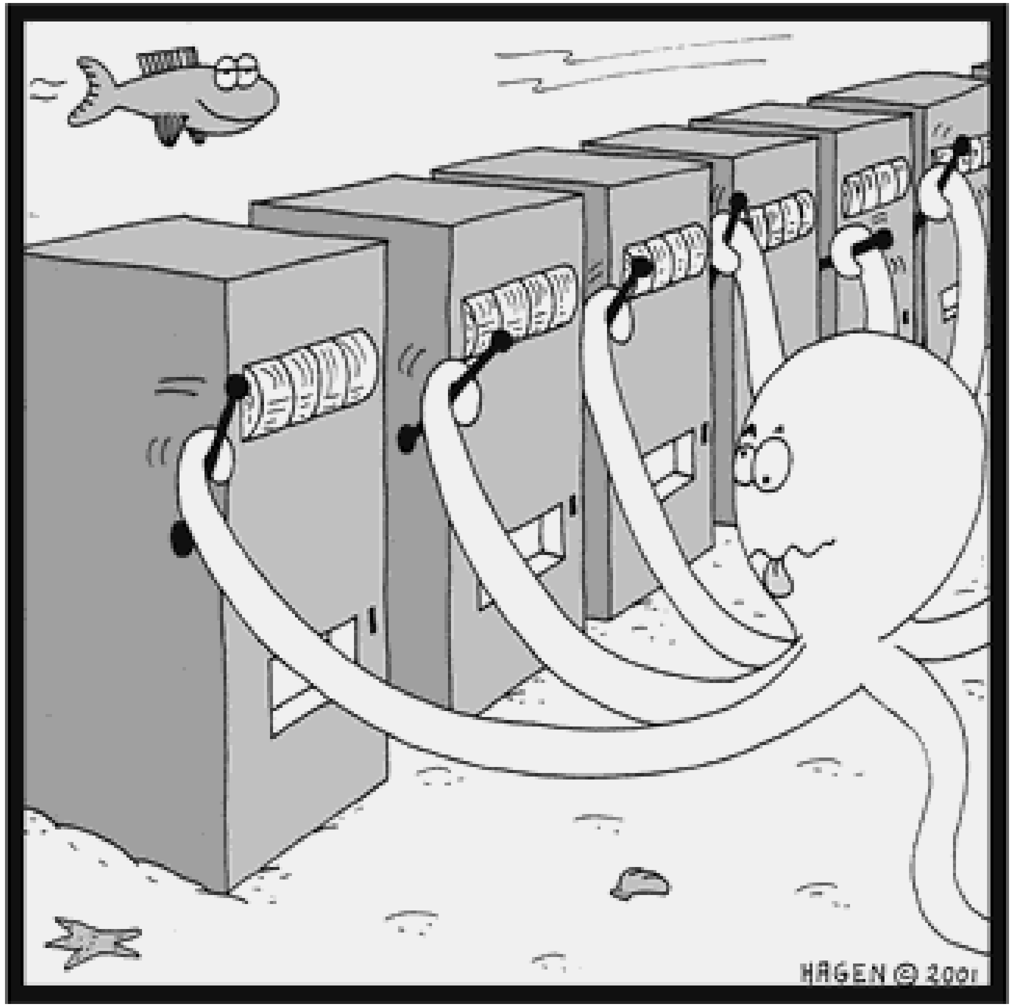
Reinforcement Learning Fundamentals

Lecture 6: Multi-armed Bandit

Dr Sandeep Manjanna
Assistant Professor, Plaksha University
sandeep.manjanna@plaksha.edu.in

In today's class…

- Value function-based methods
  - ϵ-greedy algorithms
- Performance metrics
  - Correctness
  - Convergence
  - Sample Efficiency



Value function-based Methods

- Consider that the estimated value of a given action ***a*** at timestep ***t*** is given by $Q_t(a)$

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

- Over time as the denominator goes to infinity, $Q_t(a)$ converges to $q_*(a)$ *(the true expected value of action a)*.

- We want to choose an action that gives the maximum estimated reward. What would be a greedy strategy to do this?

$$A_t \doteq \arg\max_a Q_t(a)$$

- But, what about exploration?

ϵ-greedy Algorithms

- Parameter $\epsilon \in [0, 1]$ controls the amount of exploration.

- ϵG1
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - At $t = \lfloor \epsilon T \rfloor$, identify a^{best}, an arm with the highest empirical mean.
  - If $t > \epsilon T$, sample a^{best}.

$$A_t \doteq \arg\max_a Q_t(a)$$

- ϵG2
  - If $t \leq \epsilon T$, sample an arm uniformly at random.
  - If $t > \epsilon T$, sample an arm with the highest empirical mean.

Usually unless mentioned, ϵ-greedy method refers to ϵG3

- ϵG3
  - With probability ϵ, sample an arm uniformly at random; with probability $1 - \epsilon$, sample an arm with the highest empirical mean.

Slide from Prof. Shivaram Kalyanakrishnan's course: "Foundations of Intelligent and Learning Agents."

ε-greedy Algorithms

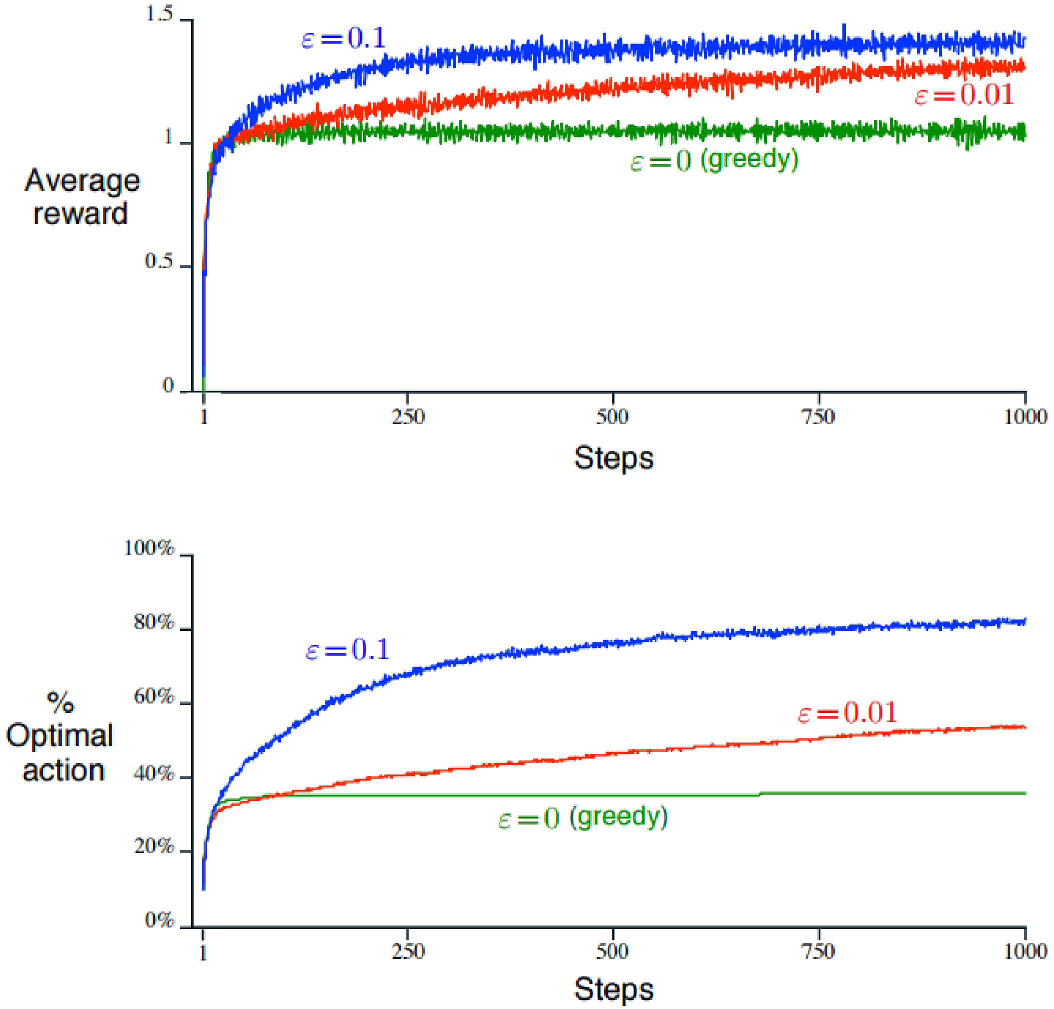


Figure 2.2: Average performance of ε-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 runs with different bandit problems. All methods used sample averages as their action-value estimates.

From the textbook: "Reinforcement Learning an Introduction", by Sutton and Barto

ϵ-greedy Algorithms

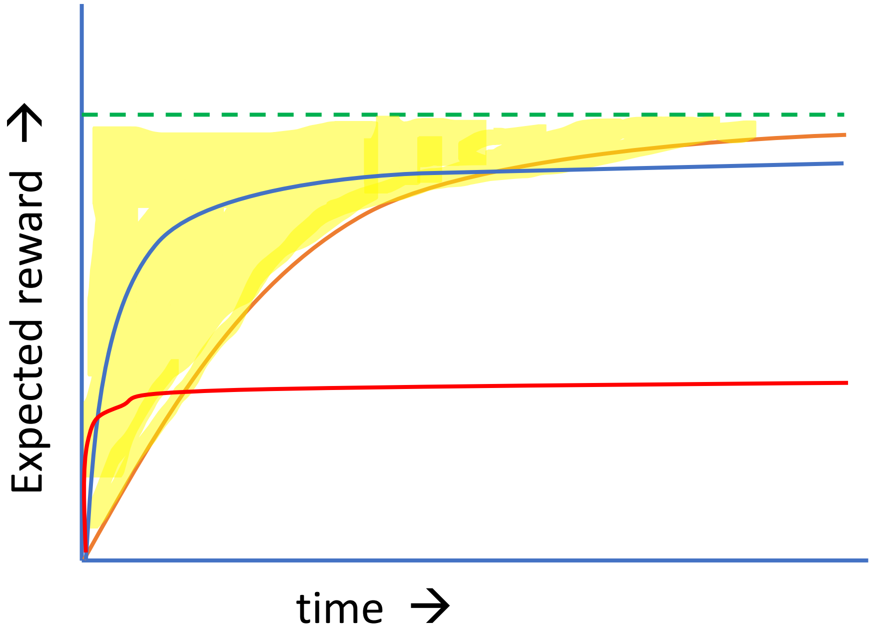
- Is ϵG2 better than ϵG1 ?
- What is the probability of picking an arm in each of these algorithms ?
- How can ϵ-greedy algorithms be written in the form of P(arm | history) ?

Performance Metrics

- **Asymptotic Correctness**
  - Gives a guarantee that eventually the algorithm will be selecting an arm that has the highest pay-off.
    - As T tends to infinity,
  - Will ϵ-greedy algorithm give asymptotic correctness?
    - Keep decreasing the ϵ value with time (cooling).

Performance Metrics

- **Regret Optimality**
  - "disappointed over (something that one has done or failed to do)" – *definition of regret from Oxford Dictionary*



  - Regret optimality refers to increasing the total reward one gets over the process of learning.