

# Lecture II - Explore Vs. Exploit

## Programming: Everyday Decision-Making Algorithms

Dr. Nils Roemer

Kühne Logistics University Hamburg - Winter 2025

### Explore Vs. Exploit

#### Some definitions...

1. Question: What does “explore” mean?
  - Explore is the gathering of new information.
2. Question: What does “exploit” mean?
  - Exploit is the utilization of already known information to obtain a known result.
3. Question: what is the relationship between both?
  - Explore and exploit are opposing alternatives.

#### Some examples I

- Clinical trials:
  - Explore: Test new drugs.
  - Exploit: Use existing drugs.
- A/B testing:
  - Explore: Test new website designs.
  - Exploit: Use existing website designs.

#### Some examples II

- Dating:
  - Explore: Go on a date with someone new.
  - Exploit: Go on a date with someone you already know.
- Social interactions:
  - Explore: Meet new people.
  - Exploit: Spend time with known people.

#### Everyday decision-making

- Explore: Do we try new things?
- Exploit: Do we stick to our favorite ones?
- Life is a trade-off, a balance:
  - between novelty and tradition.
  - between the latest and the greatest.
  - between explore and exploit.
- Question: What is the optimal balance?
- Scientists have been working on this for over 50 years.

## The problem with exploitation

Question: Any ideas what it might be?

- Exploitation is not always the best strategy.
- Especially when you have very limited information.
- When you stop exploring, you might miss better options.
- Imagine you are not able to gather new information and could only choose known alternatives.

## The problem with exploration

Question: Any ideas what it might be?

- Exploration is not always the best strategy.
- Especially when you are limited in using new information.
- When you constantly explore, you might never enjoy the fruits of your exploration.
- Imagine you could only eat each meal only once, hear each song only once, talk to each person only once...

## The Multi-Armed Bandit Problem

### Decision Support

- To provide support, computer scientists formulated the explore-exploit trade-off.
- It is known as the multi-armed bandit problem.
- Question: What is a one-armed bandit?

### One Armed Bandit

Photo by Kabir M on Unsplash

### The multi-armed bandit problem I

- A gambler is faced with a room of slot machines (one-armed bandits).
- Each slot machine has a different probability of winning.
- Question: What does the scenario have to do with explore vs exploit?

### The multi-armed bandit problem II

- By playing a slot machine, the gambler can gather information about the probability of winning.
- But each pull of a lever comes with a certain cost.
- It's the aim of the gambler to maximize his winnings.

### The multi-armed bandit problem III

- Consider the following scenario:
  - You already pulled the lever of two machines.
  - Machine A: 15 pulls, 9 wins.
  - Machine B: 2 pulls, 1 win.
- Question: Which machine should you play next?

### Expected value as a decision criterion

- The expected value of a slot machine is the average number of wins per pull.
- Expected value of machine A =  $E(A) = 9/15 = 0.6$
- Expected value of machine B =  $E(B) = 1/2 = 0.5$
- Machine A has the higher expected value.
- But 2 and even 15 pulls are not a large number (considering standard deviation).

### The multi-armed bandit problem IV

- The multi-armed bandit problem represents a lot of different real-world decisions.
- It shows us, that there might be a difference between the optimal long-term average performance and the optimal immediate performance.
- Which lever to pull next depends completely on something we haven't discussed yet:

### The multi-armed bandit problem V

- How long you plan to be in the casino?
- Question: Why is this important?
- Question: How does this influence our decision on taking machine A or machine B?

...

"I'm more likely to try a new restaurant when I move to a city than when I'm leaving it" (Chris Stucchio)

### The influence of the interval

- Let's call the time you plan to be in the casino "the interval".
- The longer the interval, the more (in general) you should explore, since you will have more opportunities to exploit the gathered information.
- The shorter the interval, the more you should exploit your current knowledge.
- The optimal strategy depends on the length of the interval.

### Interval and Exploration

- Explore when you have the time to use the resulting knowledge.

...

"I moved to Pune, India, and I just [...] eat everywhere that didn't look like it was gonna kill me" (Chris Stucchio)

### Interval and Exploitation

- Exploit when you are ready to cash in.

...

"And as I was leaving the city I went back to all my old favorites, rather than trying out new stuff [...]. Even if I find a slightly better place, I'm only going to go there once or twice, so why take the risk?" (Chris Stucchio)

## Exploration and Exploitation

### Exploration

Photo by Colin Maynard on Unsplash

### Exploitation

Photo by Cristina Gottardi on Unsplash

### Reverse Engineering

- Derivation of the interval by observing the strategy
- Among the the ten highest-grossing movies, how many were sequels?
  - 1981: 2
  - 1991: 3
  - 2001: 5
  - 2011: 8
- Question: Do you have an explanation for the trend?

### Sequels...

Photo by Universal Pictures on Wikipedia

### Reverse Engineering: A possible explanation

- Making a brand new movie is risky but has the potential to create a new fan base. (explore)
- From a Studio's perspective, a sequel is a movie with a guaranteed fan base, a cash cow, a sure thing, an exploit.
- One possible explanation for the numbers is that the studios think they are approaching the end of their interval.
- They are pulling the arms of the best machines they've got before the casino turns them out.

### The multi-armed bandit problem VI

- While the so far provided anecdotes are helping us to understand the explore-exploit trade-off, they are far away from being a satisfying "optimal" solution.
- Actually, finding an algorithm that tells us exactly how to handle the trade-off is a very hard problem.
- On the way there were many interesting approaches...

## Win-Stay Lose-Shift

### Win-Stay Lose-Shift<sup>1</sup>

- Question: What do you think, what the win-stay lose-shift strategy does?
- The win-stay lose-shift strategy is a simple strategy that is often used in multi-armed bandit problems.

---

<sup>1</sup>For more details see Robbins, H. (1952) 'Some aspects of the sequential design of experiments', Bulletin of the American Mathematical Society, 58.

- It is based on the idea that if you have won, you should stay with the same machine.
- If you have lost, you should switch to a different machine.
- This strategy is not always the best strategy, but it is simple and proven better than choose an arm at random.

### Win-Stay is a no brainer

- Question: What do you think about win-stay?
- If you decide to pull an arm and you win, you should pull the same arm again.
- Nothing changes, except the attractiveness of the arm you just pulled is higher.

### Lose-Shift is another story

- Question: What do you think about lose-shift?
- Changing arms each time you lose might be a pretty rush move.
- Imagine you're eating at your favorite restaurant for the tenth time in a row.
- You have always been very satisfied (win), but today you are disappointed (lose).
- Should you turn your back on the restaurant?

### Like most of the time, easy answers comes with problems

- The win-stay lose-shift strategy penalizes losses too much.
- The strategy does not take into account the interval.
- But the strategy was good start to develop better approaches.

## The Bellman Approach

### The Bellman approach I

- Few years later, Richard Bellman, found an exact solution to the problem for all cases with finite and known intervals.
- Bellman found that under the given assumptions, exploit vs explore can be formulated as an optimal stopping problem.
- Where the question is, when to stop exploring and start exploiting.
- The solution is based on dynamic programming and backward calculation starting from the final pull (analogous to the secretary problem with full information).

### The Bellman approach II

- Bellman found that the following equation provides the optimal strategy (when the assumptions hold):

$$\mathbb{E}[B] = \frac{w + 1}{w + l + 2}$$

- where w is the number of wins and l is the number of losses.

### Question time

$$\mathbb{E}[B] = \frac{w + 1}{w + l + 2}$$

- Question: What is  $E[B]$ ?
- Question: What is  $E[A]$ ?
- Question: What machine shall we play according to the Bellman approach?

### The Bellman approach III

The following table shows the expected value for different win-lose scenarios.

Losses/Wins	1	2	3	4	5	6	7	8	9	10
1	0.50	0.60	0.67	0.71	0.75	0.78	0.80	0.82	0.83	0.85
2	0.40	0.50	0.57	0.63	0.67	0.70	0.73	0.75	0.77	0.79
3	0.33	0.43	0.50	0.56	0.60	0.64	0.67	0.69	0.71	0.73
4	0.29	0.38	0.44	0.50	0.55	0.58	0.62	0.64	0.67	0.69
5	0.25	0.33	0.40	0.45	0.50	0.54	0.57	0.60	0.63	0.65
6	0.22	0.30	0.36	0.42	0.46	0.50	0.53	0.56	0.59	0.61
7	0.20	0.27	0.33	0.38	0.43	0.47	0.50	0.53	0.56	0.58
8	0.18	0.25	0.31	0.36	0.40	0.44	0.47	0.50	0.53	0.55
9	0.17	0.23	0.29	0.33	0.38	0.41	0.44	0.47	0.50	0.52
10	0.15	0.21	0.27	0.31	0.35	0.39	0.42	0.45	0.48	0.50

### The Bellman approach IV

- The calculation of the optimal strategy is very extensive when there are many arms and a long interval.
- And yet the approach does not help us in many scenarios because we do not know the exact length of the interval (time in the casino).
- At this point, it looked like the multi-armed bandit problem would remain a problem without a solution.

## The Gittins Index

### The Gittins Index I<sup>2</sup>

- In the 1970s Unilever asked a young mathematician, John Gittins, to help optimize their drug trials.
- Given different compounds, what is the quickest way to determine which is likely to be effective?
- Gittins found an optimal strategy and abstracted the problem to a general level.
- He found the solution to the multi-armed bandit problem.

---

<sup>2</sup>Gittins, J. (1979) 'Bandit Processes and Dynamic Allocation Indices', Journal of the Royal Statistical Society. Series B (Methodological).

## The Gittins Index II

- A major problem with the multi-armed banded problem is that previous solutions made very critical assumptions about the underlying interval.
- For example, that the length of the interval is known at the beginning of the analysis.
- Gittins developed a charming solution to this problem. In his approach, future wins (e.g., cash flows) are discounted so that any interval length (including infinity) can be considered<sup>3</sup>.

## Reality check: Discounting I

- Does discounting future wins make sense?
- Question: Does discounting money wins make sense?
- Regarding monetary wins, it does. For example, due to interest rates and opportunity costs.

## Reality check: Discounting II

- Does discounting future wins make sense?
- Question: Does discounting non-monetary wins make sense?
- Regarding non-monetary wins, it is more difficult to justify.
- But its not counterintuitive.
- What is more important to you today, tonight's dinner, or ceteris paribus the dinner in a week's time?

## The Gittins Index III

- The Gittins index can be used for any problems of the form of the multi-armed bandit problem.
- That means it solves the explore-exploit trade-off.
- Let's consider our machine A and B example one last time.
  - Machine A: 15 pulls, 9 wins, 6 losses.
  - Machine B: 2 pulls, 1 win, 1 lose.

## The Gittins Index IV

Losses/Wins	0	1	2	3	4	5	6	7	8	9
0	.7029	.8001	.8452	.8723	.8905	.9039	.9141	.9221	.9287	.9342
1	.5001	.6346	.7072	.7539	.7869	.8115	.8307	.8461	.8588	.8695
2	.3796	.5163	.6010	.6579	.6996	.7318	.7573	.7782	.7956	.8103
3	.3021	.4342	.5184	.5809	.6276	.6642	.6940	.7187	.7396	.7573
4	.2488	.3720	.4561	.5179	.5676	.6071	.6395	.6666	.6899	.7101
5	.2103	.3245	.4058	.4677	.5168	.5581	.5923	.6212	.6461	.6677
6	.1815	.2871	.3647	.4257	.4748	.5156	.5510	.5811	.6071	.6300

---

<sup>3</sup>Gittins makes a geometric discounting assumption, but the approach can be extended to other discounting assumptions.

Losses/Wins	0	1	2	3	4	5	6	7	8	9
7	.1591	.2569	.3308	.3900	.4387	.4795	.5144	.5454	.5723	.5960
8	.1413	.2323	.3025	.3595	.4073	.4479	.4828	.5134	.5409	.5652
9	.1269	.2116	.2784	.3332	.3799	.4200	.4548	.4853	.5125	.5373

Question: Choose machine A or B according to the Gittins index?

- The index for machine B (0.6346) is higher than for machine A (0.6300).
- The index shows a clear win-stay pattern.
- There is a relaxed lose-shift pattern.
- At the (0,0) point we see the exploration bonus (premium).
- The index converges to 1/2 for a 50/50 chance game.

### The Gittins Index V

- The problem with the Gittins index is that it is very difficult to calculate.
- See the following equation:

$$G_i(s_i, f_i) := \sup_{\tau \geq 1} \frac{\mathbb{E} \left[ \sum_{t=0}^{\tau-1} \beta^t \cdot r_i^t \mid s_i, f_i \right]}{\mathbb{E} \left[ \sum_{t=0}^{\tau-1} \beta^t \right]}$$

- Where  $G_i(s_i, f_i)$  is the Gittins index for machine  $i$ ,  $s_i$  is the number of wins,  $f_i$  is the number of losses,  $\beta$  is the discount factor, and  $r_i^t$  is the reward for machine  $i$  at time  $t$ .

## Explore vs Exploit: Summary

### Explore vs Exploit: Summary

- Consider an explore vs exploit decision situation.
- As you learned exploiting comes with a known (expected) outcome for example  $E(A) = 0.6$
- Exploring comes with an unknown outcome  $E(B) = ?$
- What should you do according to decision science?

### Explore vs Exploit: Anecdotaly

- If you have a long interval, you should explore, choose B untill you are sure about  $E(B)$ .
- If you have a short interval, you should exploit, choose A.

### Explore vs Exploit: Mathematically

- If  $E(A)$  and  $E(B)$  are known, choose higher expected value.
- If  $E(B)$  is unknown, but you know the length of the interval, the Bellman-approach provides the optimal strategy.
- If  $E(B)$  is unknown, and you do not know the length of the interval, the Gittins index provides the optimal strategy.



## Explore vs Exploit: Key Takeaways

“The grass is always greener on the other side of the fence.”

- The math tells us why:
- Exploration in it self has a value, since trying new things increases our chance of finding the best.
- Your todays takeaway from the lecture should be: Be sensitive to how much time you have left in the casino and explore, explore, explore...

## Literature

### Interesting literature to start

- Christian, B., & Griffiths, T. (2016). Algorithms to live by: the computer science of human decisions. First international edition. New York, Henry Holt and Company.<sup>4</sup>
- Ferguson, T.S. (1989) ‘Who solved the secretary problem?’, Statistical Science, 4(3). doi:10.1214/ss/1177012493.

### Books on Programming

- Downey, A. B. (2024). Think Python: How to think like a computer scientist (Third edition). O’Reilly. [Here](#)
- Elter, S. (2021). Schrödinger programmiert Python: Das etwas andere Fachbuch (1. Auflage). Rheinwerk Verlag.

...

#### Note

Think Python is a great book to start with. It’s available online for free. Schrödinger Programmiert Python is a great alternative for German students, as it is a very playful introduction to programming with lots of examples.

## More Literature

For more interesting literature, take a look at the [literature list](#) of this course.

---

<sup>4</sup>The main inspiration for this lecture. Nils and I have read it and discussed it in depth, always wanting to translate it into a course.