# NCHU-2025-spring-drl-hw3

CHI HUNG YANG

May 2025

## 1 Introduction

This document describes three commonly used Multi-Armed Bandit (MAB) strategies in reinforcement learning: the $\varepsilon$-Greedy algorithm, the Upper Confidence Bound (UCB) algorithm, and the Softmax (Boltzmann) algorithm. Each balances exploration and exploitation to maximize total rewards over a limited number of decisions.

## 2 $\varepsilon$-Greedy Algorithm

### 2.1 Algorithm Overview

The $\varepsilon$-Greedy algorithm is a straightforward exploration–exploitation strategy. At each decision round $t$:

- With probability $1 - \varepsilon$, select the arm with the highest estimated average reward (exploitation).
- With probability $\varepsilon$, select an arm uniformly at random (exploration).

Here, $\varepsilon \in [0, 1]$ represents the exploration rate, typically chosen between 0.01 and 0.1 based on the number of arms and total decision rounds.

### 2.2 Action Selection Probability

In round $t$, the probability of selecting arm $i$ is given by:

$$P\big(a_t = i\big) = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{K}, & \text{if } i = \arg\max_j \hat{\mu}_j(t-1), \\ \dfrac{\varepsilon}{K}, & \text{otherwise,} \end{cases}$$

where:

- $K$ is the total number of arms.
- $\hat{\mu}_j(t-1)$ is the estimated average reward of arm $j$ up to round $t-1$.

### 2.3 Reward Update Rule

When arm $i$ is selected at round $t$ and reward $r_t$ is observed, its pull count and estimated average reward are updated as follows:

$$n_i(t) = n_i(t-1) + 1, \tag{1}$$

$$\hat{\mu}_i(t) = \hat{\mu}_i(t-1) + \frac{1}{n_i(t)}\big(r_t - \hat{\mu}_i(t-1)\big). \tag{2}$$

### 2.4 Summary of $\varepsilon$-Greedy

The $\varepsilon$-Greedy algorithm is widely used in MAB and other reinforcement learning scenarios requiring exploration–exploitation trade-offs due to its simplicity and low computational cost. Adjusting $\varepsilon$ or using decay strategies can further improve long-term performance.

## 3 Upper Confidence Bound (UCB) Algorithm

### 3.1 Algorithm Overview

The UCB algorithm applies the principle of optimism in the face of uncertainty by constructing an upper confidence bound for each arm's estimated value. At each round $t$, the arm with the highest bound is selected, naturally balancing exploration and exploitation.

### 3.2 UCB Index Formula

For arm $i$ at round $t$, the UCB index is defined as:

$$\mathrm{UCB}_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{2\ln t}{n_i(t)}},$$

where:

- $\hat{\mu}_i(t)$ is the estimated average reward of arm $i$ up to round $t$.
- $n_i(t)$ is the number of times arm $i$ has been selected before round $t$.
- $t$ is the total number of decision rounds so far.

### 3.3 Pseudocode

```
Initialize: for each arm i, set n_i = 0, \hat\mu_i = 0
// Ensure each arm is tried once
for i = 1 to K do
  pull arm i, observe reward r
  n_i = 1, \hat\mu_i = r
```

```
end for

for t = K+1 to T do
  for i = 1 to K do
    UCB_i = \hat\mu_i + sqrt((2*ln t)/n_i)
  end for
  select arm i* = argmax UCB_i
  pull arm i*, observe reward r
  n_{i*}++
  \hat\mu_{i*} += (r - \hat\mu_{i*})/n_{i*}
end for
```

### 3.4 Theoretical Guarantee

UCB achieves a logarithmic regret bound:

$$\mathbb{E}[R(T)] = O\Big(\sum_{i:\Delta_i>0} \frac{\ln T}{\Delta_i}\Big),$$

where $\Delta_i = \mu^* - \mu_i$ is the gap between the optimal arm's mean reward and arm $i$'s mean reward.

## 4 Softmax (Boltzmann) Algorithm

### 4.1 Algorithm Overview

The Softmax algorithm assigns a selection probability to each arm based on a Boltzmann distribution over estimated values, allowing smooth trade-off between exploration and exploitation.

### 4.2 Temperature Parameter

A temperature parameter $\tau > 0$ controls the randomness:

- As $\tau \to 0^+$, the selection becomes more greedy (higher-value arms get almost all probability).
- As $\tau \to \infty$, the selection becomes more uniform across arms.

### 4.3 Action Selection Probability

In round $t$, the probability of selecting arm $i$ is:

$$P_i(t) = \frac{\exp\big(\hat{\mu}_i(t)/\tau\big)}{\sum_{j=1}^{K} \exp\big(\hat{\mu}_j(t)/\tau\big)}.$$

### 4.4 Reward Update Rule

After selecting arm $i$ and observing reward $r_t$, update its estimate:

$$\hat{\mu}_i(t) = \hat{\mu}_i(t-1) + \frac{1}{n_i(t)}\big(r_t - \hat{\mu}_i(t-1)\big),$$

where $n_i(t)$ is the pull count including this round.

### 4.5 Summary of Softmax

Softmax exploration provides a differentiable selection mechanism that smoothly interpolates between greedy and random policies, and can be tuned via the temperature parameter for desired exploration behavior.

## 5 Conclusion

This document presented the $\varepsilon$-Greedy, UCB, and Softmax algorithms for solving the Multi-Armed Bandit problem, including their motivations, formulas, and key properties. These strategies are fundamental tools in reinforcement learning for efficient decision-making under uncertainty.