# **Exploration in RL**

**Traditional Exploration Strategies for Bandits** 

#### Marius Lindauer







Winter Term 2021

### Recap: The Bandit Problem

- ► Simplified RL setting with no states
- lacktriangle Simply try to identify which action  $a^* \in \mathcal{A}$  is the best one
  - of course, we want to be efficient in doing that!
- ▶ Reward is drawn from some unknown distribution

### Recap: The Bandit Problem

- Simplified RL setting with no states
- lacktriangle Simply try to identify which action  $a^* \in \mathcal{A}$  is the best one
  - of course, we want to be efficient in doing that!
- ▶ Reward is drawn from some unknown distribution
- $\sim$  That's exactly the problem you face in every state s again. Let's assume that we fix s for the moment

Lindauer RL: Exploration, Winter Term 2021 2

## **Optimistic Initialization**

- lacktriangle Simple idea: initialize  $\hat{Q}_0(a)$  to high values
- ▶ Update action value by incremental Monte Carlo evaluations
- ightharpoonup Starting with N(a)>0

$$\hat{Q}_t(a_t) = \hat{Q}_{t-1} + \frac{1}{N_t(a_t)}(r_t - \hat{Q}_{t-1})$$

- Encourages systematic exploration early on
- ▶ But can still lock onto suboptimal action

- $\blacktriangleright$  play best known action  $\hat{a}$  with probability  $1-\epsilon$
- $lackbox{}$  play a random action  $a\in\mathcal{A}$  with probability  $\epsilon$
- Question: Is this a zero-regret strategy?

Lindauer RL: Exploration, Winter Term 2021

- lacktriangle play best known action  $\hat{a}$  with probability  $1-\epsilon$
- $lackbox{}$  play a random action  $a\in\mathcal{A}$  with probability  $\epsilon$
- Question: Is this a zero-regret strategy?
  - No, since with probability  $\epsilon$  we will obtain a non-zero regret and therefore, even in the limit the regret will not go to zero, but obtain a linear regret.

Lindauer RL: Exploration, Winter Term 2021 4

- $\blacktriangleright$  play best known action  $\hat{a}$  with probability  $1-\epsilon$
- lacktriangle play a random action  $a \in \mathcal{A}$  with probability  $\epsilon$
- Question: Is this a zero-regret strategy?
  - No, since with probability  $\epsilon$  we will obtain a non-zero regret and therefore, even in the limit the regret will not go to zero, but obtain a linear regret.
- $\triangleright$  Solution: Anneal  $\epsilon$  over time
- ▶ If we linearly anneal  $\epsilon$  over time to 0, do we have a zero-regret strategy?

Lindauer RL: Exploration, Winter Term 2021

- $lackbox{ play best known action } \hat{a} \ \mbox{with probability } 1-\epsilon$
- lacktriangle play a random action  $a \in \mathcal{A}$  with probability  $\epsilon$
- Question: Is this a zero-regret strategy?
  - No, since with probability  $\epsilon$  we will obtain a non-zero regret and therefore, even in the limit the regret will not go to zero, but obtain a linear regret.
- $\triangleright$  Solution: Anneal  $\epsilon$  over time
- ▶ If we linearly anneal  $\epsilon$  over time to 0, do we have a zero-regret strategy?
  - ▶ No, because of the linear annealing, we have only a finite amount of observations which might not suffice to identify the best action
- $\sim$  Anneal  $\epsilon$  proportional  $\sqrt{t}$  or 1/t

Lindauer RL: Exploration, Winter Term 2021 4

## **Upper Confidence Bounds**

- lacktriangle Track all rewards you obtained by playing each action  $a_k$  and compute mean  $\mu(a_k)$  and standard deviation  $\sigma(a_k)$  to estimate the underlying reward distribution
- Optimistic in face of uncertainty by upper confidence bound:

$$\mu(a_k) + \kappa \cdot \sigma(a_k) / \sqrt{N(a)}$$

▶ Idea: Over time, we get more and more evidence for the best actions until we are sure that the best known is really the best

## **Upper Confidence Bounds**

- lacktriangle Track all rewards you obtained by playing each action  $a_k$  and compute mean  $\mu(a_k)$  and standard deviation  $\sigma(a_k)$  to estimate the underlying reward distribution
- Optimistic in face of uncertainty by upper confidence bound:

$$\mu(a_k) + \kappa \cdot \sigma(a_k) / \sqrt{N(a)}$$

- ▶ Idea: Over time, we get more and more evidence for the best actions until we are sure that the best known is really the best
- ► To prevent premature convergence: Use optimistic initialization of each action s.t. all actions are played in the beginning

#### UCB 1

$$a_t \in \argmax_{a \in A} Q(a) + \sqrt{\frac{2 \log t}{N_t(a)}}$$

ightharpoonup Condition: Rewards have to be i.i.d random variables in [0,1].

#### UCB 1

$$a_t \in \operatorname*{arg\,max}_{a \in A} Q(a) + \sqrt{\frac{2 \log t}{N_t(a)}}$$

- lacktriangle Condition: Rewards have to be i.i.d random variables in [0,1].
- ▶ Theorem: The UCB 1 algorithm achieves logarithmic asymptotic total regret

$$\lim_{t\to\infty} L_t \leq 8\log t \sum_{a|\Delta_a>0} \Delta_a$$

- $\blacktriangleright$  where  $L_t$  is the regret after t trials and  $\Delta_a = V^* Q(a)$
- Using Hoeffding's Inequality

# **Thompson Sampling**

- lacktriangle Track all rewards you obtained by playing each action  $a_k$  and compute mean  $\mu(a_k)$  and standard deviation  $\sigma(a_k)$  to estimate the underlying distribution
- Draw from each estimated distribution a single realization and simply play the action with the best one

$$s_k \sim \mathcal{N}(\mu(a_k), \sigma(a_k))$$
 
$$a \in \mathop{\arg\max}_{a_k} s_k$$

In the limit, only the best performing action will be played with high probability