# Multi-Armed Bandit Assignments

*Author: Diederik M. Roijers*

*These assignment consists of must-haves and optional assignments. Perfectly doing the must-haves leads to a 12/20 score. The remaining 8 points can be attained by doing optional assignments. Each optional assignment has a maximum number of additional points. NB: even though your number of points can exceed 20, the maximum possible score is capped at 20. However, note that it is possible to lose points by making mistakes in the assignments, therefore, when aiming for a 20, it is advised to aim for a total number of attainable points higher than 20.*

## *Assignments*

This assignment is about multi-armed bandits in which the payoffs for each arm are Bernoulli-distributed. An explanation of this problem and some of the methods can be found in the following paper:

> Chapelle, Olivier, and Lihong Li. "An empirical evaluation of Thompson sampling." In *NIPS: Advances in neural information processing systems*, pp. 2249-2257. 2011. http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf

Please read sections 1, 2, and 3.

- **Must-have:** use the code from https://dataorigami.net/blogs/napkin-folding/79031811-multi-armed-bandits to obtain a version of the Bernoulli-distributed MAB problem, or implement it yourself.
- **Must-have:** implement the epsilon-greedy, soft-max, and optimistic initialisation action-selection strategies, and compare their cumulative reward curves (without discounting) and cumulative regret curves, for max. 2500 arm pulls, as a function of the number of pulls.
- **Optional assignments:** implement the following methods and compare them in terms of cumulative undiscounted reward and cumulative regret. The number of points for each method are listed:
  - From Auer et al 2002 (https://homes.di.unimi.it/~cesabian/Pubblicazioni/ml-02.pdf):
    - UCB1 (2 points)
    - UCB2 (2 point)
    - $\varepsilon_n$-greedy (2 points)
  - From Olivier and Li, 2011
    - Thompson sampling (3 points)
    - UCB using the Chernoff bound (Equation 1) (2 points)
  - Other exploration methods you can find in the literature (2 points each, max 6 points (i.e., 3 methods) in total)
- **Optional assignment (8 points)**: Implement Bayes-optimal exploration. Until what number of pulls, $T_{max}$, is it possible to run this within 20 minutes computation time? How much reward on average (over 100 random runs; NB: you only need to construct the search tree once) does Bayes-optimal exploration accrue for $T_{max}$? How much reward can the other method accrue on average for $T_{max}$?