Prof. Dr. Leif Döring

André Ferdinand, Sara Klein

Reinforcement Learning

3. Excercise Sheet

1. Best Baseline

The variance of a random vector X is defined by to be $\mathbb{V}[X] := \mathbb{E}[||X||_2^2] - ||E[X]||_2^2$. Show by differentiation that

$$b_* = \frac{\mathbb{E}_{\pi_\theta}[X_A ||\nabla \log \pi_\theta(A)||_2^2]}{\mathbb{E}_{\pi_\theta}[||\nabla \log \pi_\theta(A)||_2^2]}$$

is the baseline that minimises the variance of the unbiased estimators

$$(X_A - b)\nabla \log(\pi_\theta(A)), \quad A \sim \pi_\theta,$$

of $\nabla J(\theta)$.

2. Programming task, Gradient Bandit Methods

Implement the Gradient Bandit algorithm from example 1.2.15 of the lecture. The probability that an arm is drawn is given by the soft-max distribution

$$\mathbb{P}_\pi(A_t = a) = \frac{\exp(\theta_t(a))}{\sum_{b=1}^{k} \exp(\theta_t(b))} =: \pi_t(a), a \in \mathcal{A}.$$

The weights are updated as follows

$$\theta_{t+1}(a) = \begin{cases} \theta_t(a) + \alpha(x_t - \overline{x}_t)(1 - \pi_t(a)) & , a = A_t \\ \theta_t(a) - \alpha(x_t - \overline{x}_t)\pi_t(a) & , \text{otherwise} \end{cases},$$

where $\alpha > 0$ is a step-size parameter and $\overline{x}_t \coloneqq \frac{1}{t-1} \sum_{i=1}^{t-1} x_t$ is the mean reward until time $t-1$. Implement the Gradient Bandit algorithm with and without the baseline term $\overline{x}_t, t \in \{1, \ldots, n\}$ and test both algorithm on a Gaussian Bandit. Play around with the mean parameters of the Gaussian Bandit. What do you notice?

3. Programming task, Boltzmann Exploration

In this task we want to implement different variants of the Boltzmann Exploration Algorithm.

(a) Use the implementation from Algorithm 1.

(b) Use the Gumbel trick from Lemma 1.2.11 to implement the algorithm.

(c) Use the implementation of Boltzmann exploration from algorithm 2. Source: paper section 4.

Input : Initialization $\hat{Q}_a(0), a \in \mathcal{A}$, number of total timesteps $n \in \mathbb{N}$, number of arms $k \in \mathbb{N}$ and parameter $\theta > 0$

Output: Trajectory of Rewards $(x_t)_{t\in\{1,\ldots,n\}}$ and actions $(a_t)_{t\in\{1,\ldots,n\}}$

begin

  for $t \leftarrow 1$ **to** n **do**

    Sample a_t from $\mathrm{SM}\big(\theta, (\hat{Q}_a(t))_{a\in\mathcal{A}}\big)$;

    Obtain reward x_t by playing arm a_t;

    Update the estimated action value functions $\big(\hat{Q}_a(t), a \in \mathcal{A}\big)$;

  end

  return $(x_t)_{t\in\{1,\ldots,n\}}, (x_t)_{t\in\{1,\ldots,n\}}$;

end

Algorithm 1: Boltzmann exploration algorithm

Input : Initialization $\hat{Q}_a(0), a \in \mathcal{A}$, number of total timesteps n, number of arms $k \in \mathbb{N}$ and parameter $C \in \mathbb{R}$

Output: Trajectory of Rewards $(X_t)_{t\in\{1,\ldots,n\}}$ and actions $(A_t)_{t\in\{1,\ldots,n\}}$

begin

  for $t \leftarrow 1$ **to** n **do**

    Simulate $z_a, a \in \mathcal{A}$ independently identically standard Gumbel;

    Set $a_t = \underset{a\in\mathcal{A}}{\arg\max}\big\{\hat{Q}_a(t) + \sqrt{\frac{C^2}{N_a}} z_a\big\}$;

    Obtain reward x_t by playing arm a_t;

    Set $N_{a_t} = N_{a_t} + 1$;

    Update the estimated action value functions $\big(\hat{Q}_a(t)\big)$;

  end

  return $(X_t)_{t\in\{1,\ldots,n\}}, (A_t)_{t\in\{1,\ldots,n\}}$;

end

Algorithm 2: Boltzmann exploration algorithm modified

(d) Test the different algorithms on a multi-armed bandit. Which algorithm with which parameter configuration leads to the best results (minimum reward, maximum best action probability)?