Investigating Optimism in the Exploration-Exploitation Dilemma

Matthew Gorton
Supervisor: Alan Wise

September 5, 2019

The Multi-Armed Bandit Problem

Set-up

K ‘arms’, with different rewards:

- Rewards follow a probability distribution $\rightarrow$ **explore** different arms
- Want to maximise reward $\rightarrow$ **exploit** arms which give a higher reward

The Multi-Armed Bandit Problem

Set-up

K 'arms', with different rewards:

- Rewards follow a probability distribution $\rightarrow$ **explore** different arms
- Want to maximise reward $\rightarrow$ **exploit** arms which give a higher reward

Example applications

- Recommender systems (e.g. targeted advertising)
- Adaptive clinical trials

Mathematically

Our aim is to minimise the 'regret' R_n after n runs,

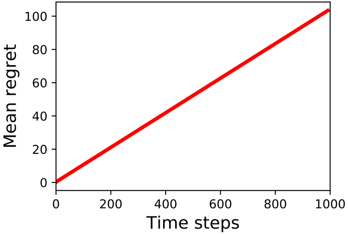
$$R_n \equiv n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right].$$

- t - number of time steps
- μ^* - mean of the optimum arm
- X_t - reward at time t

What is a good algorithm?

Consider the definition of regret,

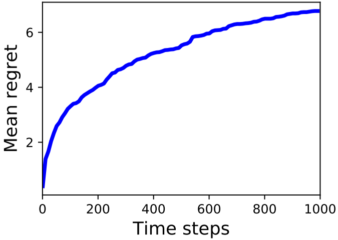
$$R_n \equiv n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right].$$



Sub-linear regret

A good algorithm achieves sub-linear regret,

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



Baseline strategy: ε_n-Greedy

Probability	**Arm selected A_t**
$1 - \varepsilon_n$	$\arg\max_k \left[\hat{\mu}_k(t-1)\right]$
ε_n	random

- ε_n decreases with time $\rightarrow$ less exploration at late times.
- Two tuning parameters, c and d $(c > 0, 0 < d < 1)$

Optimism in the face of uncertainty

Optimism Principle: act as if the environment is as nice as plausibly possible [Lattimore and Szepesvári, 2018].

A chain

Not a chain

UCB algorithms

Select the arm maximising the 'upper confidence bound', which is usually of the form

$$\underbrace{\hat{\mu}_k(t-1)}_{\textit{exploitation}} + \underbrace{f(T_k(t-1))}_{\textit{exploration}}.$$

- $\hat{\mu}_k(t-1)$ - observed mean of arm k
- $f(T_k(t-1))$ - decreasing function of T_k

UCB algorithms

Select the arm maximising the 'upper confidence bound', which is usually of the form

$$\underbrace{\hat{\mu}_k(t-1)}_{\textit{exploitation}} + \underbrace{f(T_k(t-1))}_{\textit{exploration}}.$$

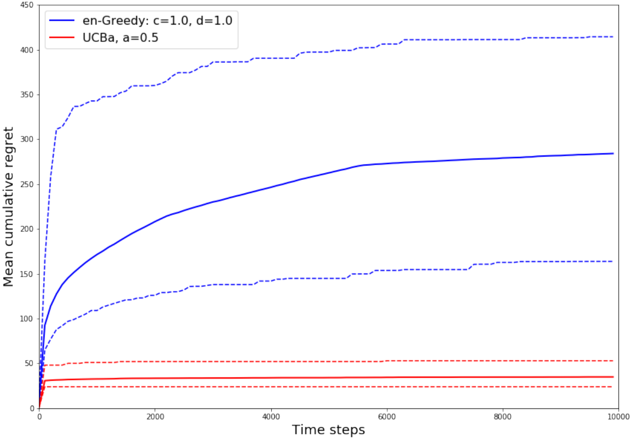
- $\hat{\mu}_k(t-1)$ - observed mean of arm k
- $f(T_k(t-1))$ - decreasing function of T_k

Example forms of the upper confidence bound:

- UCB(α): $\hat{\mu}_k(t-1) + \sqrt{\frac{\alpha \ln(t-1)}{T_k(t-1)}}$
- KL-UCB: $\max\left\{q : T_k(t-1)\mathrm{kl}\left(\hat{\mu}_k(t-1), q\right) \leq \ln(t-1) + c\ln(\ln(t-1))\right\}$

ε_n-Greedy vs UCB

Best values: UCB(α): $\alpha \approx 0.5, \varepsilon_n$-Greedy: $c \approx 1, d \approx 1$

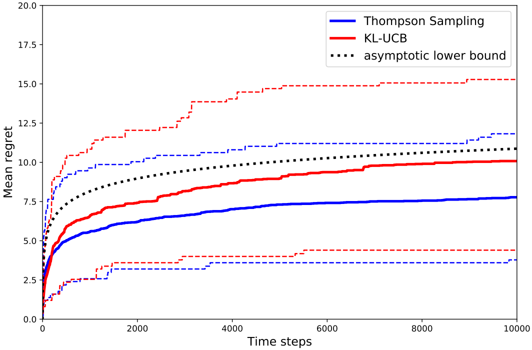


Performance on a 9-armed Bernoulli bandit: dashed lines represent the 95% confidence interval.

Optimal approaches

Lower regret bound

- Can calculate the asymptotic lower bound on the regret [Lai and Robbins, 1985]
- UCB(α) does not match lower regret bound. Other algorithms (KL-UCB and Thompson Sampling) can match the lower bound in the case of a Bernoulli bandit

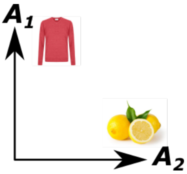


Performance on a 2-armed Bernoulli bandit: dashed lines represent the 95% confidence interval

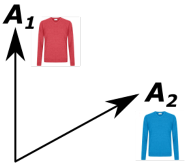
Contextual bandits

Key points

- Arms $\boldsymbol{A_t}$ are now normalised vectors
- The extent to which arms 'point' in the same direction shows their similarity



Arms orthogonal: no information about $\boldsymbol{A_2}$ from $\boldsymbol{A_1}$.



Arms give information about one another.

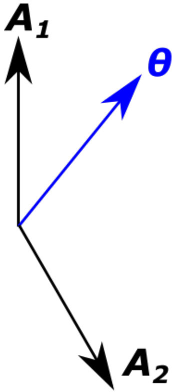
Contextual Bandits: Regret

Unknown context vector $\boldsymbol{\theta}$: represents optimal arm 'direction'

Regret is now defined as

$$R_n = \sum_{t=1}^{n} \left[\langle \boldsymbol{A}^*, \boldsymbol{\theta} \rangle - \langle \boldsymbol{A_t}, \boldsymbol{\theta} \rangle \right],$$

where $\boldsymbol{A}^*$ is the optimum arm.



E.g. Music streaming

Imagine you are a music streaming service, producing playlists.

E.g. Music streaming

Imagine you are a music streaming service, producing playlists.

- Arms (playlists) $\boldsymbol{A_k}$ are made of songs from d artists, e. g. for $d = 5$

$$\boldsymbol{A_k} = \left[0, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0, \frac{1}{\sqrt{3}}\right]$$

E.g. Music streaming

Imagine you are a music streaming service, producing playlists.

- Arms (playlists) $\boldsymbol{A_k}$ are made of songs from d artists, e. g. for $d = 5$

$$\boldsymbol{A_k} = \left[0, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0, \frac{1}{\sqrt{3}}\right]$$

- On average, the user listens to each artist's songs $i = 1, ..., d$ for a time θ_i, so $\boldsymbol{\theta}$ may look like

$$\boldsymbol{\theta} = [53, 23, 9, 39, 16]$$

E.g. Music streaming

Imagine you are a music streaming service, producing playlists.

- Arms (playlists) $\boldsymbol{A_k}$ are made of songs from d artists, e. g. for $d = 5$

$$\boldsymbol{A_k} = \left[0, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0, \frac{1}{\sqrt{3}}\right]$$

- On average, the user listens to each artist's songs $i = 1, ..., d$ for a time θ_i, so $\boldsymbol{\theta}$ may look like

$$\boldsymbol{\theta} = [53, 23, 9, 39, 16]$$

- Mean time user listens to playlist: $\langle \boldsymbol{A_k}, \boldsymbol{\theta} \rangle = 27.7$

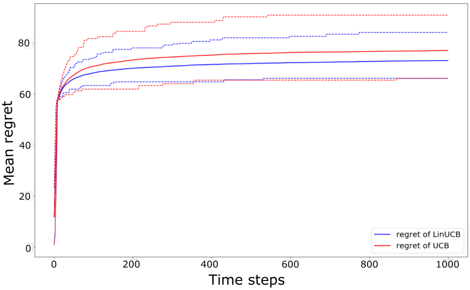
LinUCB

Same idea as previous UCB algorithms. The upper confidence bound is now

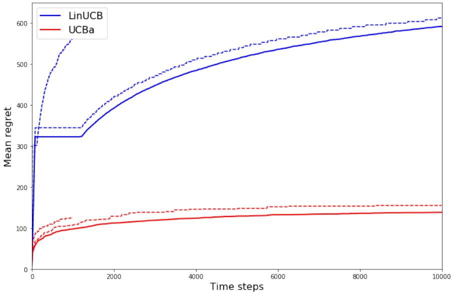
$$\langle \hat{\boldsymbol{\theta}}_{\boldsymbol{t},\boldsymbol{\lambda}}, \boldsymbol{A_k} \rangle + B_t(\delta)||\boldsymbol{A_k}||_{G_{t,\lambda}^{-1}},$$

- $\hat{\boldsymbol{\theta}}_{\boldsymbol{t},\boldsymbol{\lambda}}$ - estimated context vector
- $B_t(\delta)$ - encourages exploration
- $||\boldsymbol{A_k}||_{G_{t,\lambda}^{-1}}$ - encourages exploitation (like a standard deviation)

LinUCB: Performance



Expected regret curve. Produced by Alan Wise.)



Current output

Future Work

- Does algorithmic performance depend on arm pdfs?
  - Only explored sub-Gaussian distributions
- Extensions to LinUCB
  - Simple LinUCB requires great fine tuning

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.

Extra: Equivalent expression for regret

Using the regret decomposition lemma, the regret can also be expressed as

$$R_n = \sum_{k \in [K]} \Delta_k \mathbb{E}[T_k(n)],$$

- $[K] =$1,2,...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

Extra: Expression for asymptotic lower regret bound

[Lai and Robbins, 1985] find the asymptotic lower regret bound is given by

$$\left(\sum_{k, \Delta_k > 0} \frac{\Delta_k}{\mathrm{kl}(\mu_k, \mu_\star)} \right) \ln(T),$$

- $\mathrm{kl}(\mu_k, \mu_\star)$ - Kullback-Leibler divergence between the pdfs of the optimum arm and arm k
- T - total running time