# Explore-then-Commit and Upper Confidence Bound Algorithm

Henry Vu

Jan 22, 2024

# Today’s agenda

- Recap
- Optimism in the face of uncertainty
- Sub-gaussian distribution
- Explore-Then-Commit algorithm
- Upper Confidence Bound (UCB) algorithm
- Adversarial Bandits

# Recap: Regret

- **Regret**: The cost of not always playing the best arm.
- The regret $R_n$ after n plays $I_1, I_2, \ldots, I_n$ is defined by

$$R_n = \max_{i=1,\ldots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t} \ .$$

# Recap: Stochastic Multi-armed Bandits

- The rewards of arm i are i.i.d according to a fixed probability distribution $v_1$, $v_2$, …, $v_K$ on [0, 1]. These distributions are <u>unknown</u> to the algorithm.

- Let:

$$\mu^* = \max_{i=1,\ldots,K} \mu_i \qquad \text{and} \qquad i^* \in \operatorname*{argmax}_{i=1,\ldots,K} \mu_i \ .$$

- In the stochastic setting, pseudo-regret can be written as

$$\widetilde{R}_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} \mu_{I_t}\right]$$

# Recap: Another perspective of regret

- Let $\Delta_i = \mu^* - \mu_i$, and let $T_i(s)$ denote the number of times the algorithm chose arm i on the first s rounds. Regret is also a function of $T_i(s)$ and $\Delta_i$.

$$\overline{R}_n = \left(\sum_{i=1}^{K} \mathbb{E}\, T_i(n)\right) \mu^* - \mathbb{E} \sum_{i=1}^{K} T_i(n)\mu_i = \sum_{i=1}^{K} \Delta_i \,\mathbb{E}\, T_i(n)$$

- We now minimize the weighted sum $\mathbb{E}[T_i(n)]$, where the weights are the respective action gaps.

# Recap: simple heuristics

- **Naive**:

  Greedily plays the arm with the highest empirical mean ⇒ may get stuck due to lack of exploration, regret is linear n.

  Play all arms an equal number of times ⇒ pure exploration, regret is linear in n
- **e-greedy**:

  Exploitation: greedily plays the arm with the highest empirical mean (observed rewards) so far with probability 1-$\epsilon$,

  Exploration: plays a random arm (including empirically best arm) with probability $\epsilon$.

  ⇒ O(log(n)) regret

# Sub-Gaussian Distribution

- To show the concentration results, a fundamental assumption is that reward $X_{i,t}$ follows a sub-gaussian distribution.

- Random variable X follows a $\sigma^2$-subgaussian distribution if for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

# Sub-Gaussian distribution (cont.)

**Lemma:** Suppose that $X$ is $\sigma_1^2$-subgaussian and $X_1$ and $X_2$ are independent and $\sigma_2^2$-subgaussian respectively, then:

- $\mathbb{E}[X] = 0$ and $\mathbb{V}[X] \leq \sigma^2$.
- $cX$ is $c^2\sigma^2$-subgaussian for all $c \in \mathbb{R}$.
- $X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$-subgaussian.

**Theorem:** If $X$ is $\sigma^2$-subgaussian, then

$$\mathbb{P}(X \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

# Hoeffding's Bound

- Combining the above Theorem and Lemma, we get

$$\hat{\mu} - \mu = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \text{ is } \frac{\sigma^2}{n}\text{-subgaussian.}$$

Assume that $X_i - \mu$ are independent, $\sigma^2$-subgaussian random variables. Then, their average $\hat{\mu}$ satisfies

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$
$$\mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

# Explore-then-Commit (ETC)

- Each arm is explored m times, then fully commit to the arm with the highest empirical mean. For simplicity, assuming $X_t - \mathbb{E}[X_t]$ is 1-subgaussian.

- Formally,

$$I_t = \begin{cases} i, & \text{if (t mod K)} + 1 = i \text{ and } t \leq mK; \\ \underset{i}{\operatorname{argmax}}\, \hat{\mu}_i(mK), & t > mK, \end{cases}$$

# ETC Regret

- $$R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$$
  $$= m \sum_{i=1}^{K} \Delta_i + (n - mK) \sum_{i=1}^{K} \Delta_i \mathbb{P}\left(i = \operatorname*{argmax}_j \hat{\mu}_j(mK)\right)$$
  $$\leq m \sum_{i=1}^{K} \Delta_i + (n - mK) \sum_{i=1}^{K} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)$$

- If m is large, the first term will be too large.
- If m is too small, then the probability that the algorithm commits to the wrong arm will grow and the second term becomes too large.

# ETC Regret

- For K = 2, $\Delta_1 = 0$ and $\Delta_2 = \Delta$ and choose minimizing $m = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil$

$$R_n \leq \Delta + \frac{4}{\Delta}\left(1 + \log\left(\frac{n\Delta^2}{4}\right)\right)$$

- Notice that $R_n \leq n\Delta$, we can take the minimum of the two bounds so that

$$R_n \leq \min\left\{n\Delta,\ \Delta + \frac{4}{\Delta}\left(1 + \log\left(\frac{n\Delta^2}{4}\right)\right)\right\}$$

# Optimism in the face of Uncertainty

- Random exploration (i.e $\epsilon$-greedy) might take inefficient actions. One approach is to decrease $\epsilon$ over time, the other is to be *optimistic* about actions with *high uncertainty*.

- **Intuition**: If the optimism was justified, the algorithm is acting optimally. If the optimism was not, the algorithm learns the true payoff after a sufficient number of time steps.
  ⇒ UCB algorithm: $I_t$ = argmax$_i$ ($u_i$_hat + bound)

# UCB Algorithm

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$
$$\mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

- By Hoeffding's inequality,

$$\mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \leq \delta$$

- UCB policy is as follows:

$$I_t = \begin{cases} \operatorname*{argmax}_i \left(\hat{\mu}_i(t-1) + \sqrt{\frac{2\log f(t)}{T_i(t-1)}}\right), & \text{if } t > K; \\ t, & \text{otherwise.} \end{cases}$$

The term inside argmax is called the **index** of arm i

# UCB Algorithm (cont).

$$\underset{i}{\operatorname{argmax}} \left( \underbrace{\hat{\mu}_i(t-1)}_{\color{red}{\text{Exploitation}}} + \underbrace{\sqrt{\frac{2 \log f(t)}{T_i(t-1)}}}_{\color{green}{\text{Exploration}}} \right)$$

- $T_i(t)$ small ⇒ *larger* bound ⇒ uncertain, needs exploration
- $T_i(t)$ large ⇒ *smaller* bound ⇒ more confident to exploit

# UCB Regret

**Corollary (Lattimore & Szepesvari):** The regret of UCB is bounded by

$$R_n \leq \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{1}{\Delta_i}\left(8\log f(n) + 8\sqrt{\pi \log f(n)} + 28\right)\right).$$

and in particular there exists some universal constant $C > 0$ such that for all $n \geq 2$,

$$R_n \leq \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{C\log n}{\Delta_i}\right).$$

- This regret bound is unimprovable

- Proof: Bound $\mathbb{E}[T_i(n)]$

# UCB Regret Proof Sketch

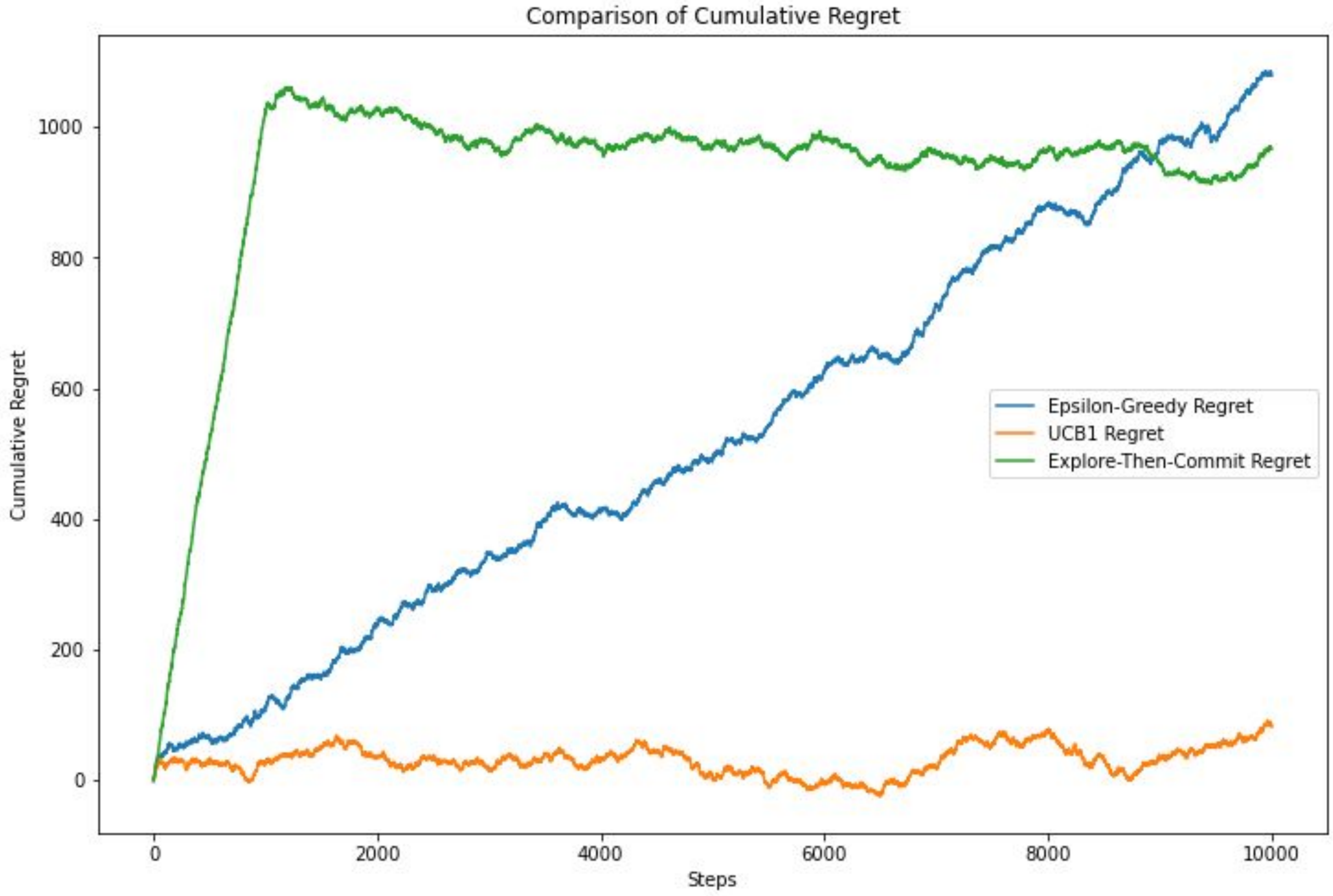
- To estimate $\mathbb{E}[T_i(n)]$, notice that arm i is chosen when $UCB_i$ is either too high OR $UCB_1$ is too low. In math terms:

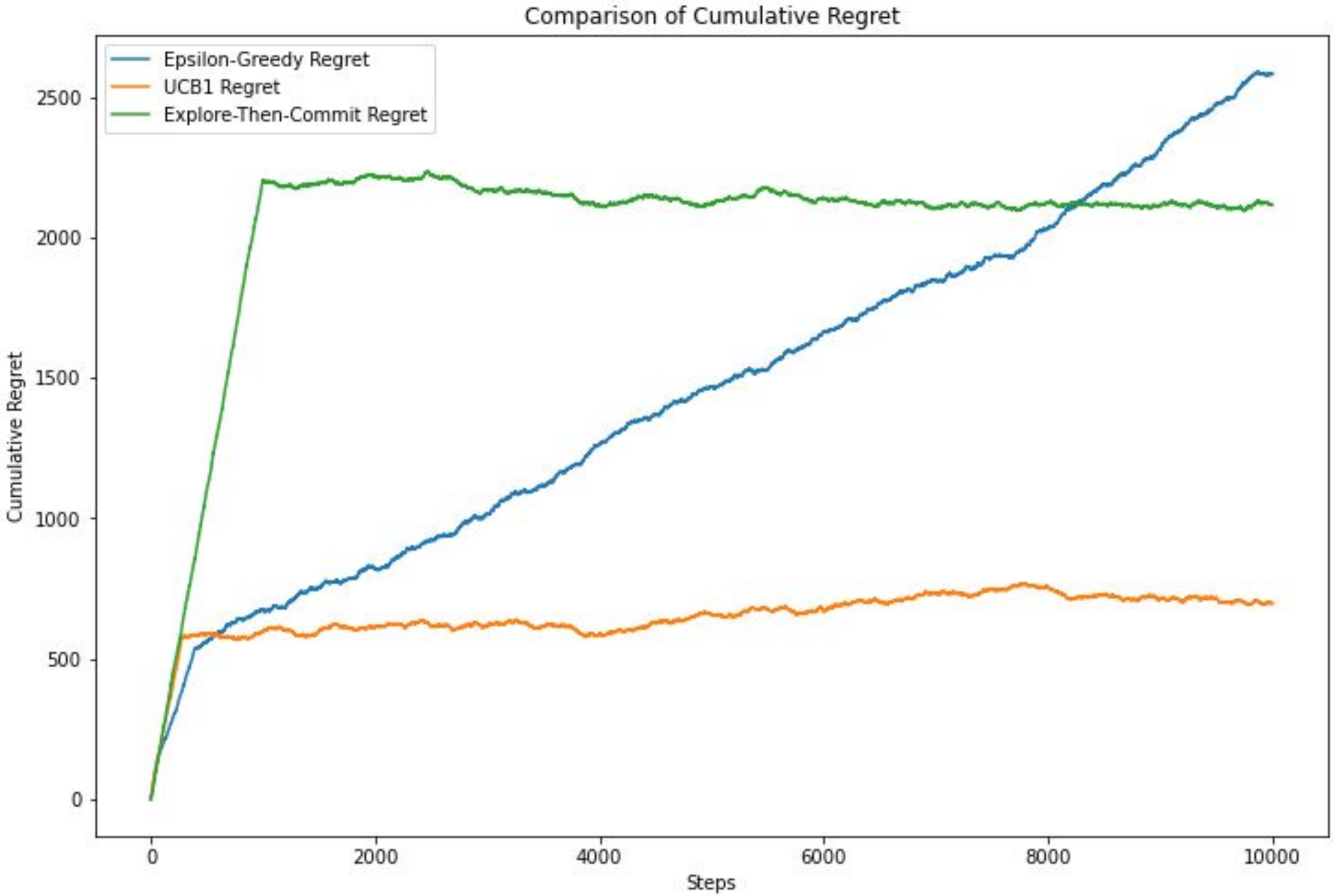
$$
\begin{aligned}
T_i(n) &= \sum_{t=1}^{n} \mathbb{I}\{A_t = i\} \\
&\le \sum_{t=1}^{n} \left\{ \mathbb{I}\left\{ \hat{\mu}_1(t-1) + \sqrt{\frac{2\log f(t)}{T_1(t-1)}} \le \mu_1 - \varepsilon \right\} \right\} \\
&+ \sum_{t=1}^{n} \left\{ \mathbb{I}\left\{ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log f(t)}{T_i(t-1)}} \ge \mu_1 - \varepsilon \text{ and } A_t = i \right\} \right\}.
\end{aligned}
$$

# Regret

For stochastic bandits, so far, we have seen:

- **$\epsilon$-greedy**: O(n). O(log(n)) if $\epsilon$ is a decreasing function of time
- **Explore-then-Commit**: O(log(n)). However, this requires prior knowledge or assumptions about the rewards distribution
- **Upper Confidence Bound**: Balances exploration and exploitation. O(log(n))


Comparison of Cumulative Regret
Cumulative Regret
Steps
0
200
400
600
800
1000
0
2000
4000
6000
8000
10000
Epsilon-Greedy Regret
UCB1 Regret
Explore-Then-Commit Regret


Comparison of Cumulative Regret
Epsilon-Greedy Regret
UCB1 Regret
Explore-Then-Commit Regret
Cumulative Regret
Steps
0
500
1000
1500
2000
2500
0
2000
4000
6000
8000
10000

# Adversarial Bandits

Henry Vu

Feb 9, 2024

# Stochastic Bandits

**Given** $A = \{1, 2, \ldots, K\}$ the set of action and (possibly) number of rounds $n \geq K$

**for** $t = 1, 2, \ldots, n$ **do**:

Algorithm pulls arm $I_t \in A$

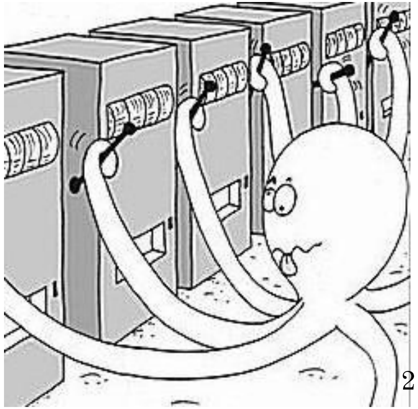
A reward vector $(X_{1, t}, X_{2, t}, \ldots, X_{n, t})$ is generated, usually scaled to $[0, 1]$

Algorithm observes reward $X_{It, t}$

**end for**

**Goal**: Minimizing the regret

Simple formulation, but no known tractable optimal solution



# Adversarial Bandits: Problem Settings

**Given** A = {1, 2, …, K} the set of action and (possibly) number of rounds $n \geq K$
**for** $t = 1, 2, \ldots, n$ **do**:

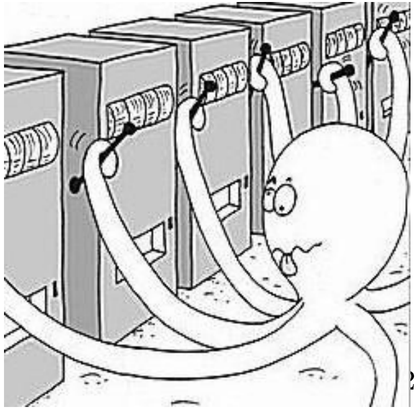
Algorithm pulls arm $I_t \in A$

A reward vector $(x_{1,t}, x_{2,t}, \ldots, x_{n,t})$ is given by the adversary (no underlying distribution)

Algorithm observes reward $x_{I_t,t}$

**end for**

**Goal**: Minimizing the <u>regret</u>



# Why adversarial?

- No assumptions about reward distribution ⇒ more robust algorithms

- Why regret vs **fixed arm** while losses are changing?

  ⇒ switching/dynamic regret

- For now, we still study the **static regret**

# Need for Randomization

Example:

- If algorithm chooses action A, $\text{reward}_A = 0$, $\text{reward}_B = 1$
- If algorithm chooses action B, $\text{reward}_A = 1$, $\text{reward}_B = 0$

⇒ Linear regret

⇒ The algorithm needs to randomize its actions to achieve sublinear regret

# Adversarial vs Stochastic

- In stochastic bandits, total expected reward to compared to the maximum *expected* reward

- In adversarial bandits, total expected reward to compared to the maximum reward. If randomization is present, compared to the expected maximum reward.

$$
\begin{aligned}
R_n(\pi, \nu) &= \max_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^{n} (X_{t_i} - X_{t_{I\,t}})\right] \\
&\leq \mathbb{E}\left[\max_{i \in [K]} \sum_{t=1}^{n} (X_{t_i} - X_{t_{I\,t}})\right] \\
&= \mathbb{E}\left[R_n(\pi, X)\right] \leq R_n^*(\pi),
\end{aligned}
$$

# Next Week

Exp3 Algorithm

Thank you!