RLearning:
Short guides to reinforcement learning

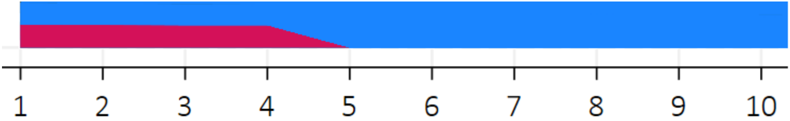
Unit 1-2: Greedy, ε-greedy, decaying ε-greedy

Davud Rostam-Afschar (Uni Mannheim)

How much to learn about the average return?

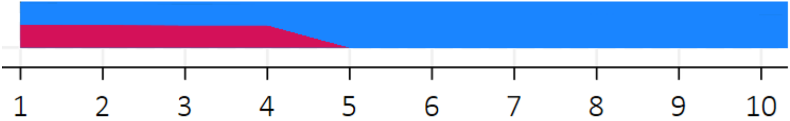
ε-First + Greedy Policy

ε-First + Greedy Policy



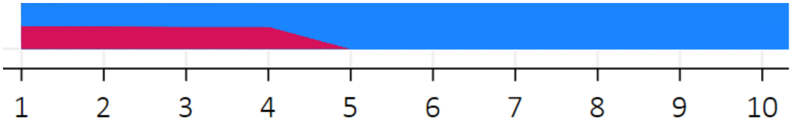
- Epsilon-first is widely known as A/B testing
- Often applied to two-armed bandits

Fixed Exploration Period + Greedy



1. Allocate a fixed time period to exploration, during which you try all bandits uniformly at random.

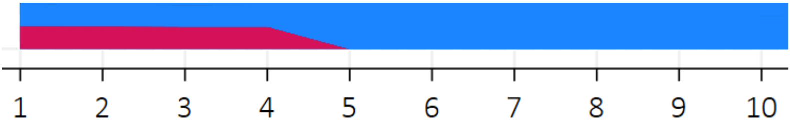
Fixed Exploration Period + Greedy



1. Allocate a fixed time period to exploration, during which you try all bandits uniformly at random.
2. Estimate mean rewards for all actions:

$$Q_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^{t-1} R_i \cdot \mathbb{1}(A_i = a)$$

Fixed Exploration Period + Greedy



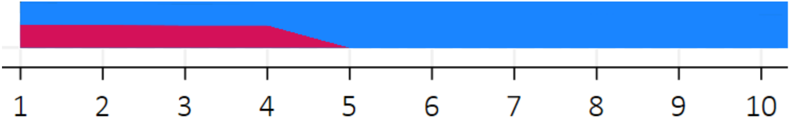
1. Allocate a fixed time period to exploration, during which you try all bandits uniformly at random.
2. Estimate mean rewards for all actions:

$$Q_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^{t-1} R_i \cdot \mathbb{1}(A_i = a)$$

3. Select the action that is optimal for the estimated mean rewards (breaking ties randomly):

$$a_t = \arg\max_{a \in \mathcal{A}} Q_t(a)$$

Fixed Exploration Period + Greedy



1. Allocate a fixed time period to exploration, during which you try all bandits uniformly at random.
2. Estimate mean rewards for all actions:

$$Q_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^{t-1} R_i \cdot \mathbb{1}(A_i = a)$$

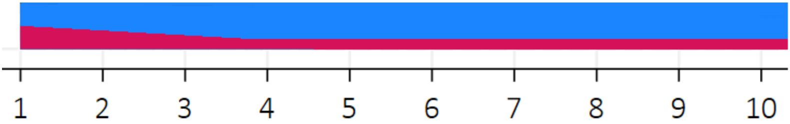
3. Select the action that is optimal for the estimated mean rewards (breaking ties randomly):

$$a_t = \arg\max_{a \in \mathcal{A}} Q_t(a)$$

4. Repeat step 3 for all future time steps.

ε-greedy

ε-greedy



- Explores for the entire number of trials of the experiment
- simple and popular heuristic (**???**)

ε-greedy

Idea: Exploit, but explore a random arm with ε probability

1. **Initial phase:** Try each arm and observe the rewards

Idea: Exploit, but explore a random arm with ε probability

1. **Initial phase:** Try each arm and observe the rewards
2. **For each round** $t = n + 1, \ldots, T$:
   - Estimate action values from *sample averages* for each arm a:

$$Q_t(a) = \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}(A_i = a)}{\sum_{i=1}^{t-1} \mathbb{1}(A_i = a)}$$

ε-greedy

Idea: Exploit, but explore a random arm with ε probability

1. **Initial phase:** Try each arm and observe the rewards
2. **For each round** $t = n+1, \ldots, T$:
   - Estimate action values from *sample averages* for each arm a:

$$Q_t(a) = \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}(A_i = a)}{\sum_{i=1}^{t-1} \mathbb{1}(A_i = a)}$$

   - With probability $1 - \varepsilon$, play the arm with highest $Q_t(a)$

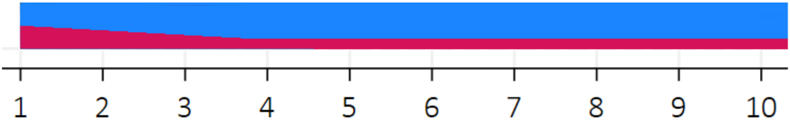
Idea: Exploit, but explore a random arm with ε probability

1. **Initial phase:** Try each arm and observe the rewards
2. **For each round** $t = n + 1, \ldots, T$:
    - Estimate action values from *sample averages* for each arm a:

$$Q_t(a) = \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}(A_i = a)}{\sum_{i=1}^{t-1} \mathbb{1}(A_i = a)}$$

    - With probability $1 - \varepsilon$, play the arm with highest $Q_t(a)$
    - With probability ε, choose an arm uniformly at random

A Simple ε-Greedy Bandit Algorithm



- **Initialize:** For each action $a = 1$ to k:
  - $Q(a) \leftarrow 0$
  - $N(a) \leftarrow 0$
- **Loop forever:**

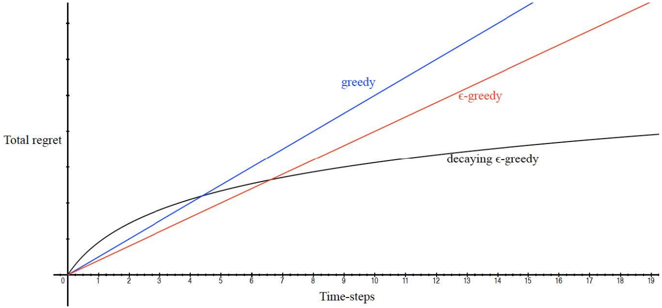
$$A = \begin{cases} \arg\max_a Q(a) & \text{with probability } 1 - \varepsilon \\ \text{random action} & \text{with probability } \varepsilon \end{cases}$$

  - Receive reward: $R \leftarrow \text{bandit}(A)$
  - Update count: $N(A) \leftarrow N(A) + 1$
  - Update estimate:

$$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}(R - Q(A))$$

Exploration vs Exploitation

Regrets of Greedy Policies



Source: David Silver

Greedy Policy	**ε-Greedy**	**Decaying ε**
Never explores	Always explores with probability ε	Decreases exploration over time
Locks on sub-optimal policy	See decomposition lemma	Requires careful tuning
Linear regret	*Linear* regret	*Sub-Linear* regret

$\Rightarrow$ *Convergence rate depends on ε choice (**?**)*

$$\text{Loss}_T = \sum_{t=1}^{T} \text{loss}_t = \sum_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{i=1}^{t-1} \mathbb{1}\{A_i = a\}\right](r^* - q(a))$$

- When ε is constant, probability to explore in each step t is ε
- Each action is selected with probability $1/\mathcal{A}$
- Probability of choosing a suboptimal action $\mathbb{P}(a_t \neq a^*) = \varepsilon/\mathcal{A}$
- Expected regret: $\text{loss}_t \geq \frac{\varepsilon}{\mathcal{A}} \sum_{a \in \mathcal{A}} (r^* - q(a))$

Theoretical Guarantees

$$\text{Loss}_T = \sum_{t=1}^{T} \text{loss}_t = \sum_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{i=1}^{t-1} \mathbb{1}\{A_i = a\}\right] (r^* - q(a))$$

- When ε is constant, probability to explore in each step t is ε
- Each action is selected with probability $1/\mathcal{A}$
- Probability of choosing a suboptimal action $\mathbb{P}\left(a_t \neq a^*\right) = \varepsilon/\mathcal{A}$
- Expected regret: $\text{loss}_t \geq \frac{\varepsilon}{\mathcal{A}} \sum_{a \in \mathcal{A}} (r^* - q(a))$
- Expected number of times action a is selected due to exploration over T steps $\frac{\varepsilon T}{\mathcal{A}}$
- Expected cumulative regret: $\text{Loss}_T = \frac{\varepsilon T}{\mathcal{A}} \sum_{a \in \mathcal{A}} (r^* - q(a)) = \mathcal{O}(T)$
- Linear regret

Theoretical Guarantees

- When $\varepsilon \propto 1/t$
  - For large enough $t : \mathbb{P}\left(a_t \neq a^*\right) \approx \varepsilon_t = \mathcal{O}(1/t)$
  - Expected cumulative regret: $\text{Loss}_T \approx \sum_{t=1}^{T} 1/t = \mathcal{O}(\log T)$
  - Logarithmic regret

AUER, P., N. CESA-BIANCHI, AND P. FISCHER (2002): "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, 47, 235–256.

BUBECK, S., AND N. CESA-BIANCHI (2012): "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," *Foundations and Trends® in Machine Learning*, 5(1), 1–122.

BURTINI, G., J. LOEPPKY, AND R. LAWRENCE (2015): "A survey of online experiment design with the stochastic multi-armed bandit," *arXiv preprint arXiv:1510.00757*.

SUTTON, R. S., AND A. G. BARTO (2018): "Reinforcement learning: An introduction," *A Bradford Book*, Available at http://incompleteideas.net/book/the-book-2nd.html.

Takeaways

What does the ε-greedy algorithm?

- ε-greedy algorithm balances exploration and exploitation
- With probability ε, it explores randomly
- With $1 - \varepsilon$, it chooses action with highest empirical mean
- A constant ε ensures ongoing exploration but leads to linear regret
- A decaying ε enables convergence to the optimal arm and may achieve logarithmic regret