# COMP-767: Reinforcement Learning - Assignment 2

## Posted Tuesday, February 19, 2019
## Due Tuesday, March 12, 2019

The assignment can be carried out individually or in teams of two. You have choices on both parts of the assignments.

1. **Prediction and control in RL [50 points]**

   Choose **one** of the following topics.

   (a) In this task, you will compare the performance of SARSA, expected SARSA and Q-learning on the Taxi domain from the Gym environment suite:

   `https://gym.openai.com/envs/Taxi-v2/`

   Use a tabular representation of the state space, and ensure that the starting and end location of the passenger are random. Exploration should be softmax. You will need to run the following protocol. You will do 10 independent runs. Each run consists of 100 segments, in each segment there are 10 episodes of training, followed by 1 episode in which you simply run the optimal policy so far. Pick 3 settings of the temperature parameter 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 (see the book)
   - One u-shaped graph that shows the effect of the parameters on the final testing performance (see the book)
   - Learning curves (mean and standard deviation) 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.

   (b) We discussed in class some work the complexity of exploration in reinforcement learning. The $E^3$ algorithm is one of the first attempts to provide sample complexity results for tabular reinforcement learning algorithms that do control.

   `https://www.cis.upenn.edu/~mkearns/papers/reinforcement.pdf`

   You need to write a short summary (max 3 pages in format of your choice) of the result and the main steps and ideas in the proof presented in this paper. Feel free to add some background if you think it would be necessary to understand their approach. Explain why (or why not) in your opinion this is an algorithm that generalizes to function approximation.

2. **Function approximation [50 points]**

   Choose **one** of the following topics.

(a) Implement and compare empirically the performance of Monte Carlo and TD-learning with eligibility traces and linear function approximation on a simple domain (more details to be provided)

(b) In 2000, Geoff Gordon proved an interesting result that on-policy SARSA with function approximation converges to a region. This result was in contrast with Leemon Baird's earlier counterexample on Q-learning divergence, which is discussed in the book.

`http://www.cs.cmu.edu/~ggordon/sarsa-bound.ps.gz`

You need to write a short summary (max 3 pages in latex format of your choice) of the intuition of the proof, and what is the intuition behind the difference in behavior of these two algorithms. Do you see any possibility to improve Gordon's result? Explain your answer.