CSCE 475/875

Game Day 1: Learning Day

Assigned: February 6, 2020 Game Day: February 18, 2020

Introduction

On Learning Day, students are required to practice reinforcement learning as agents. The key is to (1) learn as accurately as possible while (2) earning as much reward as possible. These two objectives might be at odds with each other and thus it is important for the students to choose the correct parameters for the learning at different stages of the learning process.

Note also that learning in multiagent systems also involve the *exploitation vs. exploration* tradeoff. Does one explore in order to learn the best alternatives accurately, or does one exploit whatever it has learned to gain rewards even though the learned information might not be optimal?

The objectives of Learning Day are to learn and familiarize with the reinforcement learning and multiagent learning mechanisms, and to learn how to observe the environment in order to change the parameters of the mechanisms to make them more effective and efficient.

More specifically, you will learn about Q-learning, its learning rate (α) and its discount factor (β).

Recall the following definition from our lectures:

```
**Definition 7.4.1 ($Q$-learning)** *$Q$-learning is the following procedure:*
Initialize the $Q$-function and $V$values (arbitrarily, for example)
**Repeat** *until convergence*
    Observe the current state $s_t$.
    Select action $a_t$ and take it.
    Observe the reward $r(s_t, a_t)$ and next state $s_{t+1}$
    Perform the following updates (and do *not* update any other Q-values):
        $Q_{t+1}(s_t, a_t) \leftarrow (1 - \alpha_t)Q_t(s_t, a_t) + \alpha_t(r(s_t, a_t) + \beta V_t(s_{t+1}))$
        $V_{t+1}(s) \leftarrow \max_a Q_t(s, a)$
**End Repeat**
```

Setup

In the environment there is a set of states, *S*, and a set of actions, *A*. An online website will be setup for you to interact with the environment. The URL of this website will be provided to you on Game Day.

1. Each team will be provided an initial state. Using this website, each team will be allowed to enter an action to perform on an input state. The website will then display the following information: (a) the resulting output state, and (b) the reward if the action is performable on the input state. Otherwise, it will display a message indicating that the action is not performable on the input state, and the team will be returned to the input state.
2. There is no cost to each team in performing an action. Unlike the states, these actions are reusable. That is, you can perform the same action again without exhausting the resources.

3. Once an input state is transitioned to an output state, the input state becomes not available to the team. However, it is possible that the team may arrive at this state at a later time by transitioning to it by carrying out other state-action pairs.
4. It may assumed that that $Q_0(s, a)$ for all performable state-action pairs is set to some random values of your choice initially, and 0.0 for all non-performable state-action pairs. It may be assumed that $V_0(s)$ for all states is set to 0 initially.
5. There will be two rounds of learning.
   a. For Round 1, each team must first determine its own α and β. (Must indicate this in the Pre-game report.) After Round 1, each team is required to submit a $|S|$ x $|A|$ matrix of Q-values (one cell for each state-action pair) by e-mailing the Game Day Monitor. For each cell, depending on the resultant output states, one would have $|S|$ entries. (See Table 1 below. That is, in all, there would be $|S|$ x $|A|$ x $|S|$ entries.)
   b. The teams then are allowed to approach other teams to inquire about their Q-value matrices. This is to allow a team to decide whether to purchase Q-value matrices of the other teams. *Each team is allowed to provide truthful or non-truthful statements about its Q-value matrices.*
   c. Each team is only allowed to buy at most two Q-value matrices of the other teams directly from the Game Day Monitor, $2000 for the first matrix, and $1000 for the second. The team whose matrix is bought will be compensated 50% of the price paid, and they will be informed such before Round 2 begins. *No direct team-to-team negotiations or transactions are allowed.*
   d. To make a purchase, a team is required to e-mail such a request (specifying the desired matrix) to the Game Day Monitor, and the Game Day Monitor will e-mail the said matrix to the team accordingly.
   e. Prior to the start of Round 2, each team must again determine its own α and β. (Must indicate this in the Mid-game report.) After Round 2, each team is again required to submit the updated $|S|$ x $|A|$ matrix of Q-values to the Game Day Monitor.
6. In terms of what will be tallied to win the game day, we will look at two aspects: (a) accuracy in the Q-values (based on the matrix), and (b) the total amount of rewards (minus any amount spent on purchasing or selling matrices) earned while interacting with the website. Each of these scores will be normalized using the best score available, and then both will be added together with equal weights. We will compute the composite score twice, after each round. Then, *using a weighted sum, 0.4 for round 1 and 0.6 for round 2, the two composite scores will then be added to arrive at the final total*. The team with the largest final total will be the winner of the Game Day.

Reward: A Distribution

Each state-action pair transitions to a state of S with a probability p. The reward of a state-action pair follows a Gaussian distribution, with a mean and a standard deviation. You may assume that the maximum reward possible is $5000.

Requirements

Each student group is required to turn in three reports: pre-game strategies, mid-game strategies, and post-game lessons learned.

- Pre-game strategies are to be handed in before the Game Day starts. Without this, your team will be disqualified from the Game Day.

- The report on mid-game strategies consists of your observations noted down on your worksheets during the Game Day.
- Post-game lessons learned are handed in at the end of the Game Day.

Some ideas on what should be included in the reports: your strategies for each round of reinforcement learning, your rationale behind the values of the learning rate and the discount factor, how you divide the members of the group to different tasks, your total rewards for each round and the final grand total, the ordering of the state-action pairs, and finally your conclusion.

Your participation on Learning Day will be graded based on:

- 50% Game Day Reports (pre-game and mid-game strategies, worksheets)
- 50% Game Day Activities (in-class participation on Game Day, team performance)

Table 1. An example Q-value matrix.

	A1	A2	A3	…	…	An
S1	→ S1: ?? → S2: ?? → S3: ?? … … → Sm: ??	→ S1: ?? → S2: ?? → S3: ?? … … → Sm: ??	→ S1: → S2: → S3: … … → Sm:	…	…	→ S1: → S2: → S3: … … → Sm:
S2	→ S1: → S2: → S3: … … → Sm:	→ S1: → S2: → S3: … … → Sm:	→ S1: → S2: → S3: … … → Sm:	…	…	→ S1: → S2: → S3: … … → Sm:
S3						
…						
…						
Sm						