Duelling Bandits

IE617: Course Project

Satyankar Chandra (22B0967) Anilesh Bansal (22B0928)
Siddharth Verma (22B2153) Harsh Jitendrakumar (23N0452)

Indian Institute of Technology Bombay

May 5, 2025
10:00 - 10:20 AM

Guide: Prof. Manjesh Kumar Hanawal

Presentation Overview

Introduction

Classical Multi Armed Bandit Problem :-

- A set of K arms, each with an unknown reward distribution.
- Maximize total expected reward (or minimize regret)
- At each time step, select an arm to pull and observe the reward.

What if we cannot get the reward of an arm directly?

Absolute rewards have no natural meaning, but we can still compare the arms in pairs. How to find the best arm?

This leads to the **duelling bandit problem**.

Very common in real world applications, hence an important problem.

The K-armed Duelling Bandit Problem [1]

- A set of K arms $\mathcal{B} = \{b_1, b_2, \ldots, b_K\}$, each with an unknown reward distribution.
- At each time step, select a pair of arms $(b_i^{(t)}, b_j^{(t)})$ to compare and observe the winner.
- Use the results of the comparisons to find the best arm b^*.
- Again minimize a (new notion of) regret.

Probability of b_i winning against b_j in a "duel" is denoted by $\mathbb{P}(b_i > b_j) = \frac{1}{2} + \epsilon(b_i, b_j)$ which is assumed to be stationary over time.

We assume that a total ordering $\succ$ exists over $\mathcal{B}$ such that $b_i \succ b_j$ implies $\epsilon(b_i, b_j) > 0$. Hence some best arm b^* exists.

[1] Formulated by Yue et al., Journal of Computer and System Sciences

Regret Model

Two types of regret are defined in the paper with respect to the best arm b^*:

1. **Strong Regret (R_T^{strong}):** This measures the expected preference gap between the best arm b^* and *both* arms chosen for comparison at each step.

$$R_T^{strong} = \sum_{t=1}^{T} \left(\epsilon(b^*, b_i^{(t)}) + \epsilon(b^*, b_j^{(t)}) \right) \tag{1}$$

2. **Weak Regret (R_T^{weak}):** This measures the expected preference gap between the best arm b^* and the *better* of the two arms chosen for comparison at each step.

$$R_T^{weak} = \sum_{t=1}^{T} \min \left\{ \epsilon(b^*, b_i^{(t)}), \epsilon(b^*, b_j^{(t)}) \right\} \tag{2}$$

Applications in Real World

The duelling bandit problem is particularly useful in scenarios where absolute rewards are difficult to obtain, but pairwise comparisons are feasible.

Some such instances include:

- **Online Advertising:** Selecting the best ad to show to users.
- **Recommendation Systems:** Choosing the best item to recommend to users.
- **Drug Discovery:** Finding the best drug from a set of candidates.
- **Clinical Trials:** Comparing different treatment options.
- **Sensory Testing:** Evaluating the best product based on user preferences.

The duelling bandit framework only requires binary feedback (that can be noisy) about the relative rewards of two strategies.

Assumptions about the Environment

We impose additional structure to the probabilistic comparisons $\mathbb{P}(b_i > b_j)$ to derive proper regret bounds.

- **Strong Stochastic Transitivity (SST):** For any triplet of bandits $b_i > b_j > b_k$, the preference margin between the best and worst is at least as large as the margins in the intermediate comparisons:

$$\epsilon(b_i, b_k) \geq \max\{\epsilon(b_i, b_j), \epsilon(b_j, b_k)\} \tag{3}$$

- **Stochastic Triangle Inequality (STI):** For any triplet of bandits $b_i \succ b_j \succ b_k$, the preference margin between the best and worst is no more than the sum of the intermediate margins:

$$\epsilon(b_i, b_k) \leq \epsilon(b_i, b_j) + \epsilon(b_j, b_k) \tag{4}$$

Commonly Satisfying Generative Models

These assumptions are not overly restrictive and are satisfied by common preference models, including:

- **Logistic/Bradley-Terry Model:** Each bandit b_i has an associated positive strength μ_i, and $\mathbb{P}(b_i > b_j) = \frac{\mu_i}{\mu_i + \mu_j}$.
- **Gaussian/Thurstone Model:** Each bandit b_i has an associated random utility $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$. Comparisons are made based on these utilities, $\mathbb{P}(b_i > b_j) = \mathbb{P}(X_i > X_j) = \mathbb{P}(X_i - X_j > 0)$. If X_i are independent Gaussians with equal variance, $X_i - X_j \sim \mathcal{N}(\mu_i - \mu_j, 2\sigma^2)$, and the assumptions hold.

The constraints only ensure that the preference margins (ϵ) behave in a consistent and predictable manner, allowing for effective learning and exploration strategies.

Algorithms for the K-armed Duelling Bandits

The high level idea is outlined in Algorithm 1 of the paper.

Algorithm 1 Explore Then Exploit Solution

1: Input: $T, \mathcal{B} = \{b_1, \ldots, b_K\}, EXPLORE$
2: $(\hat{b}, \hat{T}) \leftarrow EXPLORE(T, \mathcal{B})$
3: **for** $t = \hat{T} + 1, \ldots, T$ **do**
4: compare $\hat{b}$ and $\hat{b}$
5: **end for**

Figure: General Explore Then Exploit Framework

The main contribution of the paper is to provide a family of such *EXPLORE* algorithms which have logarithmic regret bounds and return $\hat{b} = b^*$ with high probability ofx $1 - \mathcal{O}(1/T)$.

Interleaved Filter 1 (IF1) Algorithm

The IF1 algorithm is the first such *EXPLORE* algorithm.

Algorithm 2 Interleaved Filter 1 (IF1)

Input: $T, \mathcal{B} = \{b_1, \ldots, b_K\}$
$\delta \leftarrow 1/(TK^2)$
Choose $\hat{b} \in \mathcal{B}$ randomly
$W \leftarrow \{b_1, \ldots, b_K\} \setminus \{\hat{b}\}$
$\forall b \in W$, maintain estimate $\hat{P}_{\hat{b},b}$ of $P(\hat{b} > b)$
$\forall b \in W$, maintain $1-\delta$ confidence interval $\hat{C}_{\hat{b},b}$ of $\hat{P}_{\hat{b},b}$
while $W \neq \emptyset$ **do**
  for $b \in W$ **do**
    compare $\hat{b}$ and b
    update $\hat{P}_{\hat{b},b}$, $\hat{C}_{\hat{b},b}$
  end for
  while $\exists b \in W$ s.t. $\left(\hat{P}_{\hat{b},b} > 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b}\right)$ **do**
    $W \leftarrow W \setminus \{b\}$
  end while
  if $\exists b' \in W$ s.t. $\left(\hat{P}_{\hat{b},b'} < 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b'}\right)$ **then**
    $\hat{b} \leftarrow b'$, $W \leftarrow W \setminus \{b'\}$ *//new round*
    $\forall b \in W$, reset $\hat{P}_{\hat{b},b}$ and $\hat{C}_{\hat{b},b}$
  end if
end while
$\hat{T} \leftarrow$ Total Comparisons Made
return $(\hat{b}, \hat{T})$

Key Ideas:

- Maintain empirical estimates $\hat{P}^{(t)}_{\hat{b},b}$ of $\mathbb{P}(\hat{b} > b)$
- Create $1-\delta$ confidence intervals for each arm b_i.
- In each round, duel between the best arm and the active arms.
- If some arm can never be the best arm, remove it from active set of arms.
- Update best arm if reqd and reset $\hat{P}^{(t)}_{\hat{b},b}$ and $\hat{C}^{(t)}_{\hat{b},b}$

Interleaved Filter 2 (IF2) Algorithm

Algorithm 3 Interleaved Filter 2 (IF2)

1: Input: $T, \mathcal{B} = \{b_1, \ldots, b_K\}$
2: $\delta \leftarrow 1/(TK^2)$
3: Choose $\hat{b} \in \mathcal{B}$ randomly
4: $W \leftarrow \{b_1, \ldots, b_K\} \setminus \{\hat{b}\}$
5: $\forall b \in W$, maintain estimate $\hat{P}_{\hat{b},b}$ of $P(\hat{b} > b)$
6: $\forall b \in W$, maintain $1-\delta$ confidence interval $\hat{C}_{\hat{b},b}$ of $\hat{P}_{\hat{b},b}$
7: **while** $W \neq \emptyset$ **do**
8: **for** $b \in W$ **do**
9: compare $\hat{b}$ and b
10: update $\hat{P}_{\hat{b},b}$, $\hat{C}_{\hat{b},b}$
11: **end for**
12: **while** $\exists b \in W$ s.t. $\left(\hat{P}_{\hat{b},b} > 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b}\right)$ **do**
13: $W \leftarrow W \setminus \{b\}$
14: **end while**
15: **if** $\exists b' \in W$ s.t. $\left(\hat{P}_{\hat{b},b'} < 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b'}\right)$ **then**
16: **while** $\exists b \in W$ s.t. $\hat{P}_{\hat{b},b} > 1/2$ **do**
17: $W \leftarrow W \setminus \{b\}$ *//pruning*
18: **end while**
19: $\hat{b} \leftarrow b'$, $W \leftarrow W \setminus \{b'\}$ *//new round*
20: $\forall b \in W$, reset $\hat{P}_{\hat{b},b}$ and $\hat{C}_{\hat{b},b}$
21: **end if**
22: **end while**
23: $\hat{T} \leftarrow$ Total Comparisons Made
24: return $(\hat{b}, \hat{T})$

Key Ideas:

- Similar to IF1, with an additional pruning step (lines 16 - 18)
- When a new potential best arm is found, directly remove all the empirically inferior arms (w.r.t $\hat{b}$) from the active set.
- Confidence bounds are not considered while pruning.
- Rest of the algorithm proceeds like IF1.

Theoretical Regret Bounds for IF1 and IF2

The regret bound (both strong and weak) for IF1 is given by:

Interleaved Filter 1 Total Regret

$$E[R_T^{IF1}] = \mathcal{O}\left(\frac{K \log K}{\epsilon_{1,2}} \log T\right)$$

The regret bound (both strong and weak) for IF2 is given by:

Interleaved Filter 2 Total Regret

$$E[R_T^{IF1}] = \mathcal{O}\left(\frac{K}{\epsilon_{1,2}} \log T\right)$$

Note that due to the pruning step, IF2 is more aggressive than IF1, and hence has a better regret bound. Same is observed in the experiments as well.

Experimental Findings

We implemented the IF1 and IF2 algorithms in Python and ran extensive simulations to verify the bounds presented in the paper.

Simulation Setup -

- **Model Type:** Gaussian/Thurstone Model with randomly generated means μ_i and variance 1

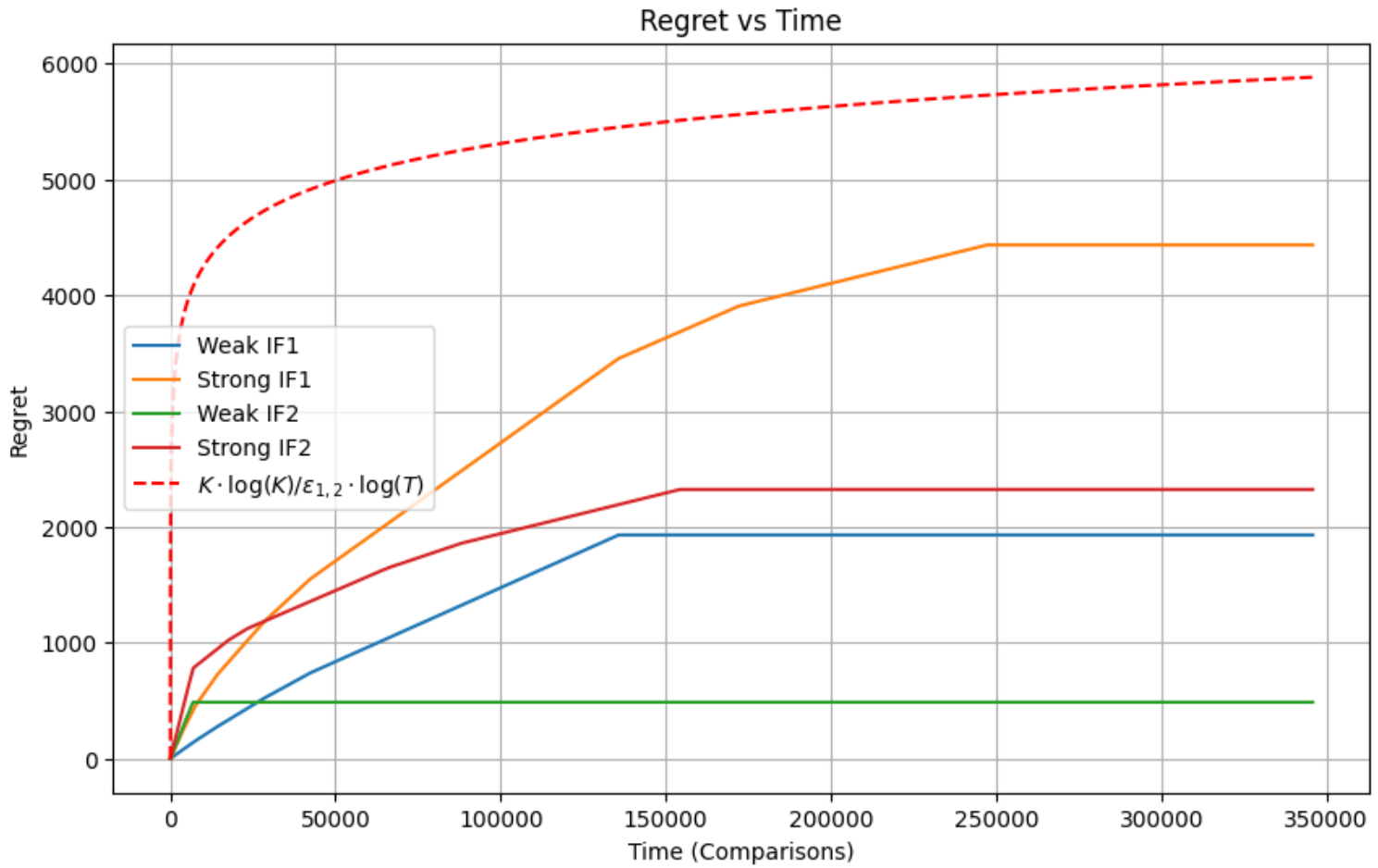
Parameters Varied -

- **Number of Arms (K):** 5, 10, 15, 20, 25
- **Number of Rounds (T):** 10000 - 500000
- **Preference Margin/Distinguishability ($\epsilon_{1,2}$):** 0.05 - 0.5

Performance Metrics Measured -

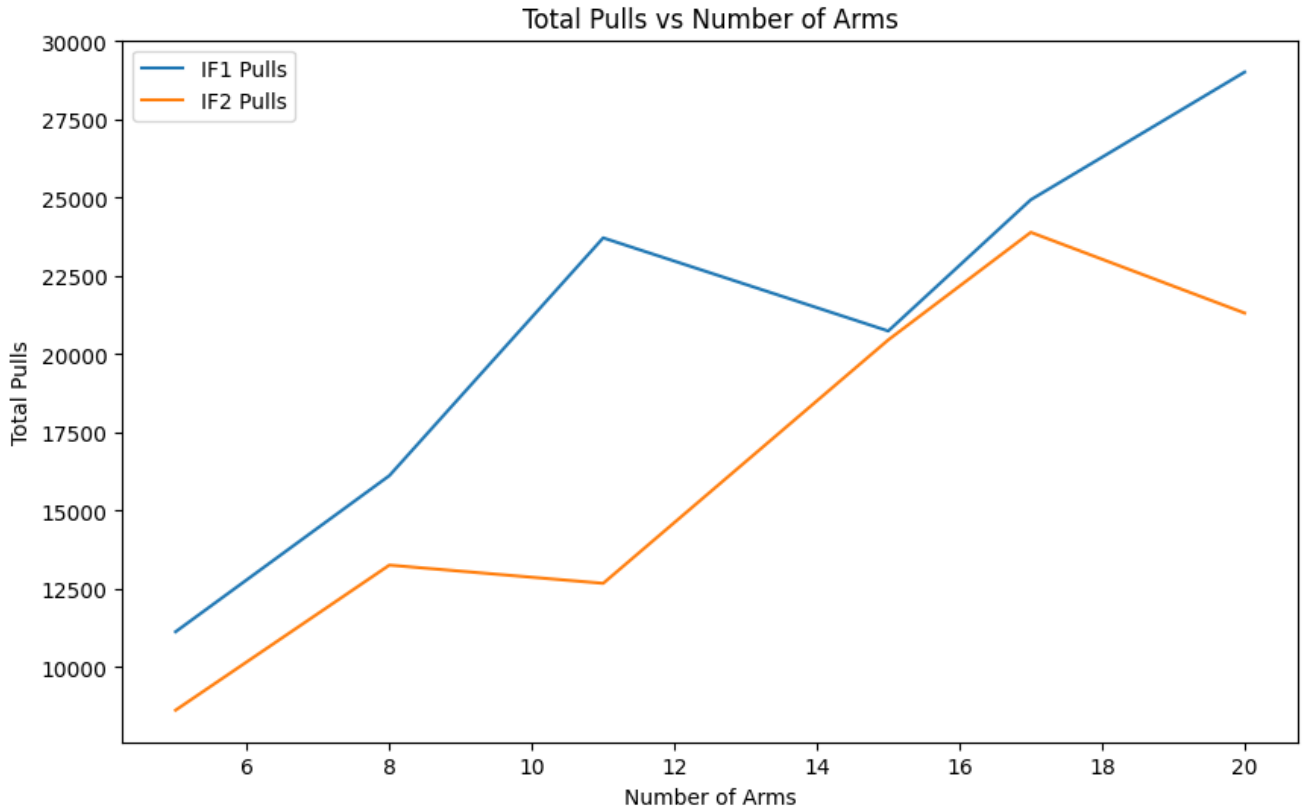
- **Total Regret:** R_T^{strong} and R_T^{weak} for both IF1 and IF2
- **Number of Comparisons:** Number of duels performed
- **Error Rate:** Number of times the best arm was not selected

Expt 1: Cumulative Regret vs Number of Rounds



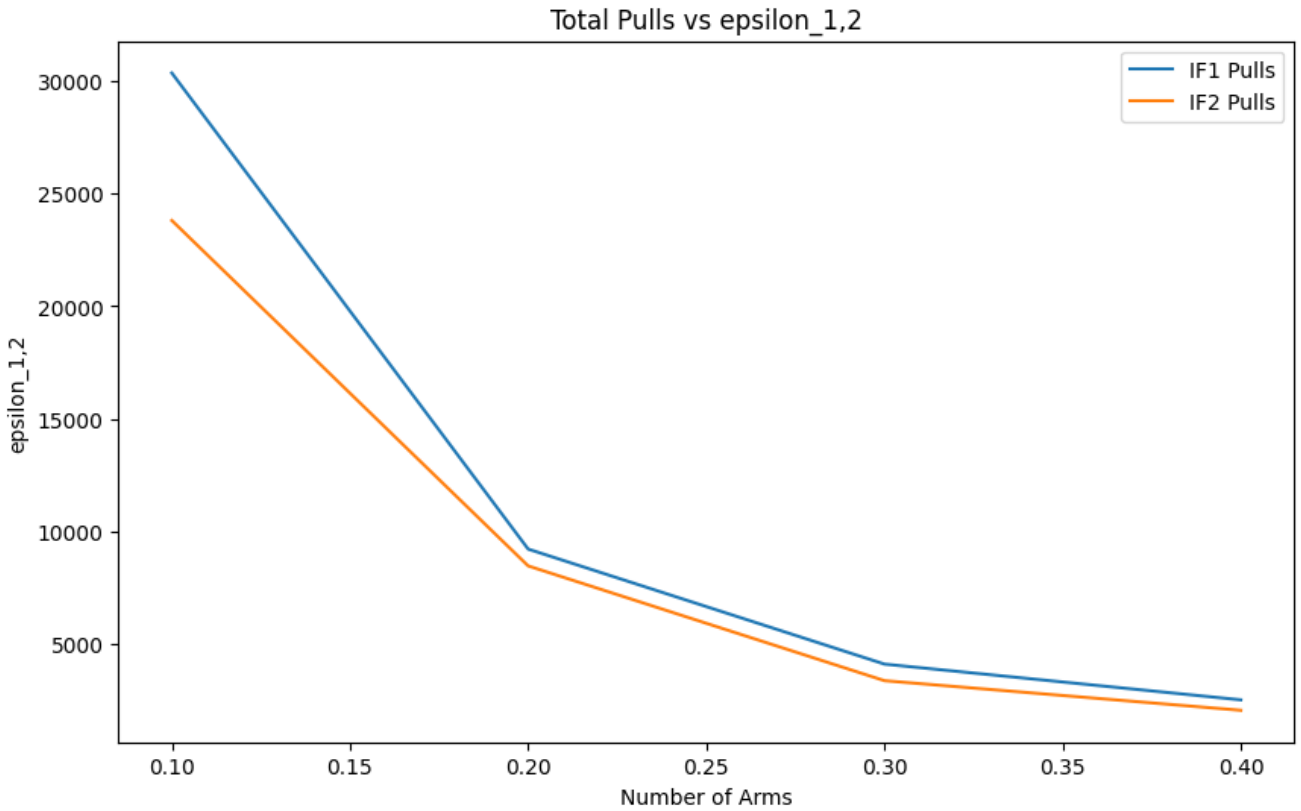
It is evident that the cumulative regret is upper bounded by $\mathcal{O}(\log T)$ for both IF1 and IF2 algorithms. The regret is also lower for IF2 as expected.

Expt 2: Number of Comparisons vs Number of Arms



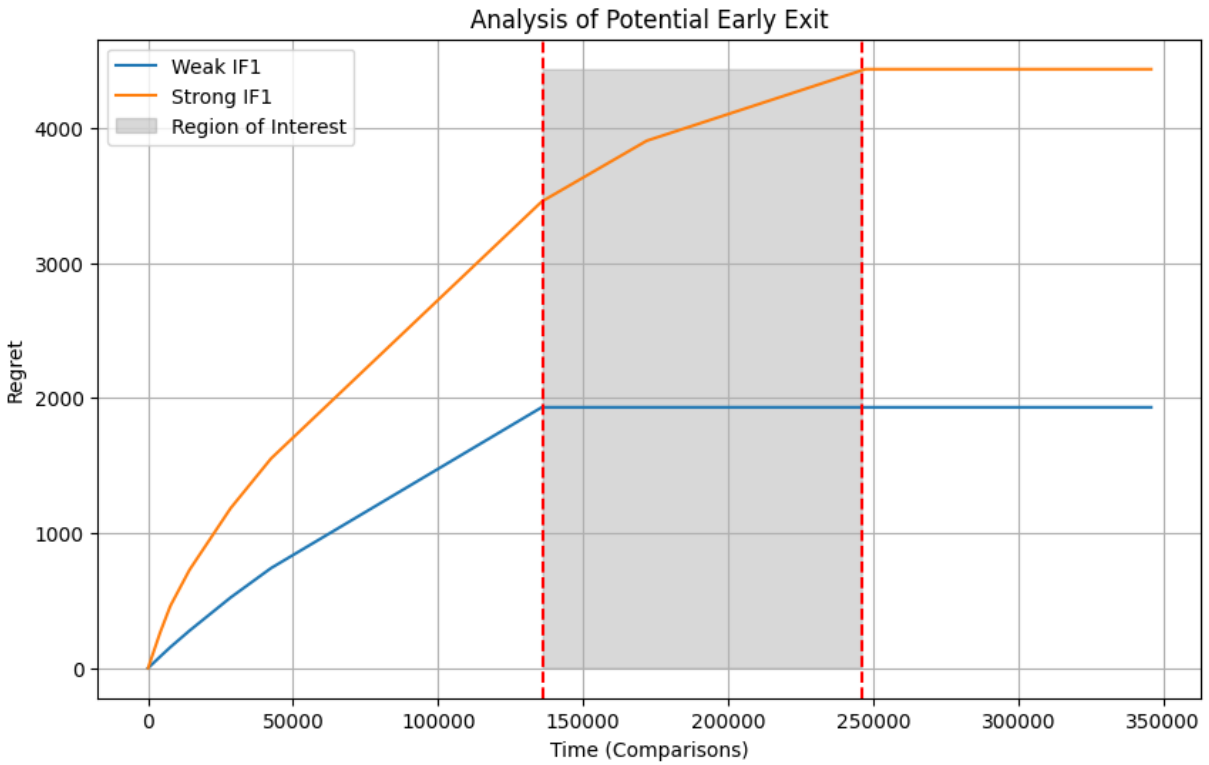
We observed that average number of comparisons made is proportional to the number of arms K. The theoretical bound for the number of comparisons is $\mathcal{O}(K \log K)$ for IF1 which is nearly $\mathcal{O}(K)$ when K is small.

Expt 3: Number of Comparisons vs $\epsilon_{1,2}$



As expected, the number of comparisons made is inversely proportional to the preference margin $\epsilon_{1,2}$. This is happening because the algorithm is able to distinguish between the arms faster when the preference margin is larger.

Expt 4: Rate of Convergence of Strong and Weak Regrets in IF1



It is evident from the figure that weak regret stabilizes much faster (at $t_1 = 140,000$) than strong regret (at $t_2 = 250,000$).

Between times t_1 and t_2, the algorithm is still exploring even when $\hat{b}$ is the final best arm. We can exploit this information to add an early stopping criterion to the algorithm and minimize strong regret.

Early Stopping Criterion

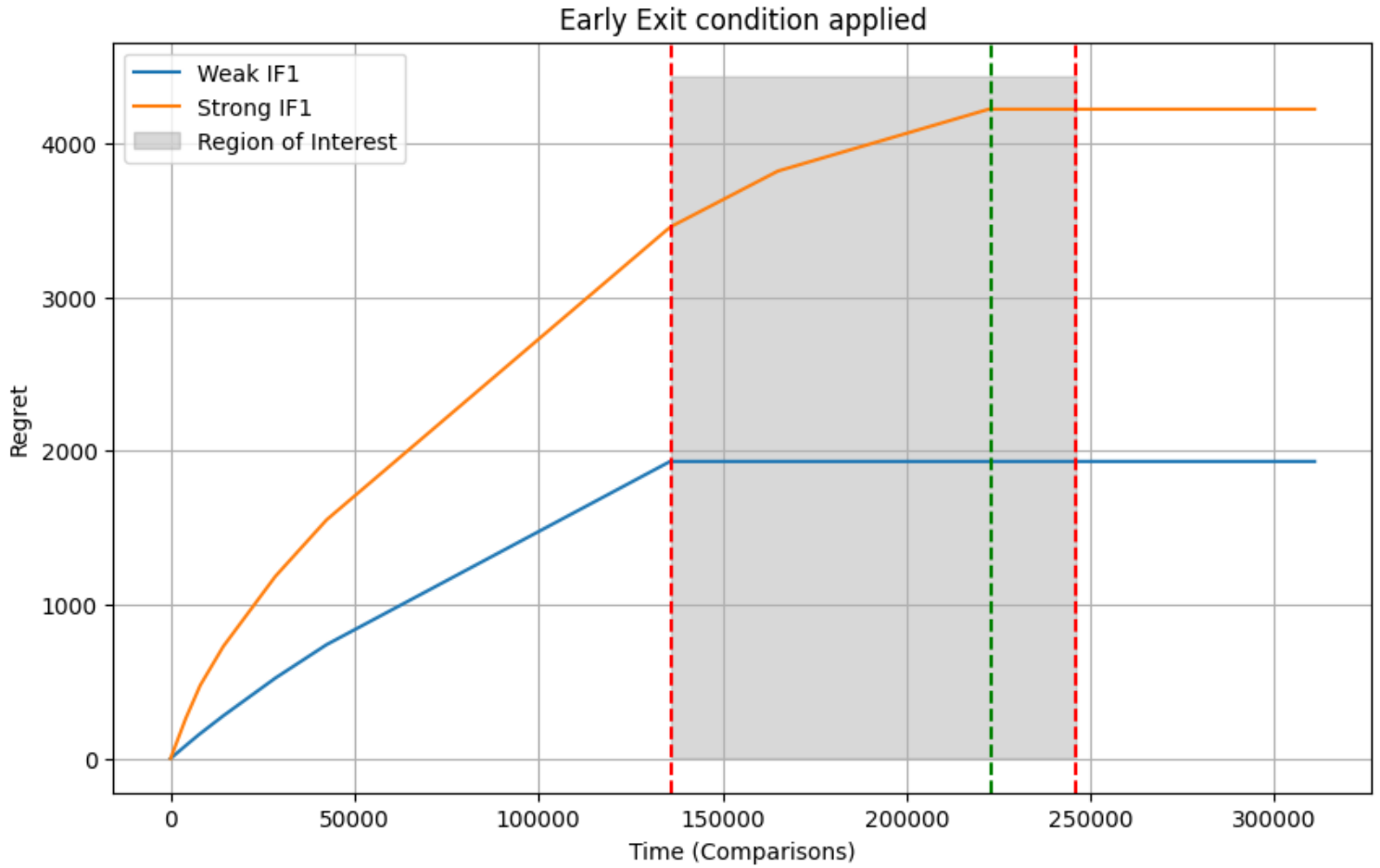
We observed that when the size of the active set of arms is small, the algorithm takes considerably more time to eliminate further arms.
Since the best arm is already found, we can immediately stop the *EXPLORE* phase and start exploiting the best arm by adding an early stopping criterion.
We decided to implement this early stopping criterion in the IF1 algorithm as follows:

- If the size of the active set of arms is less than a threshold (say $T_{stop} = 3$), consider finding the best arm by looking at the confidence intervals of the arms.
- In the removal condition, relax the confidence interval condition to $(0.3 \notin \hat{C}^{(t)}_{\hat{b},b})$ instead of $(1/2 \notin \hat{C}^{(t)}_{\hat{b},b})$.

This leads to a reduction in the regret but might also increase mistake probability in some cases. We can look for optimal hyperparameters to control this tradeoff.

Successful Early Stopping

Here is a successful example of early stopping in the IF1 algorithm. The algorithm was now able to stop at $t_2' = 222,000$ instead of $t_2 = 250,000$ with lesser regret.



In Progress: Applications to Drug Discovery

The problem of drug discovery via LSTMs is as follows:

1. Using the set of known drugs, train many LSTMs (with different hyperparameters) to generate drug candidates.
2. Independently, train a model (here a Graph Neural Network) which compares the binding affinities and relative toxicities of 2 drug molecules given their SMILES representations.
3. Define the winner of the *duel* as the drug molecule with the higher ratio of $\frac{\text{relative binding affinity}^{\alpha}}{\text{relative toxicity}^{\beta}}$.
4. Model this as a duelling bandit problem, where the arms are the LSTM models and the reward is the winner of the duel. Use any of the IF1 or IF2 algorithms to find the best LSTM model.

This is a very basic setup for this problem which we implemented given the scope of our project. We have used a simple ratio of binding affinity and toxicity which ignores many other factors.

Conclusion

We have accomplished the following in this project:

- Studied the theoretical formulation of the duelling bandit problem and its applications in real world.
- Implemented the IF1 and IF2 algorithms for the duelling bandit problem.
- Verified the theoretical regret bounds for both algorithms by running extensive simulations.
- Identified opportunities to add early stopping criteria to the algorithms to improve performance.
- Applied the algorithms to a real world problem of drug discovery (work in progress).

Thank You!