# Exam Dynamic Programming & Reinforcement Learning
# December 2021

This exam consists of **4** problems, each consisting of several questions.
All answers should be motivated, including calculations, formulas used, etc.
The minimal grade is 1. All questions give 0.5 points when answered correctly.
You are only allowed to use pen and paper.

1. Consider the deterministic shortest path problem.
a. Give the formula of the dynamic programming recursion. Pay attention to the starting conditions.
b. Give in words the definition of the value function for this specific problem.
Consider an instance with states $\{1, ..., 5\}$, with state 5 the destination, and lengths $d(x, y) = |x - y| - 1$ for $x, y \in \{1, ..., 5\}$, $x \neq y$.
c. Make a drawing of the graph with its distances.
d. Solve it using dynamic programming. Show in a table all intermediate results, i.e., all $V_t(x)$ for all $t$ and $x$.

2. Consider a Markov reward chain with states $\{1, 2, 3, 4\}$, transition probabilities $p(1, 2) = p(1, 4) = p(2, 1) = p(2, 3) = p(3, 2) = p(3, 4) = p(4, 1) = p(4, 3) = 0.5$, and reward $r(x) = x$.
a. Give the distribution $\pi_4$ for $\pi_0 = (1, 0, 0, 0)$.
b. Is this chain communicating and/or periodic? Motivate your answer.
c. Derive the long-run average distribution, using the answers to a) and b).
d. Give the set of equations (the balance equations) of which the stationary distribution is the unique solution and give this solution.
e. Formulate the average reward Poisson equation and give a solution.
f. Explain why value iteration does not converge and how this can be solved.

3. Consider a multi-armed bandit problem, where each arm represents flipping a coin that can be a misprint: it has either heads and tails on the sides (H/T) or twice heads (H/H). The reward for heads is 1, tails 0, and we are interested in maximizing the total expected discounted revenue.
a. Describe how the greedy policy, the $\varepsilon$-greedy policy, and the greedy policy with optimistic values would select arms in this situation. Give also the recursion on which they are based.
b. Assume a prior distribution for each arm that gives probability 0.5 to both H/T and H/H. Calculate for 3 consecutive pulls of the same arm all possible posterior distributions.
c. Explain in words what a Gittins index is and a stopping time.
d. Use your understanding of this problem to formulate what you think is the optimal policy.

4a. Describe the difference between a model-based approach and a model-free approach (for the model-free approach, cite the two subfamilies of approaches).
b. Similarly to supervised learning, the suboptimality (on the expected return) of an RL policy learned based on limited data can be decomposed into two terms. Cite these two terms and explain what they mean.
c. What are the two conditions for tabular Q-learning to converge (in the online setting)?
d. What is the specificity of the REINFORCE algorithm within the family of policy based methods?