# Resit Dynamic Programming & Reinforcement Learning
# February 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 a Markov reward chain with $\mathcal{X} = \{1, 2, 3, 4\}$,

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix} \text{ and } r = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

a. Is this chain periodic and/or communicating?
b. Compute the stationary distribution and the long-run average reward.
c. Give the average-reward Poisson equation.
d. Find a solution to this average-reward Poisson equation.
e. Only in state 1 there is a second action with $p(4|1, 2) = 1$ and $r(1, 2) = r^*$ (thus in state 1 you go with probability 1 to state 4 under action 2 and the reward is $r^*$). For which values of $r^*$ is this the optimal action?
f. Formulate the LP problem of the resulting Markov decision chain.

2. Consider a discounted Markov reward chain with $\mathcal{X} = \{1, 2\}$, $\beta = 0.5$,

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix} \text{ and } r = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

a. Formulate the discounted Poisson equation.
b. Solve this discounted Poisson equation.
c. Find the Gittins indices for both starting states.
d. Suppose you have a bandit problem with a number of arms as described in this exercise with different starting states. Next to that you haver one arm that always gives 1.1 as reward. What is the optimal policy?

3a. Give the state-transition diagram of the $M|M|1$ queue.
b. Formulate the average-reward continuous-time Poisson equation with as direct reward the number of customers in the queue.
c. Show that a solution is given by $x(x+1)/(2(\mu - \lambda))$.
d. Show how you can use this to find a good heuristic in case you have to choose which queue to join of multiple parallel queues with different service speeds.

4a. What is the principle of "optimism under uncertainty"?
b. Cite three elements on which it is possible to improve the generalization capabilities of RL algorithms.
c. When estimating the expected return of following a given policy, Monte Carlo rollouts can be used. Cite one advantage and one drawback of using Monte Carlo rollouts.
d. In Q-learning, what are the two potential characteristics of the state space that make is it necessary to use function approximators?