Exploration in RL

Traditional Exploration Strategies for Bandits

Marius Lindauer

Leibniz Universität Hannover

Winter Term 2021

Recap: 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!
- Reward is drawn from some unknown distribution

Recap: 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!
- 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

Optimistic Initialization

- Simple idea: initialize $\hat{Q}_0(a)$ to high values
- Update action value by incremental Monte Carlo evaluations
- 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

ϵ-greedy

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

ϵ-greedy

- play best known action $\hat{a}$ with probability $1 - \epsilon$
- play a random action $a \in \mathcal{A}$ with probability ϵ
- Question: Is this a zero-regret strategy?
  - No, since with probability ϵ 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.

ϵ-greedy

- play best known action $\hat{a}$ with probability $1 - \epsilon$
- play a random action $a \in \mathcal{A}$ with probability ϵ
- Question: Is this a zero-regret strategy?
  - No, since with probability ϵ 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.
- Solution: Anneal ϵ over time
- If we linearly anneal ϵ over time to 0, do we have a zero-regret strategy?

ϵ-greedy

- play best known action $\hat{a}$ with probability $1 - \epsilon$
- play a random action $a \in \mathcal{A}$ with probability ϵ
- Question: Is this a zero-regret strategy?
  - No, since with probability ϵ 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.
- Solution: Anneal ϵ over time
- If we linearly anneal ϵ 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

$\rightsquigarrow$ Anneal ϵ proportional $\sqrt{t}$ or $1/t$

Upper Confidence Bounds

- 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

- 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

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

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

UCB 1

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

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

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

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

Thompson Sampling

- 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

$$s_k \sim \mathcal{N}(\mu(a_k), \sigma(a_k))$$

$$a \in \arg\max_{a_k} s_k$$

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