COMP-767 **Name:** _____
Winter 2018
Midterm
March 15, 2018
Time Limit:

---

1. (30 points) Problem formulation

   You are designing a recycling robot whose job is to collect empty soda cans around the building. The robot has a sensor to detect when a can is in front of it, and a gripper to pick up the can. It also senses the level of its battery. The robot can navigate, as well as pick up a can and throw a can it is holding in the trash. There is a battery charger in the building, and the robot should not run out of battery.

   (a) Describe this problem as an MDP. What are the states and actions?

   (b) Suppose that you want the robot to collect as many cans as possible, while not running out of battery. Describe what rewards would enable it to achieve this task

   (c) Instead of thinking about the actions described above, one could describe the task of the robot as choosing between larger activities: walk randomly to find cans, wait for someone to drop a can, or go dock with battery charger. Describe the advantages and disadvantages of this problem formulation compared to the one you gave before.

2. (20 points) Returns and values

   (a) Define the discounted return G_t

   (b) Give an expression for G_t in terms of G_{t+1}

   (c) Using the answers above, show how to obtain the Bellman equation for policy

evaluation from the expected discounted return.

(d) Imagine that the rewards are at most 1 everywhere. What is the maximum value that the discounted return can attain ? Why ?

3. (20 points) Control

(a) Draw the backup diagram for 2-step Q-learning

(b) Write the corresponding learning rule for 2-step Q-learning

(c) Draw the backup diagram for 2-step Sarsa

(d) Write the corresponding learning rule for 2-step Sarsa

4. (10 points) Importance Sampling

(a) Given a trajectory $\tau = S_t, A_t, S_{t+1}, S_{t+1}...S_T$ obtained from a policy $\pi(A_t|S_t)$ in an MDP with transition matrix $P(S_{t+1}|S_t, A_t)$. Write the probability $P_\pi(\tau)$ of seeing τ under the policy π.

(b) Let b be a *behavior* policy and π a *target* policy. Write the expression for their corresponding importance sampling ratio. Show that the ratios don't depend on knowing the transition matrix.

5. (20 points) Function approximation

(a) Suppose that we have a Q-value function represented as a sigmoid function of a set of features:

$$Q(\phi, a) = \frac{1}{1 + e^{\theta^T \phi}}$$

Write down the update rule tat Sarsa would give for this function

(b) What theoretical guarantees would reinforcement learning with this type of function approximator have?

(c) Imagine that you want to implement an exploration strategy that is based on optimism under uncertainty. How would you do this in the context of function approximation?