# Exploration in RL

## Motivation

Marius Lindauer

Leibniz
Universität
Hannover

tnt

Forschungszentrum · Research Center
L3S

Winter Term 2021

# Why do we need exploration?

- Avoid getting trapped in local optima
  - If we have no convergence guarantees
- In sparse reward scenarios:
  - rare observations of rewards
  - following the Q-function or gradients might be very slow, or can lead to plateaus
- Faster convergence by discovering shortcuts

# Why do we need exploration?

- Avoid getting trapped in local optima
    - If we have no convergence guarantees
- In sparse reward scenarios:
    - rare observations of rewards
    - following the Q-function or gradients might be very slow, or can lead to plateaus
- Faster convergence by discovering shortcuts

- Risk: too much exploration could be a waste of resources

  ⤳ Exploration-exploitation dilemma

# The Bandit Problem

- Simplified RL setting with no states
- Simply try to identify which action $a^* \in \mathcal{A}$ is the best one
  - of course, we want to be efficient in doing that!
  - Practical application examples:
    clinical trials or financial portfolio design
- Reward is drawn from some unknown distribution

# The Bandit Problem

- Simplified RL setting with no states
- Simply try to identify which action $a^* \in \mathcal{A}$ is the best one
    - of course, we want to be efficient in doing that!
    - Practical application examples:
      clinical trials or financial portfolio design
- Reward is drawn from some unknown distribution

$\rightsquigarrow$ That's exactly the problem you face in every state $s$ again.
Let's assume that we fix $s$ for the moment

# The Bandit Problem (cont'd)

- Assume that $V^*$ is the expected reward from playing the best action $a^*$
- Total regret

$$\rho_T = T \cdot V^* - \sum_{t=1}^{T} r_t$$

  - where $r_t$ is the reward, we obtained at time point $t$

- Goal is to achieve zero regret in the limit:

$$\lim_{T \to \infty} \rho_T / T = 0$$

$\rightsquigarrow$ There is no offline training phase;
but we have to learn to identify $a^*$ on the fly!

# Exploration vs. Exploitation

- Exploitation: Play the action $\hat{a}$ you believe is the best based on your previous experience
- Exploration: Play an action to improve your knowledge, e.g., wrt the reward distribution of one action or the entropy of being the best

# Exploration vs. Exploitation

- Exploitation: Play the action $\hat{a}$ you believe is the best based on your previous experience
- Exploration: Play an action to improve your knowledge, e.g., wrt the reward distribution of one action or the entropy of being the best
- Do enough exploitation to ensure that we achieve zero regret
- Do enough exploration to ensure that we really identified $a^* = \hat{a}$