Introduction to Multi-Armed Bandits

Fundamentals of Reinforcement Learning

Institut für Nachrichtentechnik
Fachgebiet Kommunikationstechnik
Prof. Dr.-Ing. Anja Klein
Dr. Sabrina Klos & Dr. Andrea Ortiz

Lecture Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

simplified setiing - exxplotation/explanation

Chapter 1
Introduction

Chapters 2-3
Probability Theory

Chapters 4-6
The Multi-Armed Bandit Problem

Decisions do not influence future data

With/without context

Chapters 7-12
The full RL Problem

Decisions may influence future data

With/without knowledge of dynamics

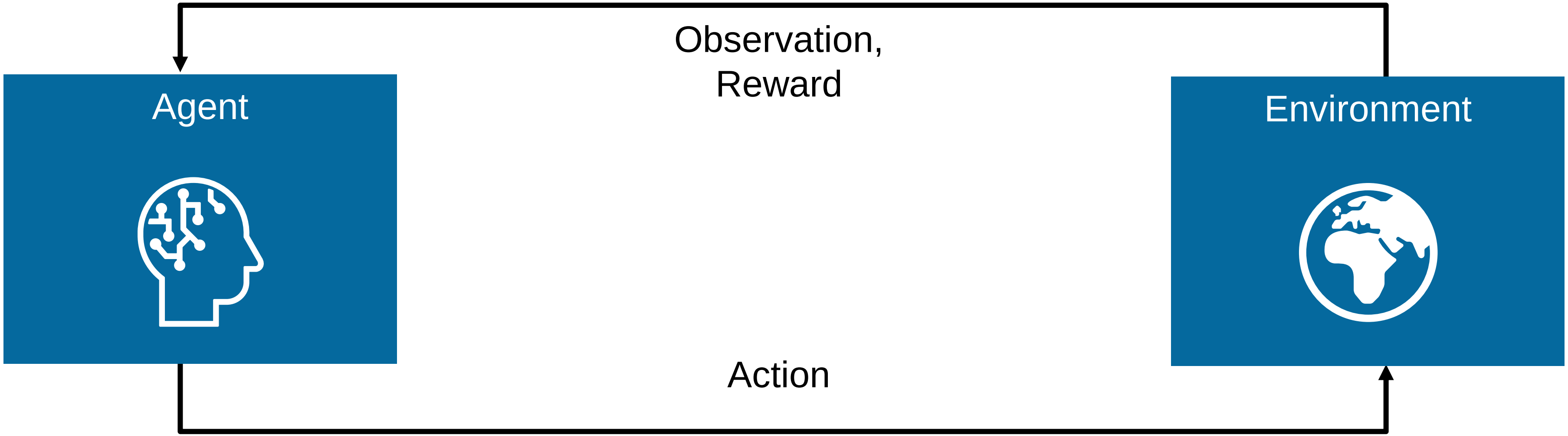
Chapter 13
Case Study

Chapter 14
Extensions

Recap: Idea of RL

RL deals with goal-directed learning from interaction

Agent–environment interaction



Recap: Characteristics of RL

RL deals with associative settings with evaluative and delayed feedback

TECHNISCHE UNIVERSITÄT DARMSTADT

Characteristics of RL	
Evaluative Feedback	There is no supervisor, only a reward signal, i.e., trial-and-error search needed.
Delayed Feedback	Reward feedback may be delayed, not instantaneous.
Sequential and Associative Setting	Time really matters, i.e., sequential non i.i.d data, and best action depends on situation.
Influence on Environment	Actions may affect subsequent situations and rewards, i.e., actions may have long term consequences.

Recap: Exploration and Exploitation

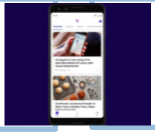
A challenge in RL is how to balance exploration and exploitation

TECHNISCHE UNIVERSITÄT DARMSTADT

Reinforcement Learning

Balance

	Exploration	**Exploitation**
Goal	Gather more information to achieve high reward in future.	Exploit current information to achieve high reward now.
Approach	Select undiscovered actions.	Select best previous actions.
Example	Show new article in news feed.	Show the so far most clicked article in news feed.



Today's topic

Study the exploration-exploitation dilemma in a simplified version of RL

TECHNISCHE UNIVERSITÄT DARMSTADT

Multi-Armed Bandits (MABs): A simplified version of RL

Evaluative Feedback	There is no supervisor, only a reward signal, i.e., trial-and-error search needed.
Immediate Feedback	Reward feedback is instantaneous.
Sequential, but Non-Associative Setting	There is only one situation, i.e., i.i.d data.
No Influence on Environment	Actions only affect immediate rewards.

Learning Goals

- You can explain the differences between full Reinforcement Learning and Multi-armed Bandits.
- You can name and explain the main modeling dimensions of Multi-Armed bandit models.
- You can model decision-making problems using the stochastic Multi-armed bandit model.

Outline

- Introduction
- Taxonomy of MAB models
- Stochastic Bandits: Model & Examples

Outline

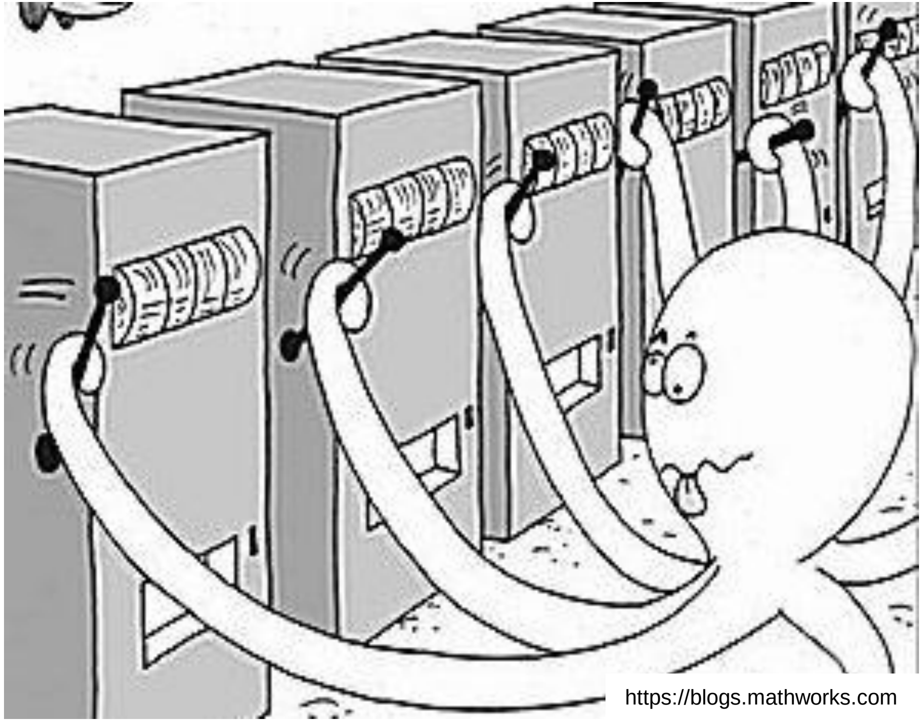
- **Introduction**
- Taxonomy of MAB models
- Stochastic Bandits: Model & Examples

Motivation

The term „multi-armed bandits“ comes from a stylized gambling scenario

lever

TECHNISCHE UNIVERSITÄT DARMSTADT

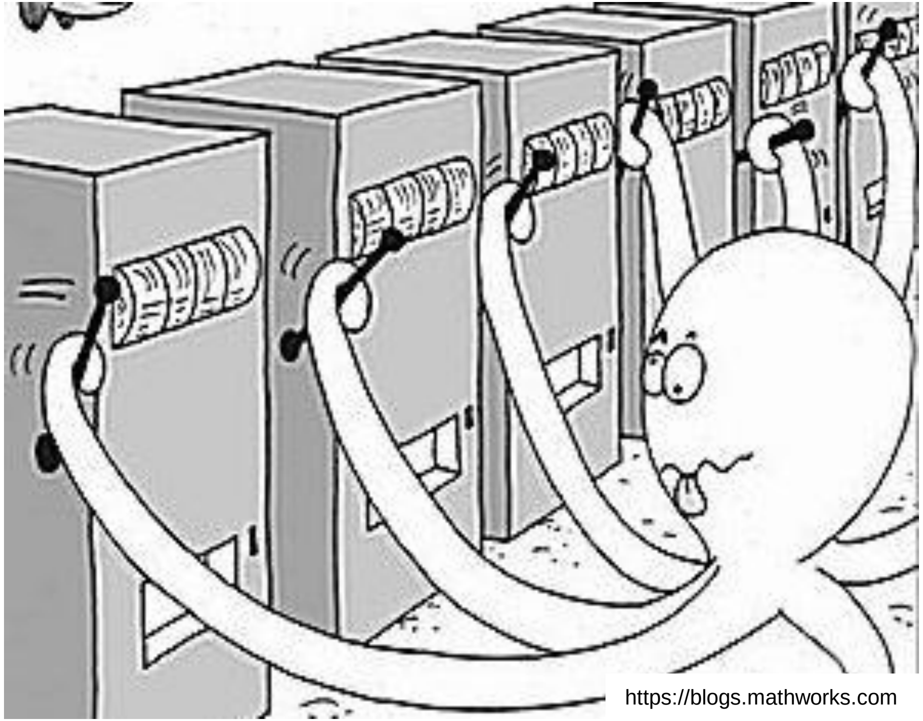


https://blogs.mathworks.com

- You face a slot machine with several levers.
- Each lever („bandit“) yields a different payout.
- You don‘t know which lever has the highest payout.
- You just have to try different levers to see which one works best.

Question

How would you go about to maximize your sum payout?

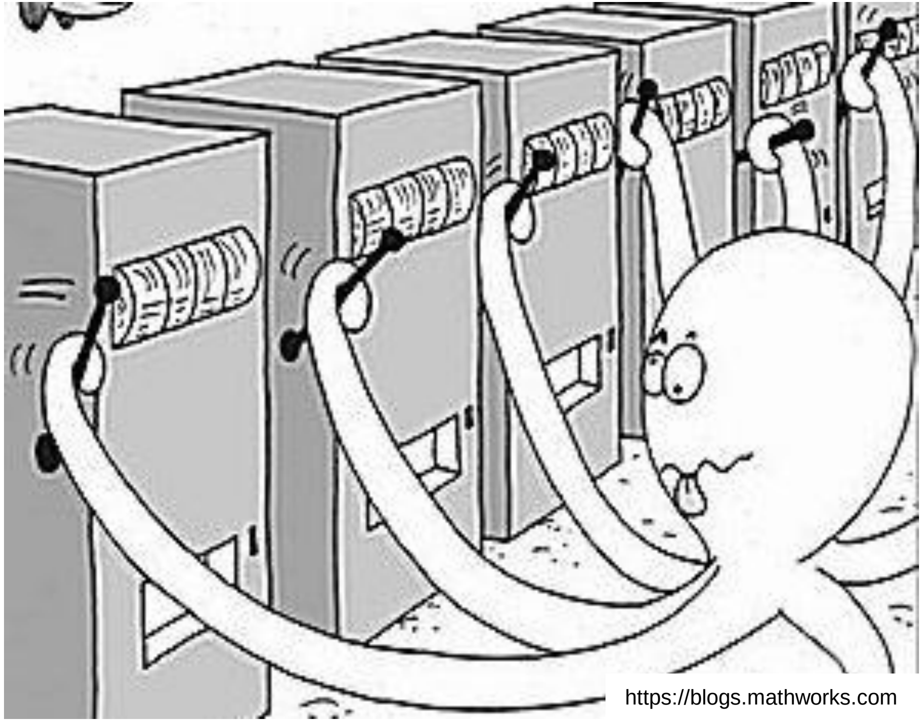


https://blogs.mathworks.com

- You face a slot machine with several levers.
- Each lever („bandit“) yields a different payout.
- You don‘t know which lever has the highest payout.
- You just have to try different levers to see which one works best.

Motivation

Difficulty for gambler is how to balance exploration and exploitation



https://blogs.mathworks.com

The balance between exploration and exploitation is essential in multi-armed bandits.

You Need Exploitation

If you keep pulling the low payout levers too often, you loose too much payout along the way.

You Need Exploration

But you won‘t know which lever is good until you try a sufficient number of times.

Applications of Multi-Armed Bandits (MABs)

Example: Maximize revenue from online banner advertisement

multi-armed : os diferentes possiveis anuncios são alavancas independentes (nesse caso)
os diferentes momentos t na vdd são diferentes usuarios possivelmente

https://www.bannerflow.com

Online banner advertisement

Task:

When a new user arrives, a website picks a banner ad to display and receives some revenue if the user clicks on this header.

Goal:

The site's goal is to maximize the revenue from the clicked ads.

Applications of Multi-Armed Bandits (MABs)

Example: Maximize followed music recommendations

TECHNISCHE UNIVERSITÄT DARMSTADT


https://hellofuture.orange.com

Music recommendation

Task:

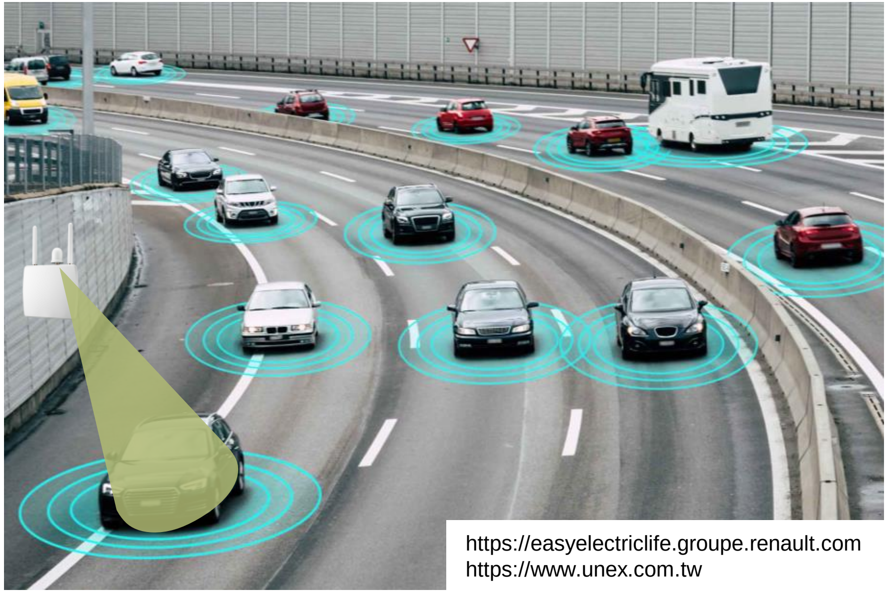
When a new user arrives, a recommender system picks a song to show the user and observes whether the user follows the recommendation.

Goal:

The system's goal is to maximize the number of followed recommendations.

Applications of Multi-Armed Bandits (MABs)

Example: Maximize throughput by beam selection in vehicular communications



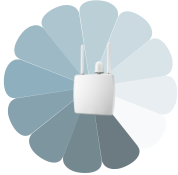
https://easyelectriclife.groupe.renault.com
https://www.unex.com.tw

Antenna beam selection

esse modelo não tem nenhum sensor para trackear o carro

Task:

When a new vehicle arrives, the base station picks one of its directional antenna beams for data transmission to the vehicle.



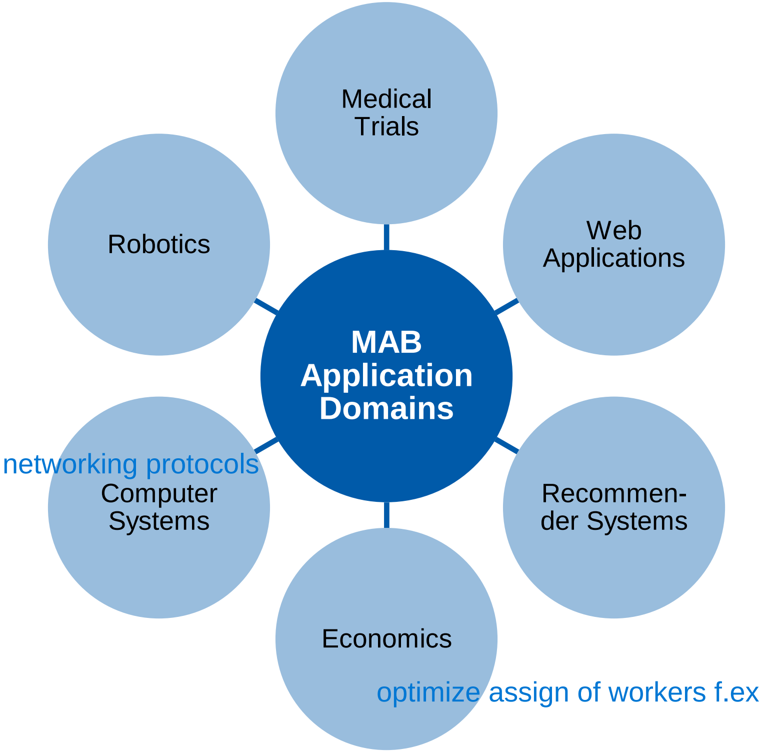
base station with beams

Goal:

The base station's goal is to maximize the amount of data received by the vehicle.

Applications of Multi-Armed Bandits (MABs)

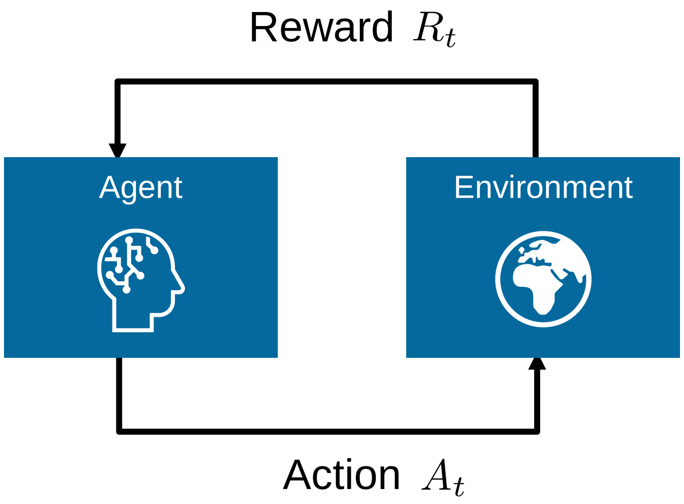
MABs have applications in several domains



Overview of Basic MAB Model

The agent and the environment interact sequentially

TECHNISCHE UNIVERSITÄT DARMSTADT



At each time step t:

The agent

- Executes action A_t
- Receives scalar reward R_t

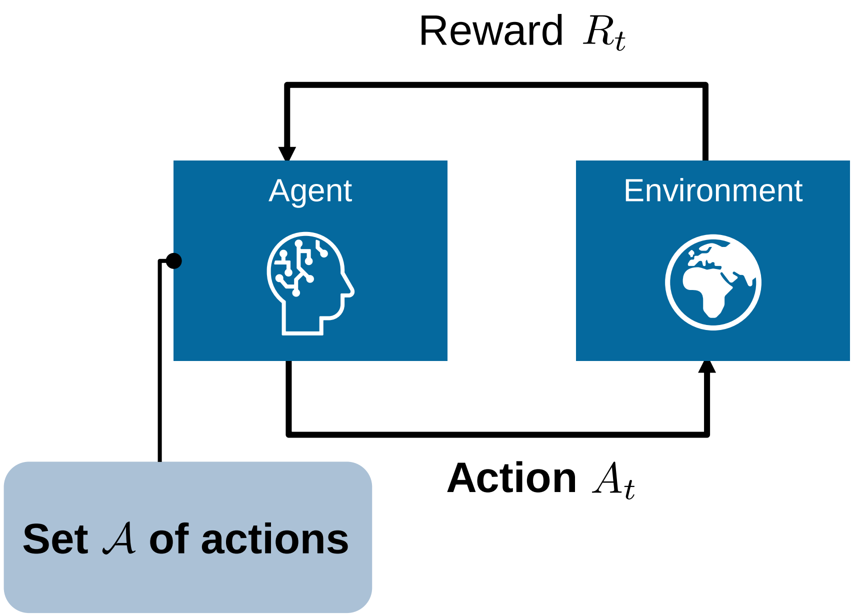
The environment

- Receives action A_t
- Emits scalar reward R_t

In MABs, there is no observation of state.

Actions

Agent selects action from fixed finite set of actions



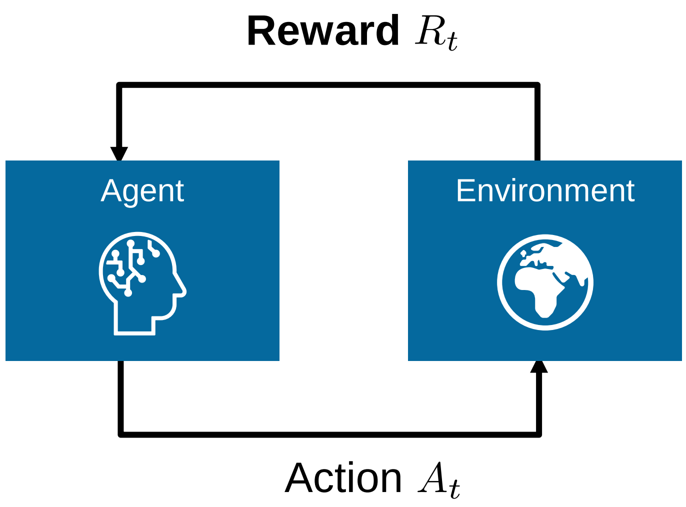
- Actions are the decisions the agent wants to learn how to make.

- **Set $\mathcal{A}$ of actions:** In each time step t, the agent selects an action (or „arm“)
  $$A_t \in \mathcal{A}$$
  from the fixed finite set $\mathcal{A}$.

- Action $A_t \in \mathcal{A}$ selected in time step t only affects immediate reward R_t, but not future rewards.

Rewards

Agent receives reward as immediate consequence for selected action



- Rewards indicate how well agent is doing in selecting actions.

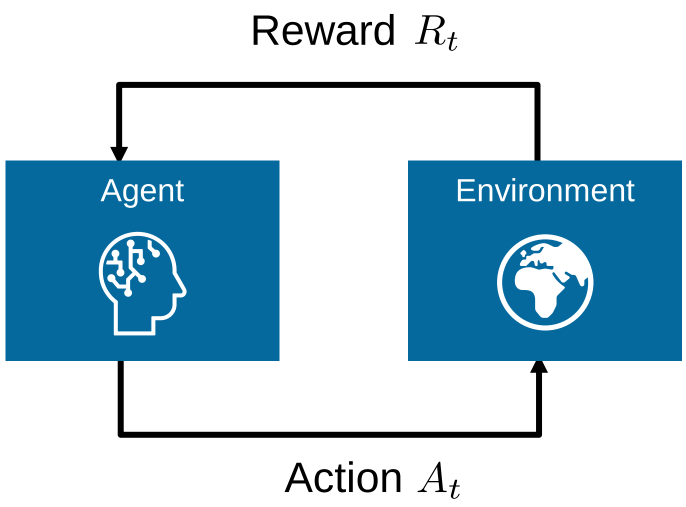
- **Reward** R_t : Scalar feedback signal received by the agent as **immediate** consequence of (only) its action in time step t.

P(Rt|At)

- Formally, reward is drawn from a **stationary probability distribution** that depends on the selected action and is **unknown to the agent**.

Goal and Challenges

Goal of sequential decision making is to maximize cumulative reward



Goal:

The agent seeks to select actions
to maximize the cumulative reward.

Challenges:

- **Lack of prior knowledge:**
  Expected reward of each action is unknown.
- **Evaluative feedback:**
  The agent only observes the instantaneous reward for the selected action, but not for other actions.
- **Balance exploration and exploitation:**
  Sacrificing immediate reward may lead to more long-term reward.

Actions and Rewards
Example: Online banner advertisement

https://www.bannerflow.com

Online banner advertisement

Action:

A banner ad to display.

Reward:

Revenue from ad if clicked,

0 otherwise.

Goal and Challenges

Example: Online banner advertisement

https://www.bannerflow.com

Online banner advertisement

Goal:

Maximize revenue from clicked banner ads.

Challenges:

- **Lack of prior knowledge:**
  Expected click rate (and hence revenue) of each banner ad is unknown.
- **Evaluative feedback:**
  The website observes only whether the user clicked on the displayed ad, but not whether the user would have clicked on others.
- **Exploration vs. Exploitation:**
  Display untried vs. so far most clicked ad.

Question
What are actions and rewards in music recommendation?

https://hellofuture.orange.com

Music recommendation

Answer
Actions are songs, rewards depend on followed recommendations


https://hellofuture.orange.com

Music recommendation

Action:

A song to recommend.

Reward:

1 if user follows recommendation,

0 otherwise.

Question
What are the goal and challenges in music recommendation?

https://hellofuture.orange.com

Music recommendation

Answer

Maximize no. of followed recommendations despite unknown song popularity.


https://hellofuture.orange.com

Music recommendation

Goal:

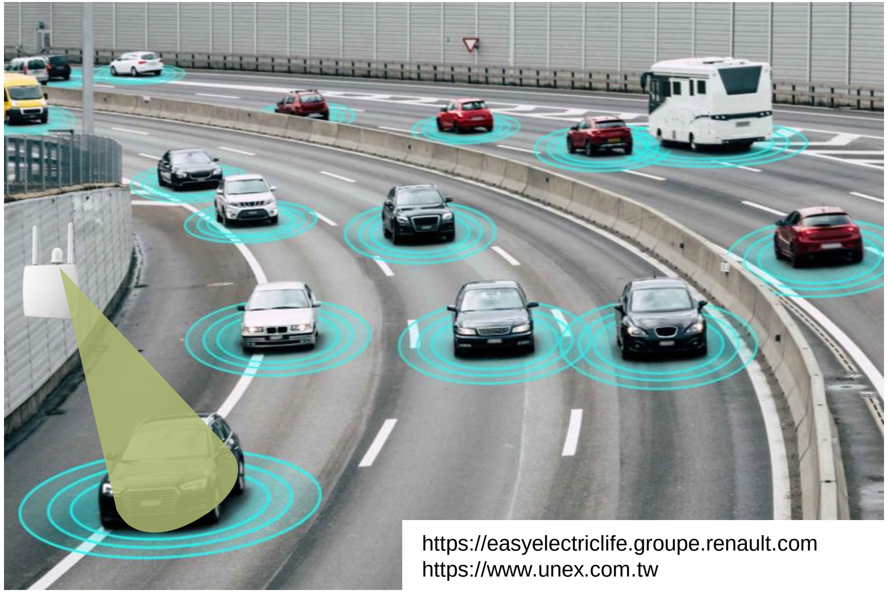
Maximize number of followed recommendations.

Challenges:

- **Lack of prior knowledge:**
  Expected popularity of each song is unknown.
- **Evaluative feedback:**
  The system only observes whether the user followed the recommendation for the shown song, but not whether the user would have followed others.
- **Balance exploration and exploitation:**
  Show untried vs. Previously most popular song.

Actions and Rewards

Example: Beam selection in vehicular communications



https://easyelectriclife.groupe.renault.com
https://www.unex.com.tw

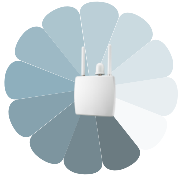
Antenna beam selection

Action:

An antenna beam to transmit data to vehicle.

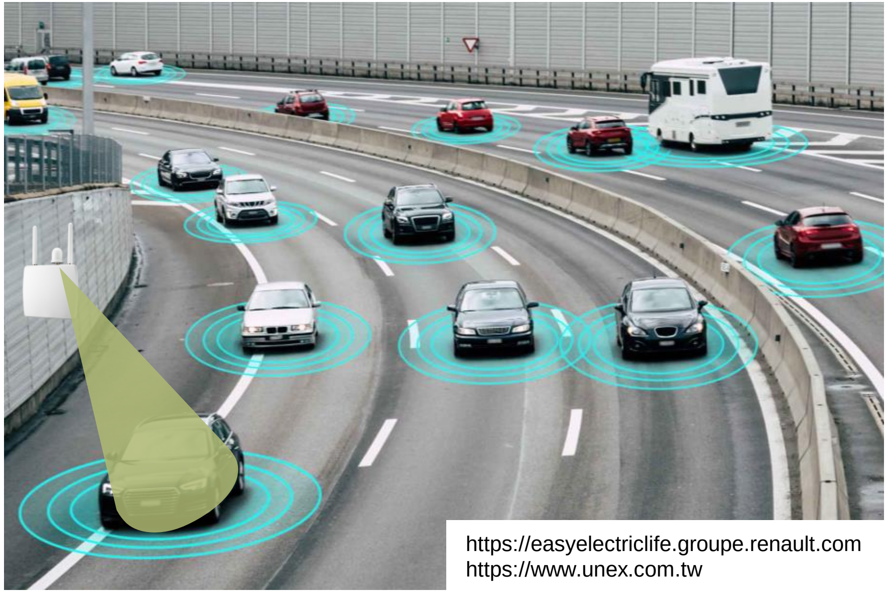
Reward:

Amount of data received by vehicle.



base station with beams

Goal and Challenges
Example: Beam selection in vehicular communications



https://easyelectriclife.groupe.renault.com
https://www.unex.com.tw

Antenna beam selection

Goal:

Maximize amount of data received by vehicle.

Challenges:

- **Lack of prior knowledge:**
  Expected data rate of each beam is unknown.
- **Evaluative feedback:**
  The base station observes only the data rate of selected beam, but not the data rate of other beams.
- **Balance exploration and exploitation:**
  Select untried beam vs. beam with previously highest data rate.

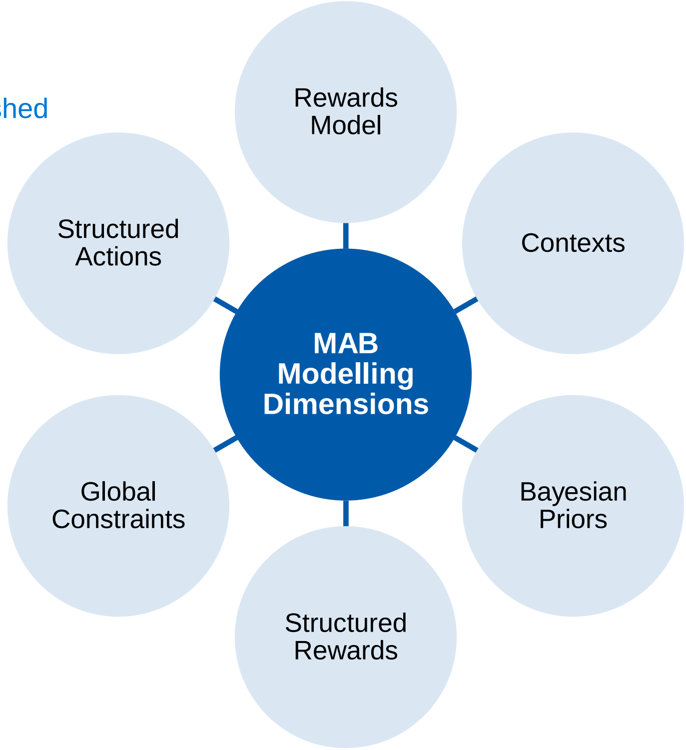
Outline

- Introduction
- **Taxonomy of MAB models**
- Stochastic Bandits: Model & Examples

Taxonomy of MAB models

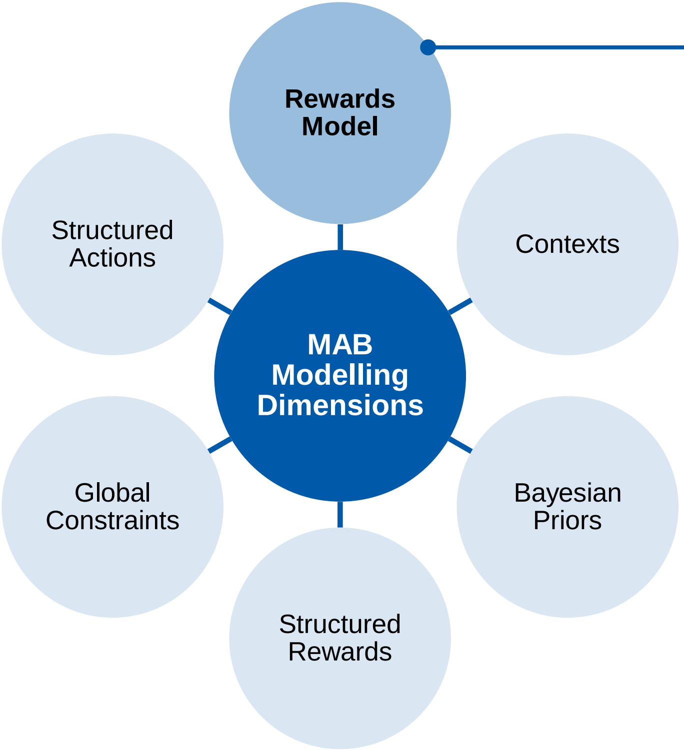
MABs cover a large problem space, with many modelling dimensions

which dimensions they can be distinguished



Taxonomy of MAB models

Rewards can be modelled as stochastic, adversarial or constrained adversarial



Stochastic Rewards

IID Rewards

Reward drawn independently from fixed distribution that depends on action, but not on the round t.

Non-IID Rewards

Rewards evolve over time as a random process, e.g., a random walk.

Adversarial Rewards

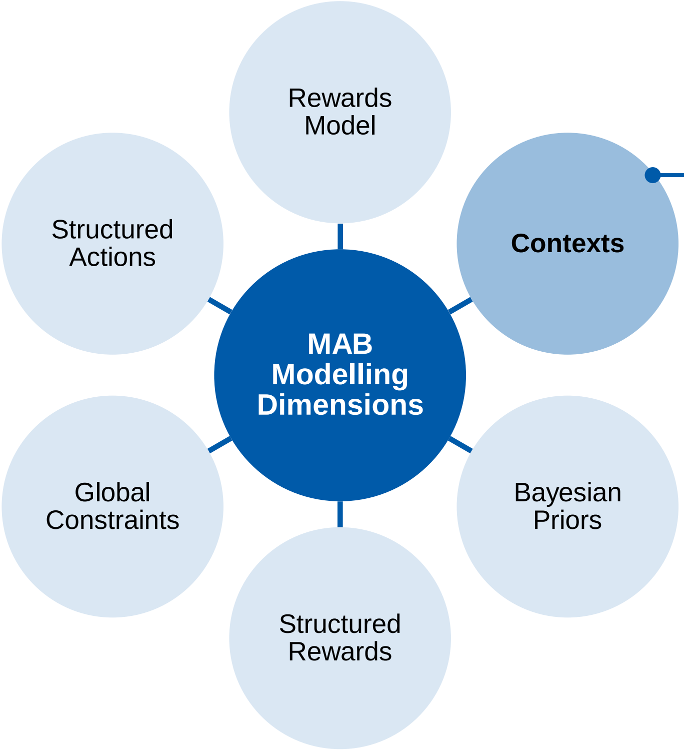
Rewards can be arbitrary, as if they are chosen by an adversary that tries to fool the agent.

Constrained Adversarial Rewards

Rewards are chosen by an adversary that is subject to some constraints. E.g., reward of each arm cannot change much from one round to another.

Taxonomy of MAB models

The agent may observe some context before selecting an action



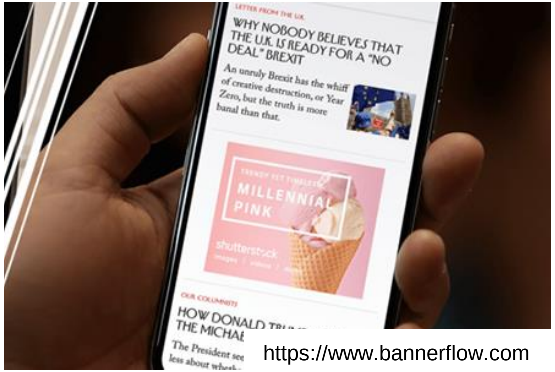
Contextual MABs

- In each round, the agent observes some context before selecting an action.
- E.g., known properties of current user.
- In contrast to the *state* in full RL, the **agent‘s actions do not influence context**.
- **Associative setting:** Best action depends on *situation*.
- **Agent‘s goal**: Learning the *best mapping from contexts to actions* (while not spending too much time exploring).

The associative setting of contextual MABs brings us one step closer to the full RL problem. ⟶ Chapter 6

Question

Can you think of possibly relevant context in the three examples?

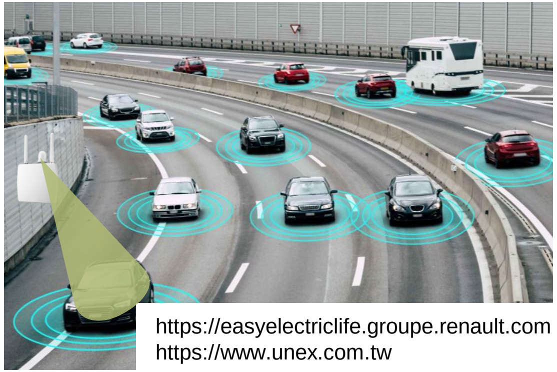


https://www.bannerflow.com

Online banner advertisement

https://hellofuture.orange.com

Music recommendation

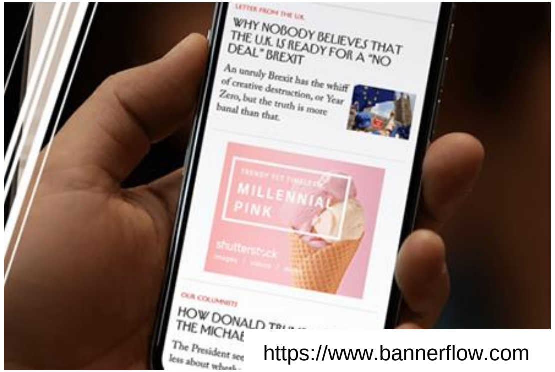


https://easyelectriclife.groupe.renault.com
https://www.unex.com.tw

Antenna beam selection

Answer

Context can be any information related to the current situation

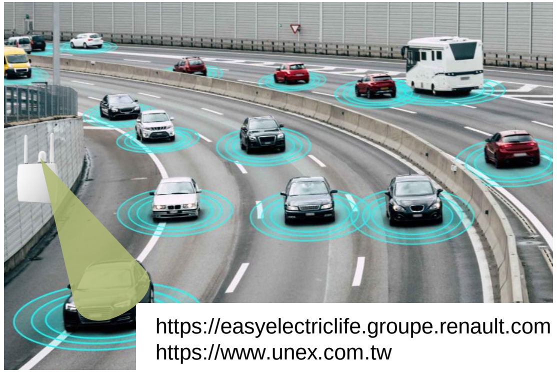


https://www.bannerflow.com

Online banner advertisement

https://hellofuture.orange.com

Music recommendation



https://easyelectriclife.groupe.renault.com
https://www.unex.com.tw

Antenna beam selection

Examples:

Online banner advertisement and music recommendation:

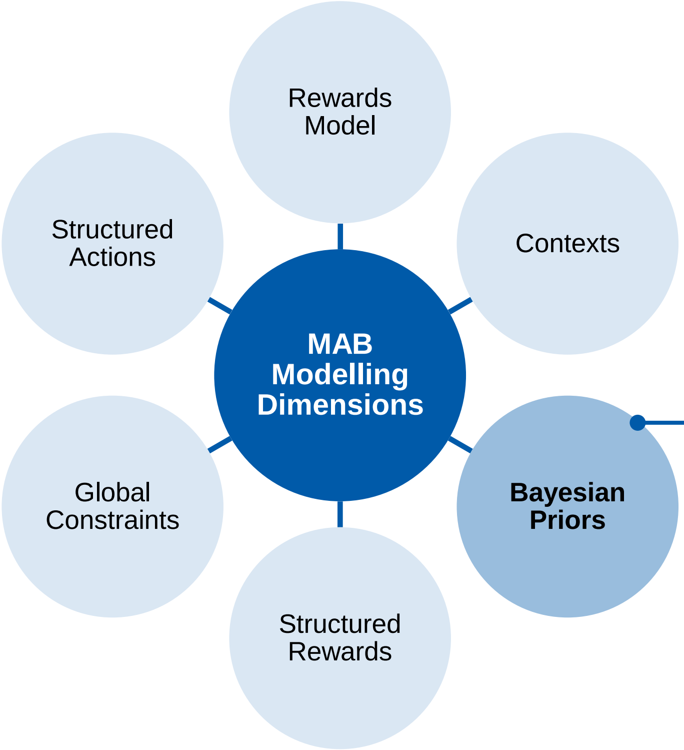
- User location
- User demographics
- Type of device

Antenna beam selection:

- Vehicle‘s direction of arrival
- Vehicle speed
- Type of vehicle
- Fluctuations of wireless channel

Taxonomy of MAB models

Each MAB problem can be studied under a Baysian approach



Bayesian Priors

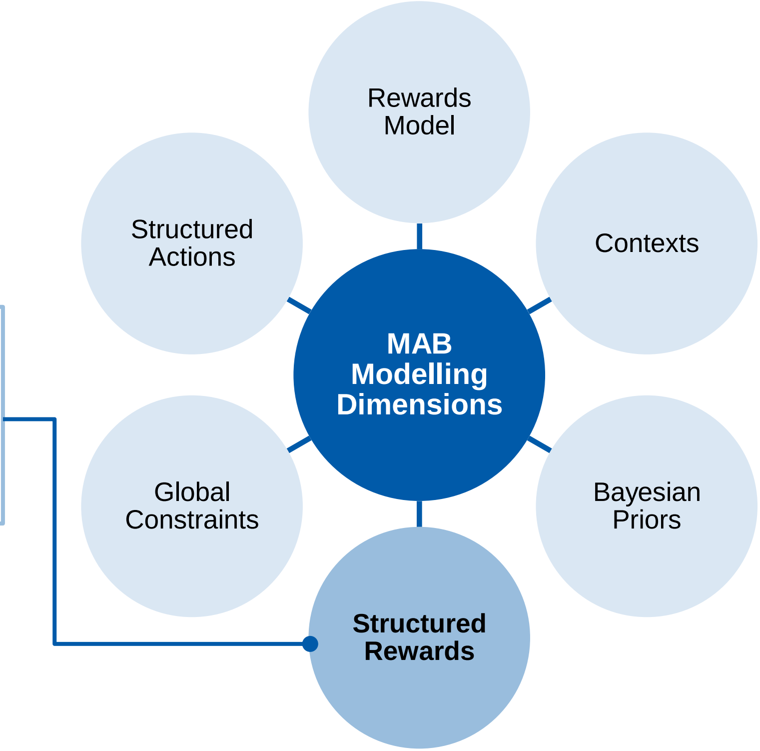
- Each MAB problem can be studied under a Bayesian approach.
- In this case, the vector of expected rewards is assumed to come from a known distribution (called *Bayesian prior*).
- One is typically interested in provable guarantees in expectation over this distribution.
- E.g., in music recommendation, the popularity of songs may be assumed to follow a Zipf distribution. Then, the vector of expected rewards could be assumed to be drawn from a distribution over different Zipf distributions.

Taxonomy of MAB models

Rewards may have a known structure

Structured Rewards

- Rewards may have a known structure.
- E.g., actions correspond to points in $\mathbb{R}^d$, and in each round the reward is a linear (or concave or Lipschitz) function of the chosen action.

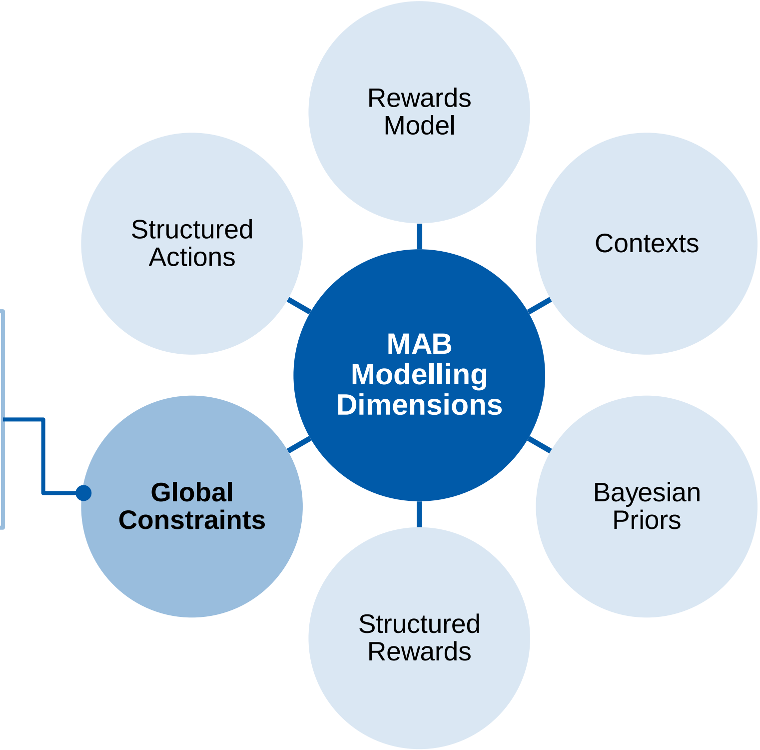


Taxonomy of MAB models

Global constraints may bind across actions and across rounds

Global Constraints

- The agent may be subject to global constraints that bind across actions and across rounds.
- E.g., a moving robot may be limited across actions and across rounds by its finite battery.

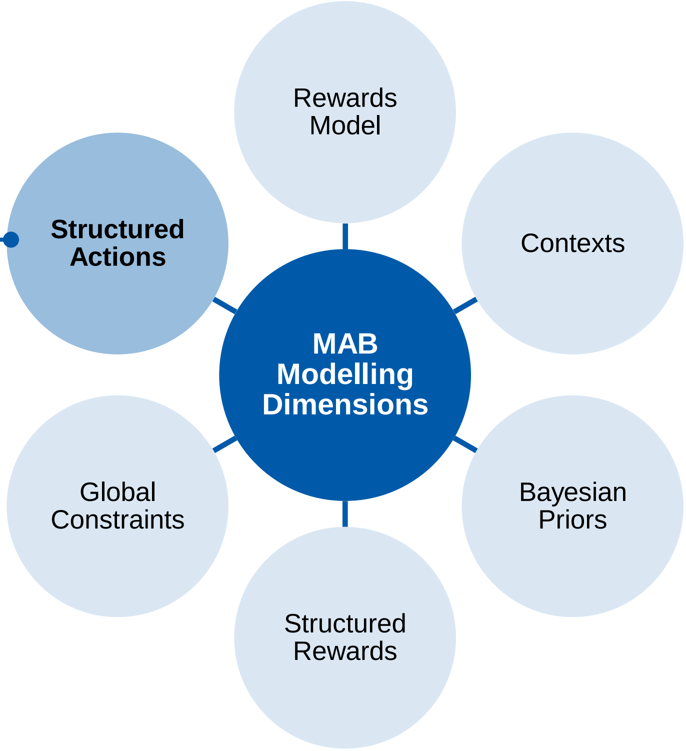


Taxonomy of MAB models

An agent may need to make several decisions at once

Structured Actions

- An agent may need to make several decisions at once and hence select several actions per round.
- E.g., a website may need to pick a set of banner ads to display simultaneously, and a base station may need to choose a set of antenna beams simultaneously.

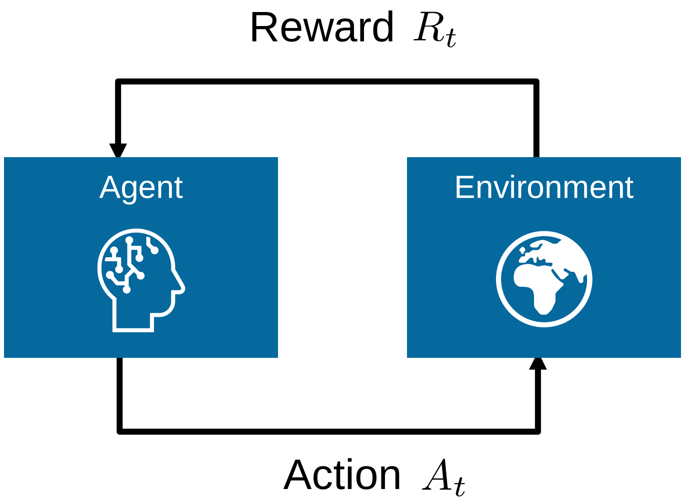


Outline

- Introduction
- Taxonomy of MAB models
- **Stochastic Bandits: Model & Examples**

Stochastic Multi-armed Bandits

We now formally define the basic MAB model with IID rewards



Given:

A set $\mathcal{A}$ of $k := |\mathcal{A}|$ actions and a time horizon T.

Interaction:

- At each step t, the agent selects an action $A_t \in \mathcal{A}$.
- The environment generates a reward R_t by drawing independently from $\mathcal{R}^{A_t}$.

Goal:

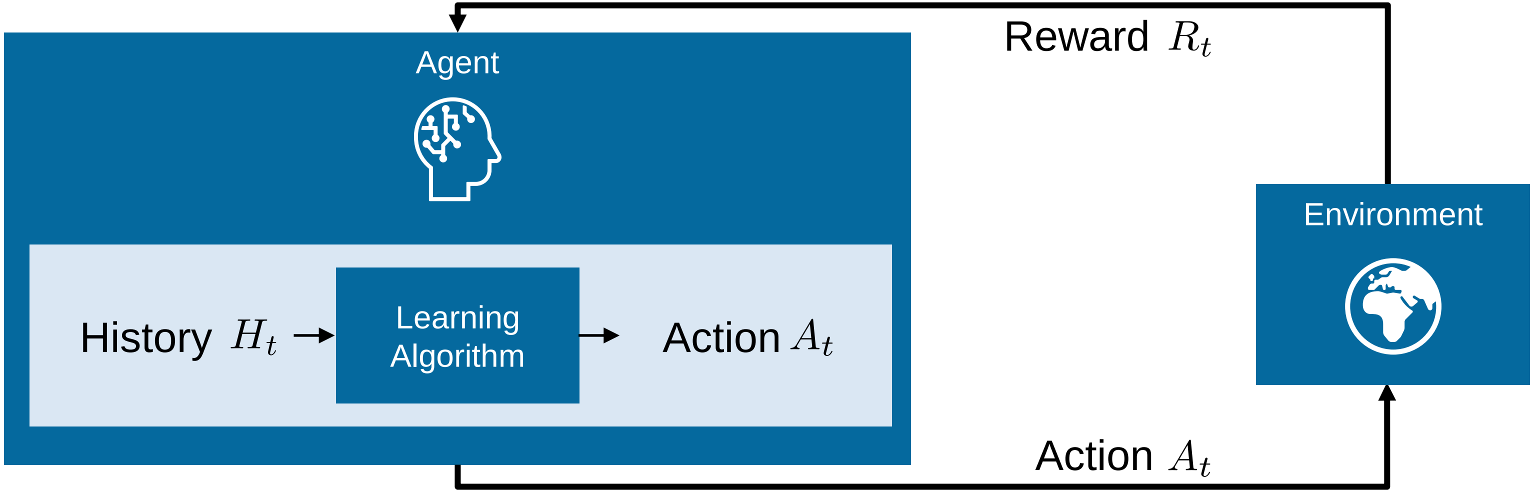
Maximize cumulative reward $\sum_{t=1}^{T} R_t$.

Unknown: r é um valor que Rt pode assumir

- $\mathcal{R}^a(r) = \mathbb{P}[r|a]$ Stationary probability distribution over bounded rewards for action $a \in \mathcal{A}$
- $Q(a) = \mathbb{E}[r|a]$ Expected reward ("action value")

Learning Algorithm

A learning algorithm tries to learn the unknown action values



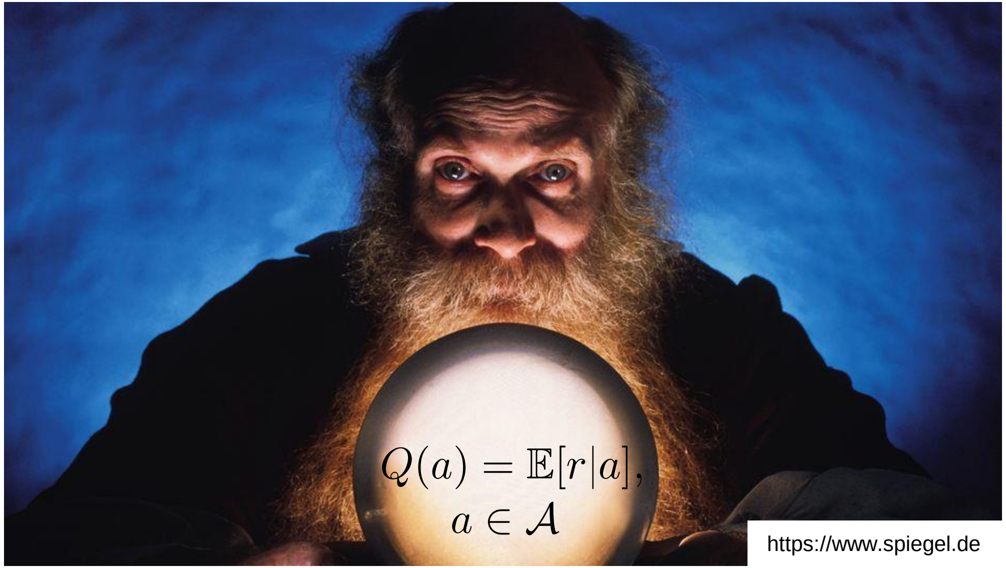
Insight legal: At é uma variavel aleatoria, pq é determinado a partir do historico, que é determinado a partir de Rt, que é probabilistico

- Not knowing action values, the agent tries to learn the unknown action values $Q(a)$.
- A **learning algorithm** maps history $H_t = A_1, R_1, ..., A_{t-1}, R_{t-1}$ to next action A_t.
- A learning algorithm yields an expected reward of $\mathbb{E}\left[\sum_{t=1}^{T} R_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} Q(A_t)\right]$.

eu tenho impressão que isso ta errado

Oracle

An oracle selects optimal actions based on prior knowledge about action values

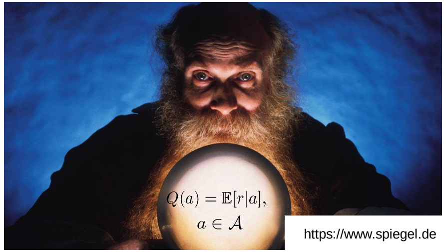


https://www.spiegel.de

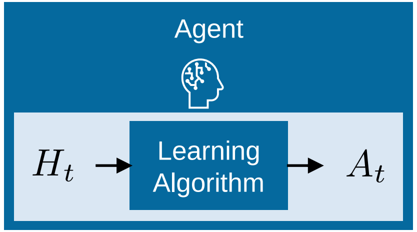
- Knowing action values $Q(a)$, Oracle selects in each step optimal action $a^* = \text{argmax}_{a \in \mathcal{A}} Q(a)$.
- Oracle yields an expected cumulative reward of $\mathbb{E}\left[\sum_{t=1}^{T} R_t\right] = T \cdot Q(a^*)$.

Regret

Regret measures loss of learning compared to Oracle‘s optimal selection



VS.



$$\mathbb{E}\left[\sum_{t=1}^{T} R_t\right] = T \cdot Q(a^*)$$

$$\mathbb{E}\left[\sum_{t=1}^{T} R_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} Q(A_t)\right]$$

- **Regret:** Loss of learning for one time step $l_t = Q(a^*) - \mathbb{E}\left[Q(A_t)\right]$
- **Total regret:** Total loss of learning $L_T = T \cdot Q(a^*) - \mathbb{E}\left[\sum_{t=1}^{T} Q(A_t)\right]$

Maximize cumulative reward = Minimize total regret

Counting Regret

We can express regret in terms of counts for large gaps

Regret is a function of gaps and counts:

- **Count** $N_{T+1}(a)$: Number of times action a is selected by the learning algorithm up to T.
- **Gap** Δ_a: Difference in value between action a and optimal action a^*, i.e.

$$\Delta_a = Q(a^*) - Q(a).$$

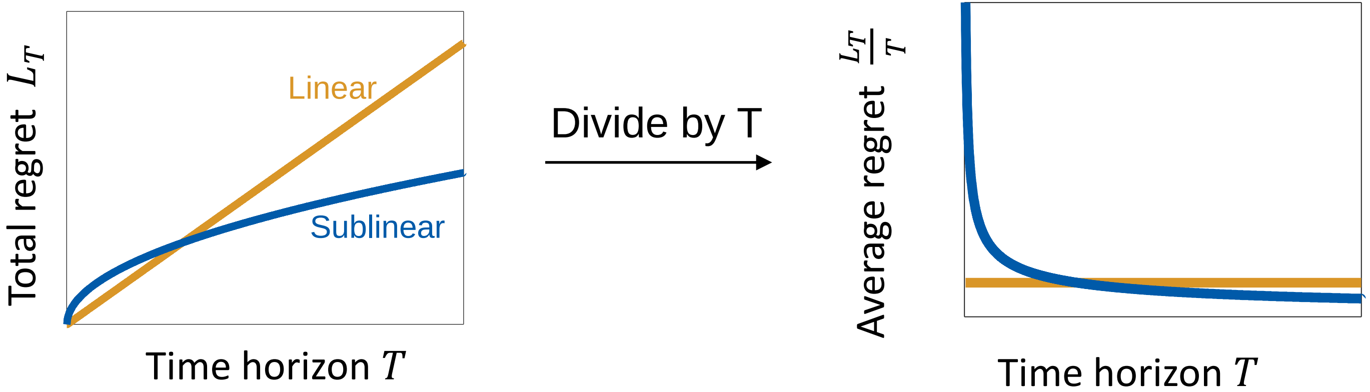
$$
\begin{aligned}
L_T &= T \cdot Q(a^*) - \mathbb{E}\left[\textstyle\sum_{t=1}^{T} Q(A_t)\right] \\
&= \mathbb{E}\left[\textstyle\sum_{t=1}^{T}(Q(a^*) - Q(A_t))\right] && \longrightarrow \text{Used def. of expectation} \\
&= \mathbb{E}\left[\textstyle\sum_{a \in \mathcal{A}} N_{T+1}(a)(Q(a^*) - Q(a))\right] && \longrightarrow \text{Rewrote reward via counts} \\
&= \textstyle\sum_{a \in \mathcal{A}} \mathbb{E}\left[N_{T+1}(a)\right] \Delta_a && \longrightarrow \text{Rewrote reward via gaps}
\end{aligned}
$$

A good learning algorithm ensures small counts for large gaps.

Problem: Gaps are not known!

Linear or Sublinear Regret

Learning algorithms need to achieve sublinear total regret to be able to learn



quanto menor conseguirmos o gama melhor

Sublinear total regret (i.e., $L_T = O(T^\gamma)$ for some $\gamma < 1$) means:

- the average regret per time step converges to zero, i.e, $\lim_{T\to\infty} \frac{L_T}{T} = 0.$
- the learning algorithm‘s action selections converge to the optimal action selection.

Lower Bound on Regret

Regret of any learning algorithm for stochastic MAB is at least logarithmic

TECHNISCHE UNIVERSITÄT DARMSTADT

- The performance of any algorithm is determined by similarity between optimal actions and other actions.
- Hard problems have similar-looking actions with different means. variancia mto grande - valores de recomepensa que se intersectam
- This is described formally by the gap Δ_a and the similarity in distributions $\mathrm{KL}(\mathcal{R}^a || \mathcal{R}^{a^*})$.

Theorem (Lai and Robbins)

Asymptotic total regret is at least logarithmic in the time horizon:

$$\lim_{T\to\infty} \frac{L_T}{\log T} \geq \sum_{a|\Delta_a>0} \frac{\Delta_a}{\mathrm{KL}(\mathcal{R}^a || \mathcal{R}^{a^*})}.$$

Question

Do you think it is possible to achieve sublinear regret?

Is it possible to find an algorithm giving us sublinear regret on any stochastic MAB?

$\longrightarrow$ Next lecture

Learning Goals

TECHNISCHE UNIVERSITÄT DARMSTADT

- You can explain the differences between full Reinforcement Learning and Multi-armed Bandits.
  - → MABs is a simplified version of RL, where actions only affect immediate rewards.
- You can name and explain the main modeling dimensions of Multi-Armed bandit models.
  - → MABs cover a large problem space and can be distinguished according to several modelling dimensions as seen in today‘s lecture.
- You can model decision-making problems using the stochastic Multi-armed bandit model.
  - → See examples of today‘s lecture and exercise 2.

Lecture Overview

Next week, we‘ll study algorithms for stochastic MABs