive clinical trials.

**7. Simplicity of Implementation**

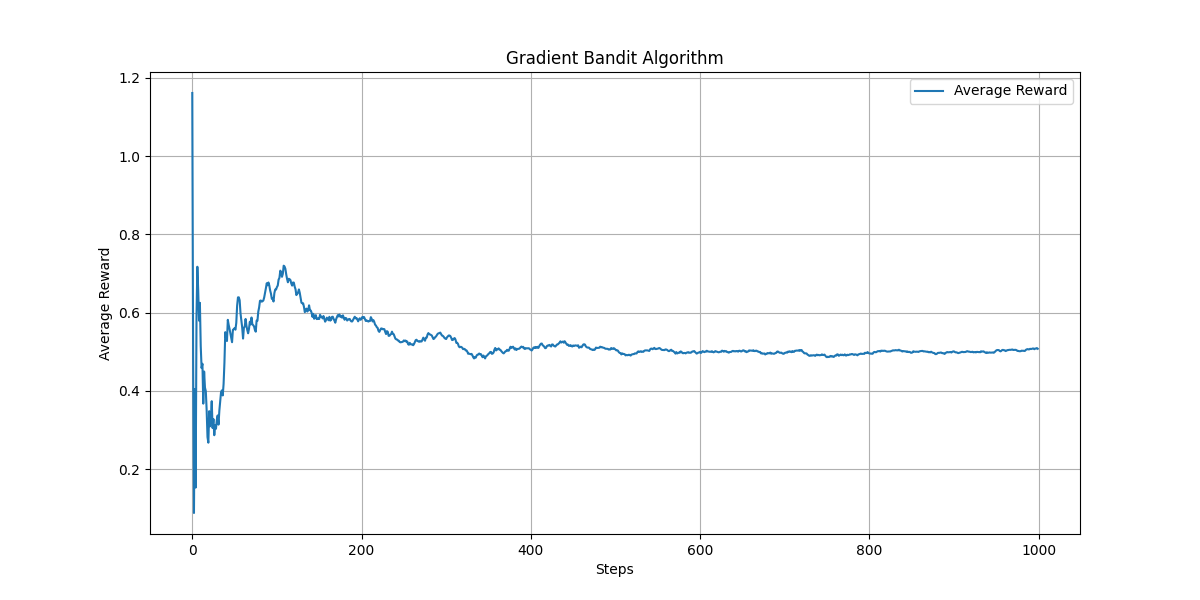
* The algorithm is relatively straightforward to implement, making it accessible for practitioners and researchers.

In summary, UCB is valuable in contexts requiring adaptive decision-making in uncertain and changing environments, ensuring efficient use of resources while maximizing rewards.

**Gradient Bandit Implementation**

**Explanation of the Code**

1. **Class Definition**:
   * GradientBandit: Initializes the number of arms, step size, preferences, action counts, and initial probabilities.
2. **Softmax Function**:
   * Computes the action probabilities based on the preferences using the softmax function.
3. **Select Arm**:
   * Chooses an arm to pull based on the computed probabilities.
4. **Pull Arm**:
   * Simulates pulling the selected arm and generates a reward based on a normal distribution centered around the arm's true value.
5. **Update Preferences**:
   * Adjusts the preferences of the arms based on the received reward and the average reward.
6. **Simulation Function**:
   * Runs the bandit simulation for a specified number of steps, recording the rewards.
7. **Plotting**:
   * Visualizes the average reward over time.



The Gradient Bandit algorithm is used in various applications, particularly when dealing with uncertain environments and decision-making processes. Here are some key uses and advantages:

**1. Online Learning**

* It adapts to changing circumstances by continuously learning from new data, making it suitable for online learning scenarios.

**2. Recommendation Systems**

* Useful in personalized recommendation systems (e.g., movies, products) where user preferences may change over time.

**3. Adaptive Clinical Trials**

* In clinical trials, it can help in dynamically selecting treatment options based on patient responses, optimizing outcomes.

**4. Marketing and Advertising**

* Helps in optimizing ad placements and content shown to users by learning which ads perform best according to user interactions.

**5. Dynamic Pricing**

* Assists businesses in adjusting prices based on customer responses and competition, maximizing revenue.

**6. Resource Allocation**

* Efficiently allocates resources in scenarios where different options have uncertain returns, such as in network routing or logistics.

**7. Exploration vs. Exploitation Trade-off**

* Effectively manages the exploration-exploitation trade-off, allowing systems to discover new options while capitalizing on known successful ones.

**8. Real-Time Decision Making**

* Suitable for applications requiring real-time decisions based on immediate feedback, such as online games or interactive platforms.

**Summary**

Overall, the Gradient Bandit algorithm is valuable in any context where decision-making must adapt to feedback over time, helping to optimize actions and improve outcomes in uncertain environments.

Determining which method has the "best value" in multi-armed bandit problems depends on various factors, including the specific application, environment characteristics, and the objective of the task. Here’s a comparison of some common methods:

**1. Epsilon-Greedy**

* **Pros**: Simple to implement and understand; good for static environments.
* **Cons**: May not explore effectively in non-stationary environments.

**2. Upper Confidence Bound (UCB)**

* **Pros**: Balances exploration and exploitation well; provides theoretical guarantees on performance.
* **Cons**: Can be computationally heavier with large action spaces; assumes stationary rewards.

**3. Thompson Sampling**

* **Pros**: Often performs well in practice; naturally balances exploration and exploitation; can adapt well to non-stationary environments.
* **Cons**: More complex to implement; requires maintaining and updating a probability distribution for each arm.

**4. Gradient Bandits**

* **Pros**: Effective in dynamic environments; adapts preferences based on rewards; good for problems with continuously changing distributions.
* **Cons**: Requires careful tuning of the step size; might converge slowly in some cases.

**5. Contextual Bandits**

* **Pros**: Incorporates user or contextual information, leading to more personalized recommendations; effective in many practical applications like ads and recommendations.
* **Cons**: More complex; requires additional data and modeling of context.

**Conclusion**

* **Best Choice**: The "best" method varies by use case:
  + For **static environments**, Epsilon-Greedy may suffice.
  + For **dynamic or non-stationary environments**, UCB and Thompson Sampling often perform better.
  + For applications requiring adaptation to user behavior (like **recommendation systems**), Gradient Bandits or Contextual Bandits might be optimal.

Ultimately, the choice should be guided by the specific goals, constraints, and characteristics of the environment in which the method will be applied.