Formulas of Four Multi-Armed Bandit Algorithms

1. Epsilon-Greedy Algorithm

Description: Epsilon-Greedy balances exploration and exploitation. With probability $1 - \varepsilon$, it selects the action with the highest estimated reward. With probability ε, it explores randomly.

$$A_t = \begin{cases} \arg\max_a Q_t(a), & \text{with probability } 1 - \varepsilon \\ \text{randomly select } a \in \mathcal{A}, & \text{with probability } \varepsilon \end{cases}$$

Explanation:

- A_t: the action selected at time t.
- $Q_t(a)$: estimated reward for action a at time t.
- ε: exploration probability.
- $\mathcal{A}$: set of all possible actions.

2. UCB (Upper Confidence Bound) Algorithm

Description: UCB selects actions based on the upper confidence bound, balancing current reward estimates with uncertainty.

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

Explanation:

- $Q_t(a)$: average reward of action a.
- $N_t(a)$: number of times action a has been selected before time t.
- t: current round.
- c: exploration constant.

3. Softmax Algorithm

Description: Softmax assigns a probability to each action using the softmax function, allowing smoother exploration based on reward estimates.

$$P(a) = \frac{\exp(Q_t(a)/\tau)}{\sum_{b \in \mathcal{A}} \exp(Q_t(b)/\tau)}$$

Explanation:

- $P(a)$: probability of selecting action a.
- $Q_t(a)$: estimated reward of action a.
- τ: temperature parameter; higher values encourage more exploration.
- $\mathcal{A}$: set of all possible actions.

4. Thompson Sampling Algorithm

Description: Thompson Sampling uses Bayesian inference to maintain uncertainty and samples from posterior distributions to select actions.

$$\theta_a \sim p(\theta_a \mid \text{history}), \quad A_t = \arg\max_a \theta_a$$

Explanation:

- θ_a: a sampled parameter from the posterior distribution for action a.
- $p(\theta_a \mid \text{history})$: posterior distribution of action a based on past observations.
- A_t: action selected with the highest sampled value.