Reinforcement Learning Fundamentals

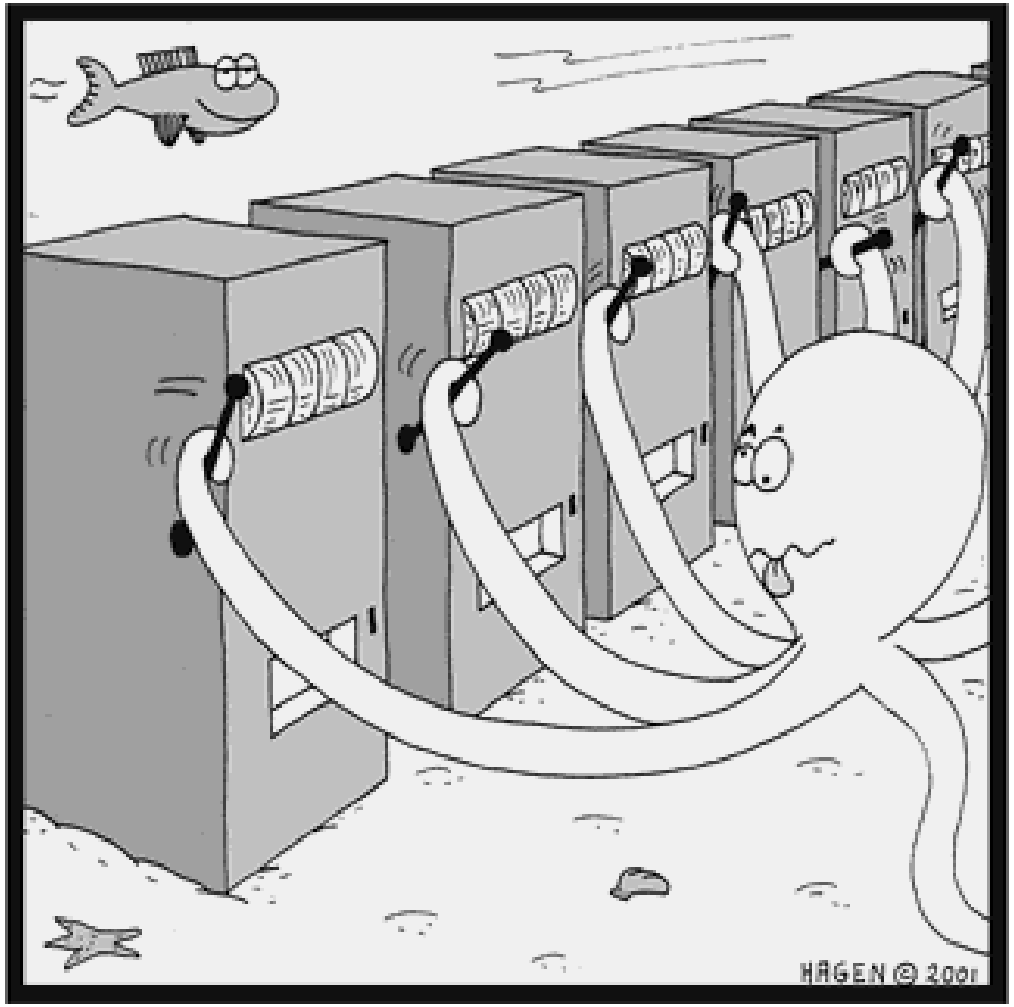
Lecture 7: Multi-armed Bandit

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

In today's class…

- Performance metrics
  - Correctness
  - Convergence
  - Sample Efficiency

- Solution methods
  - SoftMax
  - Upper Confidence Bound



Multi-armed Bandit

- n-arm bandit problem is to learn to preferentially select a particular action (arm) from a set of n actions $(1, 2, 3, \ldots, n)$
- Each selection results in Rewards derived from the respective probability distribution
- Arm i has a reward distribution with mean μ_i and

$$\mu^* = \max\{\mu_i\}$$

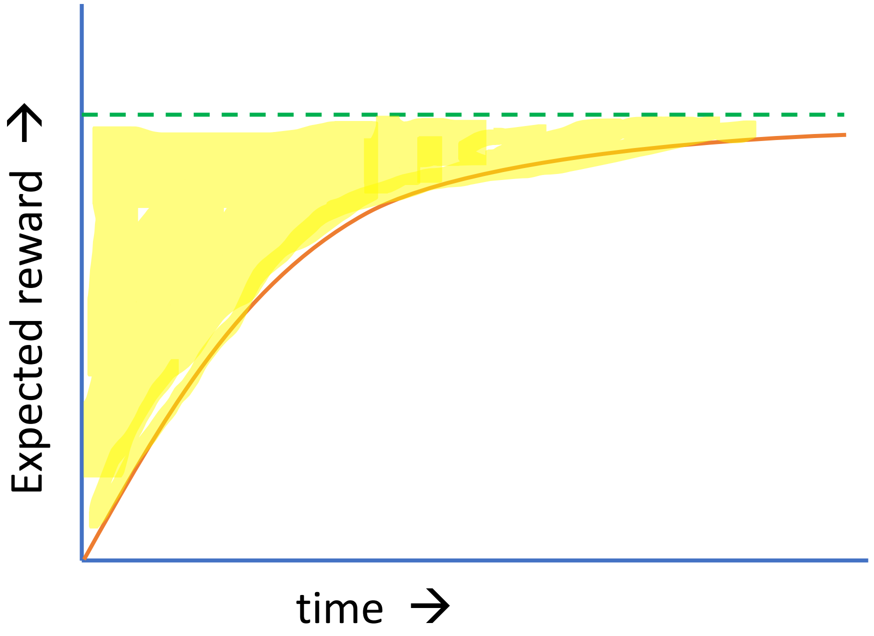
Performance Metrics

- **Asymptotic Correctness –** Identify the correct arm eventually
  - Gives a guarantee that eventually the algorithm will be selecting an arm that has the highest pay-off.
    - As T tends to infinity.

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 → Minimize regret while learning

Performance Metrics

- **PAC Optimality (Probably Approximately Correct)**

  - Approximately right in Bandit setup:
    - The arm suggested has expected payoff close to the expected payoff of the best arm.

  - Probably – It is either approximately correct or not!

  - (ε, δ) – PAC framework
    - – Identification of an ε-optimal arm with probability $1 - \delta$
    - – ε-Optimal: Mean of the selected arm satisfies

$$\mu > \mu^* - \varepsilon$$

$$\{ P (q_*(a) \geq (q_*(a_*) - \varepsilon)) \} \geq (1 - \delta)$$

*True expected value of an action **a***

True expected value of the best action a_*

Performance Metrics

- Asymptotic Correctness → measures **Correctness** of the solution
- Regret Optimality → measures the rate of **Convergence** of the solution
- PAC Optimality (Probably Approximately Correct) → **Sample efficiency** : want to minimize the sample size. Not very much used for practically implementable algorithm.

Solution Approaches

Exploration methods:

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

- **Epsilon Greedy:** Select an arm $A_t \doteq \underset{a}{\arg\max}\, Q_t(a)$ with probability (1- ε) and select any arbitrary arm with probability ε.

  - **Some problems:**
    Even if we know that for a certain a_i, $Q_t(a_i) << Q_t(a_*)$, we still sample a_i with a fixed probability.

      - Wasted trials
      - Affects the regret / increased regret

Solution Approaches

Exploration methods:

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

- **Epsilon Greedy:** Select an arm $A_t \doteq \underset{a}{\arg\max}\, Q_t(a)$ with probability (1- ε) and select any arbitrary arm with probability ε.

- **SoftMax:**
  - Converts a set of values into probability distribution.
  - Select arms with probability proportional to the current value estimates.

$$\pi_t(a_i) = \frac{\exp(Q_t(a_i)/\tau)}{\sum_j \exp(Q_t(a_j)/\tau)}$$

Temperature parameter τ

- Asymptotic convergence guarantees

Other Approaches

- **Median Elimination** (Even-Dar et at., 2006)
- **Upper Confidence Bound** (UCB by Auer et al., 1998)
- **Thompson Sampling** (Chappelle & Li, 2001)

Incremental Value-function

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

For a given action, $Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$

$$\begin{aligned} Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\ &= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\ &= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\ &= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\ &= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\ &= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big], \end{aligned}$$

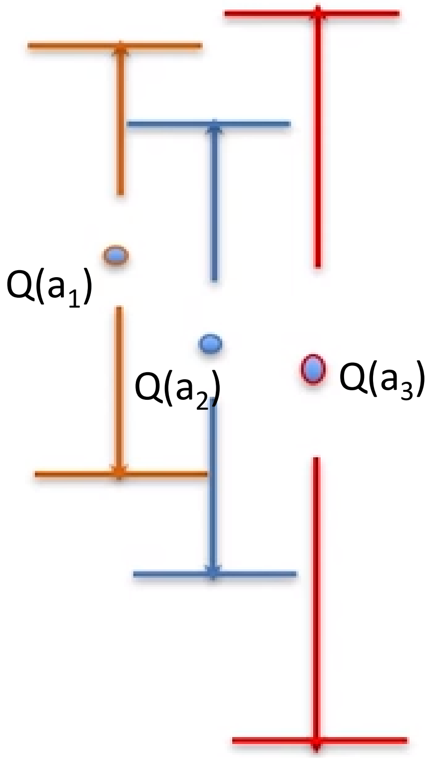
Simple ε-greedy,

Initialize, for $a = 1$ to k:
  $Q(a) \leftarrow 0$
  $N(a) \leftarrow 0$

Loop forever:
  $A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$
  $R \leftarrow bandit(A)$
  $N(A) \leftarrow N(A) + 1$
  $Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\big[R - Q(A)\big]$

$$NewEstimate \leftarrow OldEstimate + StepSize \Big[Target - OldEstimate\Big]$$

Upper Confidence Bound Action Selection



- **SoftMax** will still allocate probability to $\boldsymbol{a_2}$ and $\boldsymbol{a_3}$ as the values are close to $\boldsymbol{a_1}$. Even after convergence of value function.

- The **confidence interval** is the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or re-sample the population in the same way.

- The **confidence level** is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of the confidence interval.

- **UCB suggests**: Be greedy with respect to the upper confidence bound.

Upper Confidence Bound Action Selection
(UCB by Auer et al., 1998)

This term is a measure of the **uncertainty** or **variance** in the estimate of a's value.

$$A_t \doteq \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

- ***ln t*** denotes the natural logarithm of ***t***
- $N_t(a)$ denotes the number of times that action ***a*** has been selected prior to time ***t***
- The number ***c > 0*** controls the degree of exploration