Lab 1 - Bandits, ε-Greedy & UCB

Environments and Core Concepts

Term	Definition
Reinforcement Learning (RL)	The overarching field modeling how an agent learns to make better decisions over time by balancing **exploration** and **exploitation** to maximize overall reward.
Multi-Armed Bandits (MAB)	A simulation environment used for news recommendation where each article (arm) has an unknown click probability, and the goal is to learn the best article through interaction.
Contextual Bandits	An extension of the MAB problem where the algorithm observes a user profile (context, *x(t)*)before choosing an article. The goal is to maximize the expected number of clicks by adapting to user features.
Exploration	The act of trying different actions (arms) to gather information about the environment.
Exploitation	The act of choosing the action that currently seems best, based on gathered information.
Context (*x(t)*)	A feature vector observed at time *t* that represents user characteristics, such as age, political interest, or mobile usage.
Arm	Represents an article recommendation choice available to the agent. The lab considers *k=4* possible articles (arms).

Algorithms

Algorithm	Description	Key Components/Function
Epsilon-Greedy (ε)	An exploration strategy for MAB that uses a fixed probability (ε)	Chooses a **random arm** with probability ε; otherwise, chooses the arm with the **highest estimated value ($Q(a)$)**.

	to balance exploration and exploitation.	
Upper Confidence Bound (UCB)	An exploration strategy for MAB that uses confidence bounds to promote systematic exploration of uncertain arms.	Selects the arm a_t that maximizes the expression $\left[Q(a) + c \cdot \sqrt{\frac{\log t}{N(a)}}\right]$.
Random Action Policy	A simple benchmark policy where decisions are made randomly, used for comparison against learning algorithms.	Chooses a random arm at every time step.
Contextual Epsilon-Greedy	Extends the ε-greedy idea to the contextual setting, using a linear model ($\thetâ_a$) for each arm to predict rewards based on the context *x(t)*.	Chooses a **random arm** with probability ε; otherwise, chooses the arm that maximizes the predicted reward $r̂_a(x_t) = \thetâ_a^\top x_t$.
Linear UCB (LinUCB)	A contextual bandit algorithm that models the expected reward as a linear function of the context	$E[r_t x_t, a] = \theta_a^\top x_t$ It is designed to explore more systematically than ϵ-greedy.

Evaluation Metrics

Metric	Definition	Mathematical Notation
Average Reward	The total reward obtained up to time T, divided by T. A higher value indicates the algorithm is making better decisions on average.	$\bar{r}T = \frac{1}{T} \sum t = 1^T r_t$
Regret (Cumulative Regret)	Measures how much the algorithm loses compared to always selecting the true optimal arm ($\mu^\ast$). It quantifies the efficiency of learning.	$\mathrm{Regret}T = \sum t = 1^T (\mu^* - \mu_{a_t})$

Main Variables and Parameters

Variable/Parameter	Description	Algorithms Used In
t	Time step (e.g., *t=1* → 1000$).	MAB, UCB, Contextual Bandits
r(t)	**Reward** obtained at time step *t*. It is binary, where *1* is a user click and *0* is an ignore.	All algorithms, Evaluation
Q(a)	The **estimated mean reward** (estimated click probability) for arm *a*. It is updated online.	Epsilon-Greedy, UCB
N(a)	The **number of times arm *a* has been selected** (pulled).	Epsilon-Greedy, UCB
ε (Epsilon)	**Exploration rate** hyperparameter for ε-Greedy policies. It is the probability of choosing a random arm. (Example values: 0.1, 0.01, 0.2).	Epsilon-Greedy, Contextual Epsilon-Greedy
	Exploration hyperparameter for the UCB algorithm. Controls the degree of exploration (multiplies the confidence term). (Example values: $1, 2, 5$).	UCB
mu(i)	The **true expected reward** (click probability) of arm *i*.	Evaluation (Regret calculation)

mu^\ast	The **expected reward of the best arm** $(\max_i \mu_i)$.	Evaluation (Regret calculation)
d	The dimensionality (number) of the **contextual features** (e.g., *d=6*).	Contextual Bandits
theta_a	The **true (unknown) parameter vector** for article *a* that determines the click probability given the context *x(t)*.	Contextual Bandits (Environment definition)
thetâ_a	The **estimated linear model parameters** for arm *a*.	Contextual Epsilon-Greedy, LinUCB
A_a	A matrix maintained by LinUCB, used in computing the uncertainty bound.	LinUCB
b_a	A vector maintained by LinUCB, used in computing the estimated parameters *thetâ_a*.	LinUCB
alpha	**Exploration hyperparameter** for LinUCB. Controls the degree of exploration by scaling the confidence bound.	LinUCB
lambda	**Regularization parameter** for LinUCB, used to ensure stability in parameter updates.	LinUCB

Insight Analogy

Understanding these algorithms is like trying to find the best restaurant in a new city.

- **Epsilon-Greedy** is like choosing your favourite restaurant 90% of the time (exploitation) but forcing yourself to try a random new place 10% of the time (exploration). It guarantees you won't miss out forever, but your exploration is often wasted on bad random choices.
- **UCB/LinUCB** is like choosing a restaurant based on its currently high rating *and* how little you know about it. If a restaurant has an average rating, but only three people have ever reviewed it, UCB encourages you to try it because its true quality is highly uncertain. This makes the exploration systematic and efficient, focusing resources where information gain is highest.