Investigating Optimism in the Exploration/Exploitation Dilemma

Matthew Gorton[1] Supervisor: Alan Wise[2]

[1]Physics, Lancaster University [2]Maths & Stats, Lancaster University

excellence with impact

The Problem

Set-up: K 'arms', with different rewards. The reward from each arm follow an unknown probability distribution. Initially, we don't know which arm is best.

Our aim: Find an algorithm minimising the 'cumulative regret' after n runs,

$$R_n \equiv n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right] = \sum_{k\in[K]} \Delta_k \mathbb{E}[T_k(n)],$$

- μ^* - mean of the optimum arm
- X_t - reward at time t
- $[K] = 1, 2, \dots K$ - set of arm indices
- Δ_k - difference in mean of arm k and mean of the optimum arm
- $T_k(n)$ - number of times arm k has been played after n runs

This sort of problem arises in *e. g.* targeted advertising, clinical trials

Aim to find an algorithm achieving 'sub-linear regret', i.e.

$$\lim_{n\to\infty} \frac{R_n}{n} = 0.$$

The Exploration/Exploitation Dilemma

To minimise regret, an algorithm must balance two competing approaches:

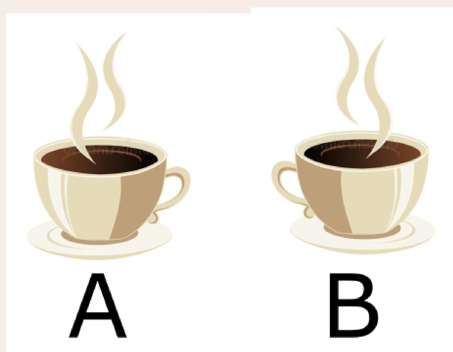
- **Exploration**: To better estimate the expected reward of each arm, we need to test different arms.
- **Exploitation**: We should select arms that give us a higher reward more often.

The Optimism Principle

Our approach is to assume that the environment is as nice as 'plausibly possible' [Lattimore and Szepesvári, 2018].

E. g. Choosing a cafe

- Choose a cafe you know well (A), or a new cafe (B) you haven't tried before?
- Optimism principle → try the new cafe several times, and update your information



Upper Confidence Bound (UCB) algorithms

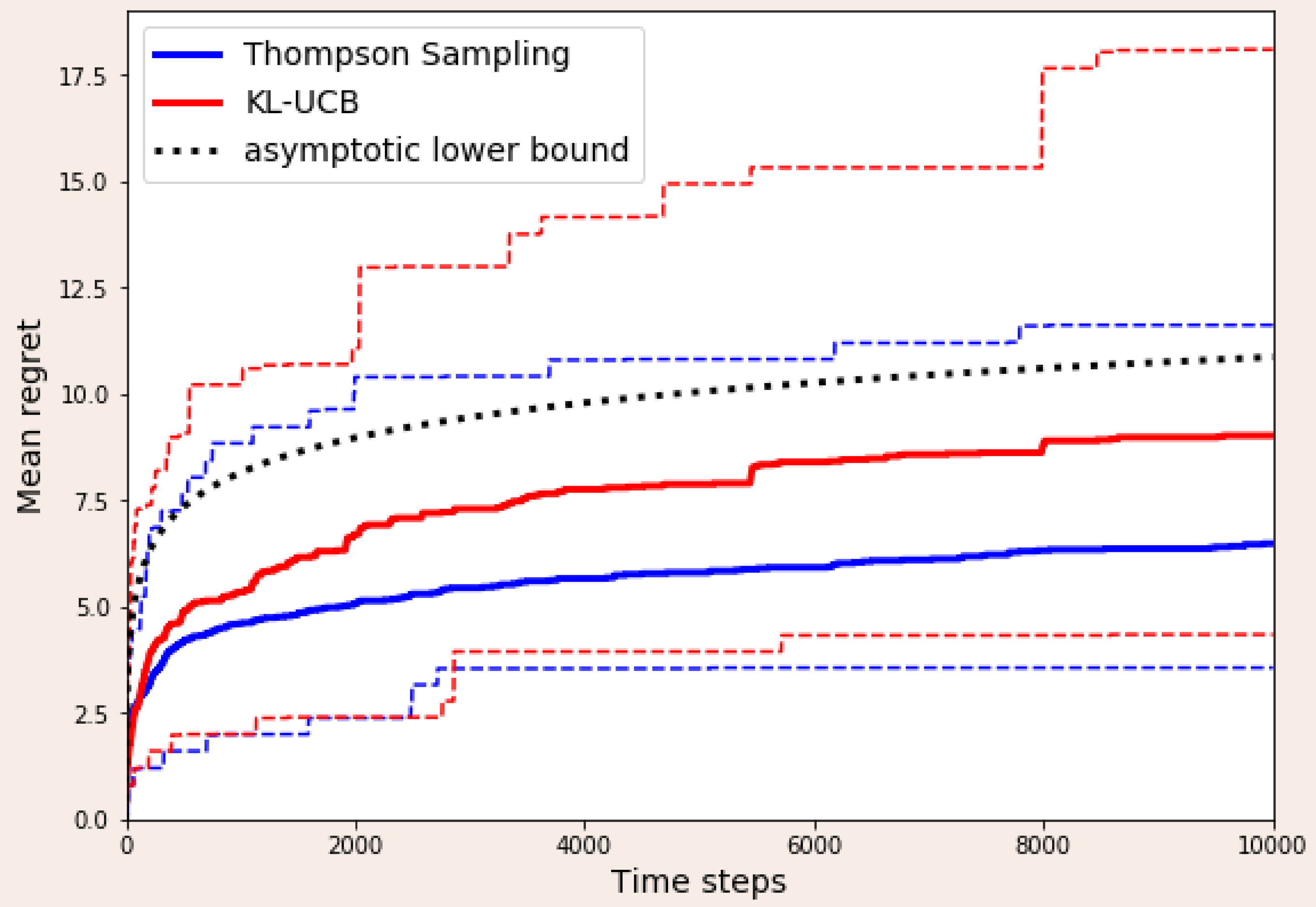
Select arm at time t, A_t, maximising the 'upper confidence bound', *i.e.*

$$A_t \in \arg\max_k \left[\hat{\mu}_k(t-1) + f(T_k(t-1))\right],$$

where $f(T_k(t-1))$ is a decreasing function of $T_k(t-1)$

- $\hat{\mu}_k(t-1)$: encourages exploitation
- $f(T_k(t-1))$ encourages exploration

Can achieve theoretical asymptotic lower regret bound [Lai and Robbins, 1985], using the KL-UCB method.



KL-UCB and Thompson Sampling achieve optimal regret for a 2-armed Bernoulli bandit. The theoretical asymptotic lower regret bound is shown in black. Dashed lines represent 95% confidence intervals, calculated from 20 test runs.

Bayesian Approach: Thompson Sampling

Approach:

1. Assume a prior probability distribution for each arm (*e. g.* Beta(α_k, β_k))
2. Sample a value θ_k from the prior for each arm, $\theta_k \sim \text{Beta}(\alpha_k, \beta_k)$
3. Select arm with maximum θ_k
4. Update α_k, β_k

Advantages

- Optimal performance (for Bernoulli bandits)
- Not sensitive to parameters in prior (tuning not required)

Disadvantages

- Only works well if prior pdf is conjugate to the true pdf

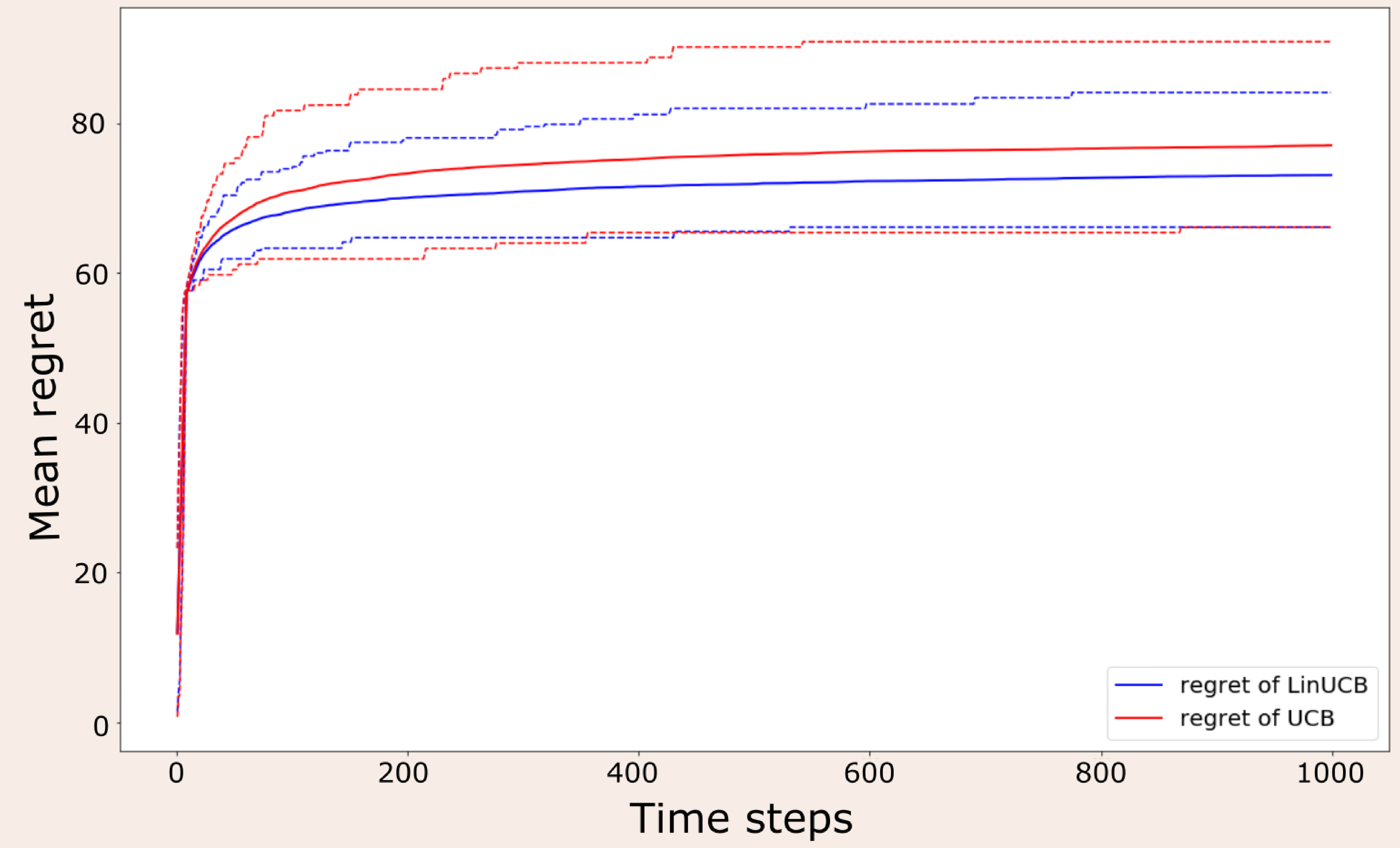
Contextual Bandits

Sometimes, we can use information about one arm to make predictions about other arms, improving algorithmic performance. *e. g.* someone who buys red jumpers is likely to buy blue jumpers as well.

- Arms are now normalised vectors.
- The extent to which arms 'point' in the same direction shows their similarity.

LinUCB algorithm

Uses information collected from observations to update a 'feature vector', guiding arm selection



LinUCB generally performs better than a simple UCB algorithm. Plot by Alan Wise.

Future Work

- Investigate whether different methods perform better for particular bandit probability distributions
- Fix LinUCB and look into extensions
- Consider more complicated examples, *e. g.* bandits in fraud detection

References

Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in applied mathematics, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2018).
Bandit algorithms.
preprint.

https://www.lancaster.ac.uk/stor-i/internships/interns m.gorton1@lancaster.ac.uk