FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
Benedikt Wille

Reinforcement Learning

2. Exercise Sheet

25.02.2025

For this assignment, knowledge from lectures 1 to 6 is assumed.

1. Upper bound on $\hat{Q}_a(t)$ for many samples

Suppose ν is a bandit model with 1-sub-gaussian arms. Show that under the UCB Algorithm $\hat{Q}_a(t) < Q_a + \Delta_a$ with probability $1-\delta$, given that $T_a(t) > \frac{2\log(1/\delta)}{\Delta_a^2}$.

2. Regret bounds for UCB on σ-subgaussian bandit models

For σ-subgaussian bandit models the UCB exploration bonus is modified as

$$\mathrm{UCB}_a(t) := \begin{cases} \infty, & T_a(t) = 0, \\ \hat{Q}_a(t) + \sqrt{\dfrac{2\sigma^2 \log(\frac{1}{\delta})}{T_a(t)}}, & T_a(t) \neq 0. \end{cases}$$

Check that the regret bound in Theorem 1.3.8 using $\delta = \frac{1}{n^2}$ changes to

$$R_n(\pi) \leq 3\sum_{a\in\mathcal{A}} \Delta_a + 16\sigma^2 \log(n) \sum_{a: Q_a \neq Q_*} \frac{1}{\Delta_a},$$

and that this leads to

$$R_n(\pi) \leq 8\sigma\sqrt{Kn\log(n)} + 3\sum_{a\in\mathcal{A}} \Delta_A$$

in Theorem 1.3.9.

3. 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 $\nabla J(\theta)$.

4. *Programming task: Algorithms, algorithms, algorithms

The aim of this task is to implement the remaining bandit algorithms from the lecture. To this end, implement the following algorithms and versions of the algorithms:

a) Greedy: Implement the purely greedy algorithm, the ϵ-greedy algorithm with fixed ϵ, and the ϵ-greedy algorithm with rates ϵ_t decreasing in time.

b) UCB: Implement the UCB algorithm as presented in the lecture, and the version adapted to σ-subgaussian bandits.

c) Boltzmann exploration: Implement the simple Boltzmann exploration, the version with the Gumbel trick, and a version allowing arbitrary distributions (at least Cauchy, Beta, Betaprime, Chi, see e.g. scipy.stats) instead of Gumbel. Additionally, implement a version where

$$A_t \sim \arg\max_{a \in \mathcal{A}} \left\{ \hat{Q}_a(t-1) + \sqrt{\frac{C}{T_a(t-1)}} Z_a \right\},$$

where Z_a are independently identically standard Gumbel and $C \in \mathbb{R}$ is a parameter.

d) Policy gradient: Implement the policy gradient method with and without the baseline trick. Use the softmax-distributions as family of probability distributions to optimize over. Its probability weights are defined as

$$\pi_\theta(a) := \frac{\exp(\theta_a)}{\sum_{b=1}^{K} \exp(\theta_b)}, \quad a \in \mathcal{A},\, \theta \in \mathbb{R}^{\mathcal{A}}.$$

Hint: Structurally, the implementation for the algorithms should not differ too much from the ETC algorithm, which you already implemented last week, if you followed the given hints. Essentially you need to only think about how to initialize certain objects you need and how to perform one given step of the algorithm at a time.

5. *Programming task: Optimal parameters

Simulate a 5-armed Bernoulli bandit with random means and play the bandit $n = 10000$ times with each of the algorithms from exercise 4. Whenever the lecture specifies them, use the optimal parameters (even if they are model-dependent). In all other cases think about compute-efficient ways to numerically determine or search for the best parameters. Average your results over $N = 1000$ iterations of the experiment and plot the following:

a) the regrets over time, including a shaded area around the curve indicating the 95-percent confidence intervals, and

b) boxplots of the following data at the end of the time-horizon n:

   i) the real means of the arms versus each algorithm's estimates,

   ii) the probabilities for playing each of the arms, and

   iii) the regrets.

Hint: You can use the structure from Exercise Sheet 1 and implement new plot functions.

The solution to the theoretical exercises will be discussed in the exercise class in B5 on March 04, 2025, at Mathelounge in B6 B301.