# Exam Dynamic Programming & Reinforcement Learning
# December 2020

This exam consists of **5** problems, each consisting of several questions.
All answers should be motivated, including calculations, formulas used, etc.
The minimal grade is 1. All questions (1a, 1b, etc.) give 0.5 points when answered correctly, except for 1b and 2a which can give 1 point.
You are only allowed to use pen and paper.

1. Consider a Markov reward chain with $\mathcal{X} = \{1, 2, 3\}$,

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

a. Compute the stationary distribution and the long-run average reward.
b. Give all solutions of the average reward Poisson equation.
c. Find the solution that also satisfies $< V, \pi_* >= 0$. What is the interpretation?
d. Explain why value iteration does not converge for the problem of this exercise and explain how this can be solved.

2. Consider a knapsack problem where you have to satisfy weight and size restrictions: for item $i \in \{1, ..., n\}$ weight $w_i$, size $s_i$ and reward $r_i$ are given. The avalaible weight and size are $W$ and $S$. Which items do you take such that the sum of rewards over the items you took is maximized while staying within the weight and size limits?
a. Formulate this as a MDC and give the backward recurrence (without solving it).

3. Consider a 2-state Markov reward process with $p(1, 2) = p(2, 1) = 1$, $T(1) \sim \exp(1)$, $T(2) = 2$ (thus deterministic), and rate rewards $r^r(1) = 0$ and $r^r(2) = 1$.
a. Determine $\pi_*$, $\nu_*$, and $\phi_*$.
b. Formulate the average-reward continuous-time Poisson equation.
c. Give a solution.

4a. Formulate the discounted Poisson equation for the MDC of problem 1.
b. Give a solution for $\beta = 0.5$. Is it unique?
c. For the same problem, calculate the Gittins Index in state 1 and $\beta = 0.5$.
d. Interpret your answer.

5a. Explain why a good exploration/exploitation tradeoff is needed in practice when using planning.
b. Cite two techniques that can be used to reduce the depth of search.
c. What is the reality gap?
d. Cite two potential advantages of learning a distribution of the discounted sum of future rewards instead of $Q^*(x, a)$.