CIN6053: Homework (due by April 18th)

This is not a team project. You do not need to handwrite your answers. Unless otherwise instructed, justify your answers fully. Upload your own self-executable codes, your answers (screenshots if you handwrite your answers), and the full pdf file of the selected paper (per each group) to the course homepage (Be concise for your codes).

1. (50 points) **(programming, use Python or Matlab)** Consider the tic-tac-toe game on a 3×3 board. The reward is 1 if one player wins the game, and 0 otherwise. The discount factor γ is 1. Suppose each player follows the random policy (i.e., choose available positions uniformly) π. For this policy, using the first-visit Monte-Carlo policy evaluation, evaluate the action-value function $q_\pi(s_0, a)$ for the first move, where s_0 is the initial state of the game. What is the best first move for the random policy π? Next, use the temporal-difference learning or TD(0) to evaluate the same action-value function. Which one does converge faster than the other?

   Try to use as few states as possible. Note that states are equivalent if their transition probabilities for the next equivalent states are the same and the rewards are the same for all possible actions. Use rotation or mirror image to identify equivalent states.

2. (50 points) **(programming, use Python or Matlab)** Implement value iteration for the gambler's problem (Example 4.3 in the Sutton's book) and solve it for $p = 0.25$ and $p = 0.55$. In programming, you may find it convenient to introduce two dummy states corresponding to termination with the capital of 0 and 100, giving them values of 0 and 1 respectively. Show your results graphically, as in Figure 4.3. Are your results stable as the threshold θ gets close to 0?