Prof. Joschka Bödecker, Gabriel Kalweit

REINFORCEMENT LEARNING
Exercise 8

## 0   Lecture

Watch *Lecture 10: Integrating Learning and Planning*[1] before the upcoming session on Friday, January 18.

## 1   UCB1

(a) Solve this task using pen and paper. Imagine a multi-armed bandit setting with three arms. Each arm has a bias $-+\frac{1}{2}$, $+\frac{1}{4}$ and $+\frac{1}{8}$, respectively. The Q-values for each arm are initialized by $1 - b$, where $b$ is the corresponding bias. The return for a pull is then $\mathrm{clip}(b + u, 0, 1)$, where $u$ is uniformly sampled noise from $[0, 1]$. Provide the first 10 iterations of the UCB1-algorithm based on the sampled noise array $[0.4, 0.7, 0.2, 0.3, 0.8, 0.5, 0.6, 0.1, 0.0, 0.9]$ for the returns. Estimate the Q-values by the mean.

(b) Implement the different exploration strategies – UCB1, $\epsilon$-greedy, random exploration and decaying $\epsilon$-greedy exploration for some schedule of your choice – in `exploration.py`. Explain your schedule. Evaluate the exploration strategies on the given toy problem by plotting the regret of UCB1, decaying $\epsilon$-greedy, $\epsilon$-greedy and random exploration and compare the results.

## 2   Exploration and Real Physical Systems

Imagine you want to apply the algorithms from this lecture on a real physical system and some actions in some states may break your robot, so you have avoid them (but you do not know those states beforehand). However, the presented algorithms **need** to explore in order to find a good solution. Which exploration strategies from the lecture lead to problems and why? How would you approach exploration?

## 3   Experiences

Make a post in thread *Week 09: Exploration and Exploitation* in the forum[2], where you provide a brief summary of your experience with this exercise, the corresponding lecture and the last meeting.

---

[1] https://ilias.uni-freiburg.de/goto.php?target=xvid_1121355&client_id=unifreiburg
[2] https://ilias.uni-freiburg.de/goto.php?target=frm_1121060&client_id=unifreiburg