# COMP-579: Reinforcement Learning - Assignment 2

## Posted Monday, February 20, 2023
## Due Friday, March 10, 2023

The assignment can be carried out individually or in teams of two.

1. **Tabular RL [100 points]**

   In this problem, you will compare the performance of SARSA and expected SARSA on the Frozen Lake domain from the Gym environment suite:

   `https://gymnasium.farama.org/environments/toy_text/frozen_lake/`

   Use a tabular representation of the state space. Exploration should be softmax (Boltzmann). You will do 10 independent runs. Each run consists of 500 segments, in each segment there are 10 episodes of training, followed by 1 episode in which you simply run the optimal policy so far (i.e. you pick actions greedily based on the current value estimates). Pick 3 settings of the temperature parameter used in the exploration and 3 settings of the learning rate. You need to plot:

   - One u-shaped graph that shows the effect of the parameters on the final training performance, expressed as the return of the agent (averaged over the last 10 training episodes and the 10 runs); note that this will typically end up as an upside-down u.
   - One u-shaped graph that shows the effect of the parameters on the final testing performance, expressed as the return of the agent (during the final testing episode, averaged over the 10 runs)
   - Learning curves (mean and standard deviation computed based on the 10 runs) for what you pick as the best parameter setting for each algorithm

   Write a small report that describes your experiment, your choices of parameters, and the conclusions you draw from the graphs.

2. **Function approximation in RL [100 points]**

   Implement and compare empirically Q-learning and actor-critic with linear function approximation on the cart-pole domain from the Gym environment suite:

   `https://gymnasium.farama.org/environments/classic_control/cart_pole/`

   For this experiment, you should use a function approximator in which you discretize the state variables into 10 bins each; weights start initialized randomly between $-0.001$ and $0.001$. You will need to use the same seed for this initialization for all parameters settings, but will have 10 different seeds (for the different runs). Use 3 settings of the learning rate parameter $\alpha = 1/4, 1/8, 1/16$. Perform 10 independent runs, each of 1000 episodes. Each episode should start at a random initial state. The exploration policy should be $\epsilon$-greedy and you should use 3 values of $\epsilon$ of your choice. Plot for each algorithm 3 graphs, one for each $\epsilon$, containing the average and standard error of the learning curves for each value of $\alpha$ (each graph will have 3 curves). Make sure all graphs are on

the same scale. Based on these graphs, pick what you consider the best parameter choice for both $\epsilon$ and $\alpha$ and show the best learning curve for Q-learning and actor-critic on the same graph. Write a small report that describes your experiment, your choices of parameters, and the conclusions you draw from this experimentation.