A Bandit in a Bandit

Adaptive Windowing for Non-Stationary Contextual Multi-Armed Bandits

Jack Kerr [1] Theo Crookes [2]

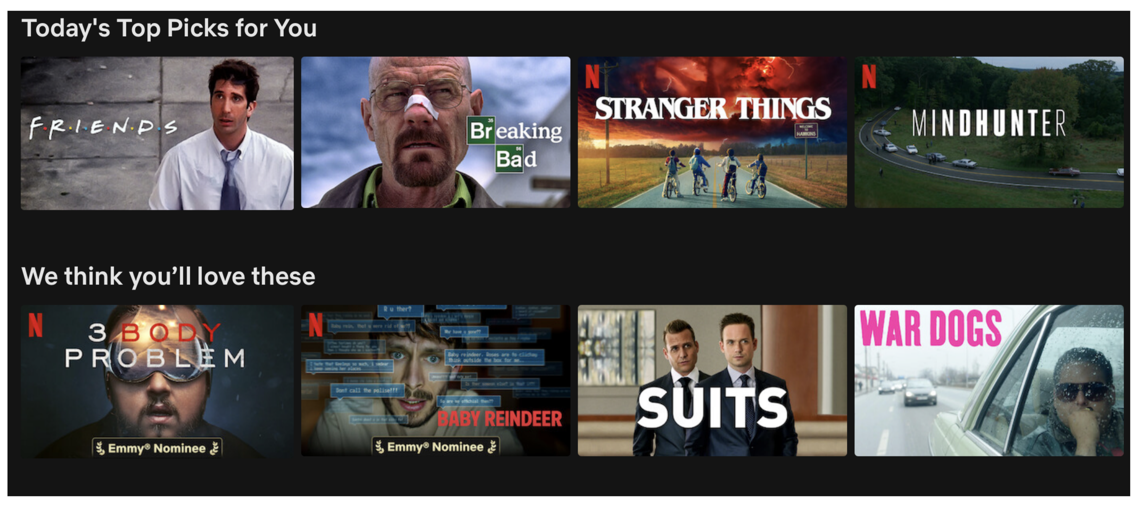
[1]University of Warwick [2]Lancaster University

excellence with impact

1. Motivation

Contextual Multi-Armed Bandit (CMAB) algorithms are an extension of Multi-Armed Bandit algorithms that **incorporate context** into each arm. This allows them to make more informed decisions about which arm to select. They are commonly used in **recommender systems**, like that used by Netflix, and **online advertising**.

This project aimed to implement these CMAB algorithms in a setting where the **mean reward** for the system is a **function of time**. This leads us to build upon CMAB algorithms to apply them to this setting.



2. Contextual Bandits

In a Contextual Multi-Armed Bandit problem, a learner must **pick an arm** from a set of arms when given context about each one. We want to pick the arm with the highest reward but balance this exploitation with the exploration of new arms.

- The **context** is denoted $\mathbf{b}_i(t) \in \mathbb{R}^d$
- $r_i(t)$ is the **reward** of arm i at time t
- $\boldsymbol{\mu} \in \mathbb{R}^d$ is known as the true but unknown parameter such that $\mathbb{E}[r_i(t)|\mathbf{b}_i(t)] = \mathbf{b}_i^T(t)\boldsymbol{\mu}$
- a_t^* is the **optimal arm** and a_t is the **chosen arm** at time t

We wish to **minimise the cumulative regret** over the time horizon T. This is defined as:

$$\mathcal{R}(T) = \sum_{t=1}^{T} \mathbf{b}_{a_t^*}^T(t)\,\boldsymbol{\mu} - \mathbf{b}_{a_t}^T(t)\,\boldsymbol{\mu}$$

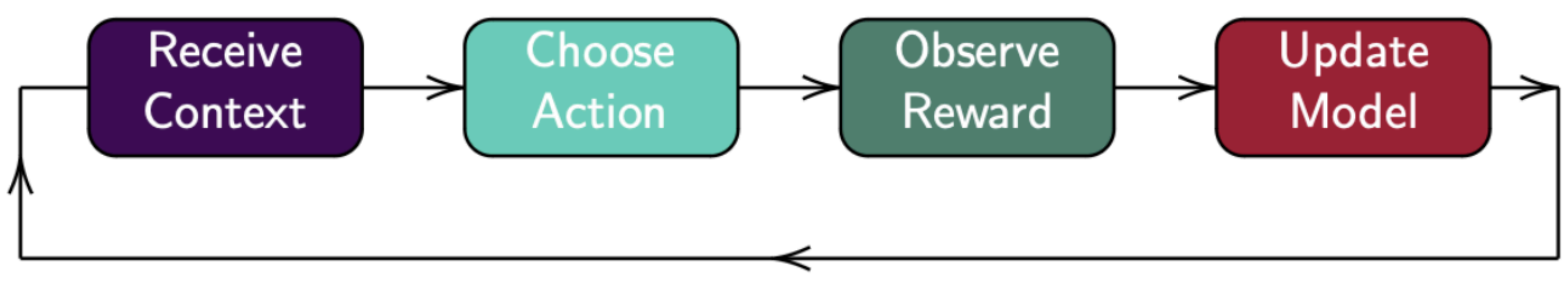


Figure 1: Contextual Multi-Armed Bandit flowchart

3. The UCB Algorithm

The algorithm that we will use as a base is **LinUCB**. It works by computing **Upper Confidence Bound** (UCB) values for each arm and playing the arm with the highest value. They take the general form of:

$$UCB_i(t) = \text{Estimated Reward} + \text{Uncertainty}$$

The UCB values use the **parameter** $\boldsymbol{\alpha}$ to balance exploration and exploitation to weight the uncertainty.

$$UCB_i(t) = \mathbf{b}_a^T(t)\hat{\boldsymbol{\mu}}(t) + \alpha\sqrt{\mathbf{b}_a^T(t)B(t)^{-1}\mathbf{b}_a^T(t)}$$

Where,

$$B(t) = \lambda \mathbf{I}_d + \sum_{k=1}^{t-1} \mathbf{b}_{a_k}(k)\mathbf{b}_{a_k}^T(k)$$

$$\hat{\boldsymbol{\mu}}(t) = B(t)^{-1} \sum_{k=1}^{t-1} \mathbf{b}_{a_k}(k) r_k$$

We use **sliding window algorithms** to deal with non-stationary settings. They involve using the history of at most τ steps before the current one to estimate $\boldsymbol{\mu}$.

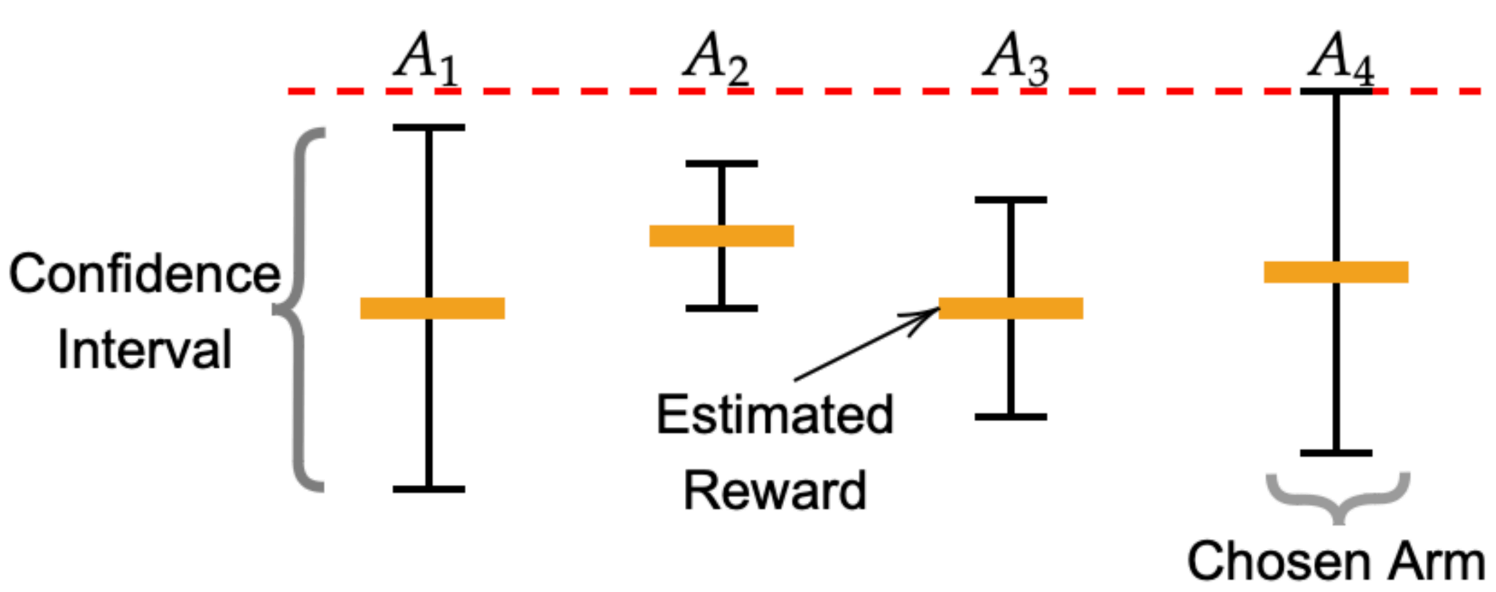


Figure 2: UCB selection diagram

4. Hedging Algorithm

Hedging relies on each window size having some **probability of being selected** and then updating this probability based on its relative performance.

Set $W_1 = \mathbf{1} \in \mathbb{R}^N, \mathbf{x}_1 = \frac{1}{N}W_1$
for all $t = 1, \dots, T$ **do**
    Sample $\tau \sim \{10, 20, 30, \dots, M\}, \quad \mathbb{P}(\tau = i) = \mathbf{x}_1$
    Observe reward r_t from playing according to SW-UCB with window size τ
    Compute the loss given by some function $\mathbf{g}(r_t, r_{t-1})$
    Update Weights $W_t(\tau) = W_{t-1}(\tau)e^{\mathbf{g}(r_t, r_{t-1})}$
    Set $\mathbf{x}_t = \frac{W_t}{\sum_j W_t(j)}$
end for

5. ϵ-Greedy (ish)

We use the same set-up as in hedging. We now track how many times τ has been used and **play the arm with the largest** β_τ, as given below, apart from an ϵ chance of playing a **random arm**.

$$\beta_\tau = \beta_\tau \left(1 - \frac{1}{P_{\tau,t}}\right) + \frac{\left((r_t - r_{t-1}) - \frac{1}{|S_{\tau,t}|}\sum_{s \in S_{\tau,t}}(r_s - r_{s-1})\right)}{P_{\tau,t}}$$

Where, $S_{\tau,t} = \{i \in \{1, \dots, t\} \mid w_i \neq \tau\}$ is the set of times where a window size of τ wasn't used.

6. Mass Updates

To speed up the learning process we update each window size that would have also selected the arm the chosen window size did. These extra windows **receive half the reward** they would have ordinarily received if they were chosen.

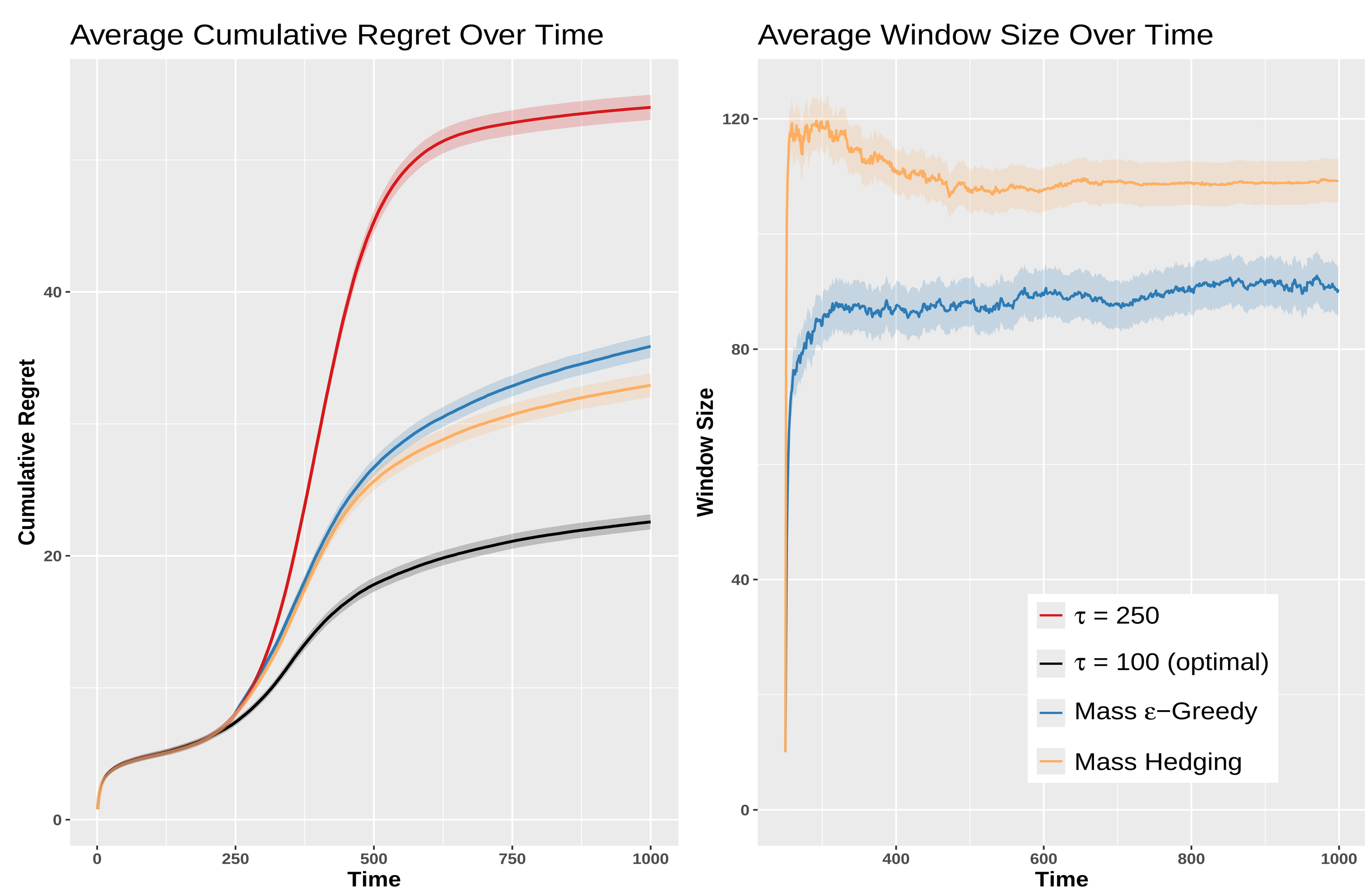


Figure 3: Algorithm comparison with a linear $\boldsymbol{\mu}$ with 95% confidence intervals over 500 iterations

7. References

- Gi-Soo Kim, Young Suh Hong, Tae Hoon Lee, Myunghee Cho Paik, and Hongsoo Kim. Bandit-supported care planning for older people with complex health and care needs, 2023.
- Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation, April 2010.