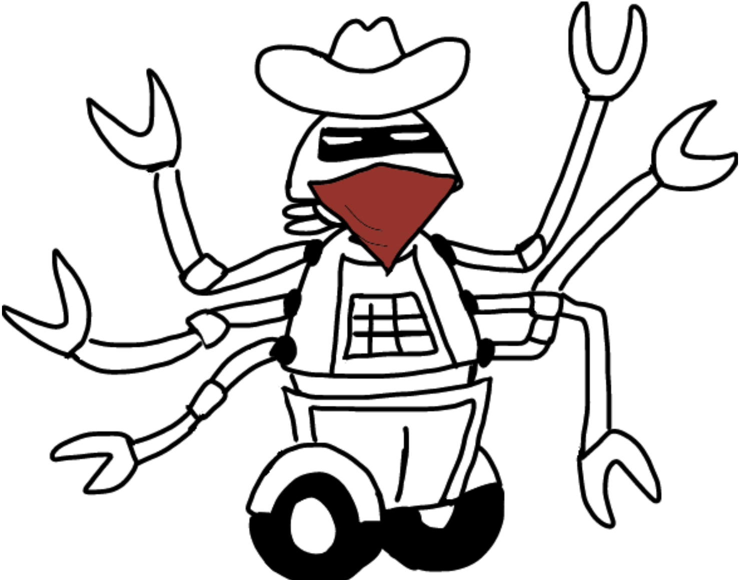
Bachelor Project: Recommending Products with Multi-Armed Bandits

Bob Rombach

Rotterdam School of Management

Febuary, 2025

Welcome to the Bandit project!



Our agenda

1. Introductions
2. Bandits for advanced dummies
3. Bandit Research Project
4. Literature Review
5. Discussion

Who am I and who are you?

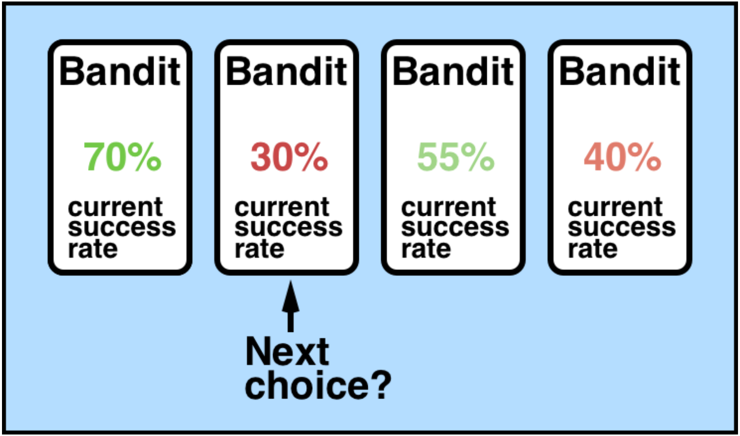
Bob Rombach (rombach@rsm.nl)

- As a PhD student, I spend the last 4 years of my life trying to understand and develop bandits
- Before that I did a BSc and an MSc in Econometrics and an MPhil in business data science

Who are you?

- Why did you pick this project?
- What do you expect (from me and the project)

Bandit choices



The Multi-Armed Bandit Problem

The end-goal: Play the actions that maximize the discounted cumulative rewards:

1. Action or Arm: An advertisement, a product, a medication
2. playing an action: Showing the add, recommending the product, administering the medication
3. Rewards: Clicks on ads, profits in the store, patients surviving
4. Cumulative: We care about ALL the rewards
5. Discounted: We put a lower weight on future rewards

The Math of the Previous Slide

Play the actions that maximize the discounted cumulative rewards:

$$\operatorname*{argmax}_{a \in A} \sum_{t=1}^{\infty} \delta^t E[R(a_t)]$$

1. $A = \{a_1, \ldots, a_J\}$ is the set of available actions
2. $a_t \in A$ is the action played at time t
3. We discount future rewards with factor δ^t, for $0 < \delta \leq 1$
   - What is the effect of the discount factor?
4. $E[R(a_t)]$ is the expected reward associated with action a_t
   - How does this relate to the actual discounted cumulative rewards?
5. Argmax means that we pick the action that maximizes the sum

Learning Versus Earning

So... Which Action do we play?

1. The action with the largest myopic reward? (myopic = based on current estimate)
2. The action with the most uncertainty (of the estimate)?
3. Another action?

This is the core of the MAB problem: Do we learn or do we earn?

- Learning/Exploring: Play actions to decrease the uncertainty in our estimates
- Earning/Exploiting: Play the actions that have the largest myopic reward

Solving a MAB Problem

A solution to the MAB problem contains:

- Estimates of the expected rewards of all arms
- A "strategy" to pick arms based on the estimates of expected rewards.
- A method to update our estimates

Let's look at the Bernoulli bandit example:

- We assume that the expected rewards of arm j follow a Bernoulli distribution with unknown parameter p_j
- We model the reward using the *Conjugate Prior* of the Bernoulli distribution: Beta(α, β), with $p_j = \frac{\alpha_j}{\alpha_j + \beta_j}$
- Upon observing reward $R(a_j) \in \{0, 1\}$ update the parameters of arm j

$$\alpha_j = \alpha_j + R(a_j)$$
$$\beta_j = \beta_j + (1 - R(a_j))$$

Bandit Strategies: Which arm do we pick?

- **Greedy algorithms**: Pick the arm with the highest estimated reward

$$a_t = \underset{j}{\operatorname{argmax}} \frac{\alpha_j}{\alpha_j + \beta_j}$$

    - Problem: We place no weight on learning!
- **Purely random algorithms**: Pick an arm at random
    - Problem: We place no weight on earning!
- **Epsilon Greedy algorithms**: Play a random arm with ε probability, otherwise play the arm with the highest expected reward
    - Higher values of ε increase learning (we explore actions)
    - Lower values of ε increase earning
    - Problem: What is the optimal value of ε?

Advanced Bandit Strategies

- **Thompson Sampling**: The frequency an arm is played is proportional to it being optimal:
  1. Draw a random value $\hat{\theta}_j$ from Beta$(\alpha_j,)$ for all arms j
  2. Play the arm with the highest θ_j $(\text{argmax}_j\, \hat{\theta}_j$
- **Upper Confidence Bound**: Place a weight on both myopic value and uncertainty (example UCB1) and pick the arm with the highest value

$$a_t = \underset{j}{\text{argmax}}\, Q_t(a_j) + c\sqrt{\frac{log(t)}{N_t(a_j)}}$$

  - $Q_t(a_j) = \frac{\alpha_j}{\alpha_j+\beta_j}$ is the myopic arm value
  - c is the confidence value that controls exploration
  - t is the total amount of actions
  - $N_t(a_j)$ is the number of times arm j is played at time t
- **Gittins Index**: The **only** optimal bandit algorithm
  - Optimally balances learning/earning, but many assumptions

Contextual Bandits

Often, both products and users/website visitors have known characteristics:

- For example, sports articles versus political articles
- Male, female, non-binary or unknown genders

We can leverage this information to increase the value of our method:

- The average reward of all political articles may be related
- Males may have preferences different from other genders

There are many contextual bandit methods that may result in superior performance to non-contextual bandits!

Measures of Bandit Performance

There are two "main" performance measures in the world of bandits for the realized rewards $R(a_t)$:

1. The average rewards of the bandit policy: (higher is better)

$$\text{average reward} = \frac{1}{T}\sum_{t=1}^{T} R(a_t)$$

2. The average regret of the bandit policy: (lower is better)

$$\text{Regret} = \frac{1}{T}\sum_{t=1}^{T} E[R(a^*)] - E[R(a_t)]$$

- $a^* = \operatorname{argmax}_j E[R(a_j)]$ is the expected reward ~~of the optimal arm~~

The Plan

- I will share a data set of 1 million Yahoo! Frontpage recommended news articles consisting of:
  1. Users and their characteristics
  2. The articles that have been recommended
  3. The articles that could have been recommended
  4. The timestamps of the recommendations
  5. The outcome (i.e. whether the user clicked on the article)

What will you do with it:

1. Build a simulation environment
2. Implement several bandits and run them in your simulation environment
3. Compare and assess the performance of your bandits.

Simulation Environment

Bandits choose what actions to play, but the data is already gathered....

- We need to simulate reality based on the underlying data!

1. Bandit first:
    - First, Let the bandit pick an action
    - Second, determine the outcome based on the data, for example:
        - Sample an observation from the data and use the outcome
        - Calculate the expected rewards and sample from bernouli distributions with those rewards
2. Data first: (e.g. replay)
    - In chronological order (i.e. based on time stamps), sample a data point
        - If the action the bandit picks is the same as the data point use the observation
        - Else, skip the observation and go to t+1

How to Design/Select the environment

The best simulation method depends on your data and your bandit method!

- This depends on the data and your method
- The goal is to reflect the real world as close as possible

The problem with data first:

- If the actions of the bandit and the data often do not line-up, we use only a small fraction of the data

The problem with bandit first:

- It is not easy to make a simulation environment that closely resembles the real world.

Choose wisely, and motivate your choices clearly:

- Limitations and assumptions!

Implementing Bandits: Basis of the Research

Within our simulation environment we will implement several bandits:

- Comparing several bandits based on your research question!

For example, we could as questions such as:

1. Does using the user context using method X improve the performance over method X without user context? (how do we incorporate context? How do we define performance?)
2. Can we make UCB perform better than Thompson Sampling by optimizing the exploration parameter c? (based on what measure)

Give enough thought to your research question, and motivate it:

- What are the underlying assumptions (e.g. the data you use)
- What sub-questions must we answer?
- Is it challenging enough? Is it too challenging?
- Is it relevant for managers? Policy makers?

Assessing Bandit Performance: Answering the Research Question

Next, we need to assess the performance of our methods:

- Usually, we run each bandit several times in our research environment with several *seeds*

The seed is the randomness in the simulation, thus, we see how much effect randomness can have on performance:

- Both the simulation environment and the bandits often contain some randomness

Report in tables and in graphs: (see examples in the literature)

- Average performance (potentially over time)
- Standard deviation of performance

No hypothesis testing please!

My advice: Start Now!

The designing of your simulation environment and bandit methods together with programming will take a lot of time and effort!

- You need to think about many choices, how they affect your research and how you can implement them.
- A simulation environment good for method A may be bad for method B. If you want to compare A to B you must design a simulation environment that works for both
- Programming and assessing the validity of your simulation environment and methods will not be easy.
- Little packages are available, if you use them make sure you understand how they work and what the assumptions are

Additional Requirement: Document on Github

In order for you to work together efficiently and for me to assess your progress please start a Github project

- Add your updated code each time after you worked on it
- Be sure to annotate and structure your code clearly
- Add results if you have them
- Invite me to the project

Note that, my coding language of preference is R. I cannot guarantee the same help for other languages.

The Use of Generative AI

I strongly encourage the use of generative AI for many purposes:

- Help with coding
- Discussing ideas
- Help with formulating ideas or finding the correct terminology
- Helping with text structuring

But... Be warned:

- The school has software that detects texts written by generative AI
- Generative AI may make up sources and provide factually incorrect answers that seem well motivated

Hence:

- Always check the validity of the answers
- Do not simply copy text; think and rewrite

The Basics

In general, no one likes reading the literature section, especially when they are unstructured

1. Have structured topics in the literature section
2. Only cite papers if you have a reason to!
    - What reasons are there to cite other literature?
3. Make sure your sources are reputable and relevant
    - Examples of reputable sources: Management Science, Marketing Science, Journal of machine learning Research, Neurips. (find out if the journal you cite from is of high reputation!)
    - Is the source relevant for this part of your literature section?

Let's consider a simple example structure

Example literature review

(This is an example, do not literally copy this)

1. First, I want to motivate that my work is relevant
   - Show that within marketing science or management science researchers implement bandits for various problem settings (and have good performance)
2. Second, we may want to motivate bandits for recommendation systems in general. What do we know about it?
   1. Let us know about why it is a good idea to implement bandits for a recommendation system based on the literature.
3. Third, we motivate the choice for the algorithms we implement, why those and not others?

Finally, lets consider an example citation

Example Reasons for Citations

Paper A shows that X is True. (use correct citation format)

1. Motivate a choice: Hence, we will use their method for Y
2. Place yourself in the literature: But they do not show Z is true
3. Show the topic is important: See, everyone uses bandits for such settings and they outperform other methods

Discussion and Questions

1. Questions?
2. Points you want to discuss?
3. preferences for next coaching sessions?