Machine Learning for Communications

Project #4

Reinforcement Learning for Dynamic Spectrum Access

Due December 15, 2019

Note: You may use outside references including books and notes and may discuss the project with your classmates in general terms. However, you may not obtain code or direct assistance from another person.

PLEASE MAKE SURE YOU ANSWER ALL PARTS OF EACH PROBLEM AND CLEARLY MARK YOUR FINAL ANSWERS. YOUR ANSWERS SHOULD BE AS COMPLETE AND CLEAR AS POSSIBLE NOT JUST A LISTING OF THE ANSWER. MATLAB CODE SHOULD BE CLEARLY DOCUMENTED AND SUBMITTED SEPARATELY FROM YOUR REPORT IN A ZIP FILE PER THE INSTRUCTIONS IN LECTURE #1.

I pledge that I have neither given nor received any unauthorized assistance on this project.

_____

(signed)

_____ _____

Name (print) Student Number

1 Mini-Project Overview

In this project you will use reinforcement learning to determine an optimal policy for a communication system attempting to share the spectrum with a second communication system. You may use either Python or Matlab to complete this project.

2 Detailed Description

There are two frequency bands which are shared by two communication systems (which cannot communicate with each other) that we will term System A and System B. You are attempting to find an optimal policy for System A. System B has a behavior that is defined by Table **??**. The state is defined by the spectrum occupancy of the two bands by System B. For example, if System B transmits in the second band only, System A observes the state 01. Based on this observed state, System A must chose an action (i.e., which frequency bands to transmit in). If System A reacts to its observation of state 01 by not transmitting (action 00), System B responds by transmitting in the second band again in the next time slot (Next State = 01). As a second example, if System A observes state 10 and chooses action 01, System B will respond by transmitting 10 with probability 0.8 and 11 with probability 0.2. That is, state 10 will transition to state 10 with probabilty 0.8 and to state 11 with probability 0.2, when action 01 is taken.

The policy of System A, is defined as choosing an action based on the observed state (i.e., the actions of System B). Note that System A does not know which state will follow the current state based on its action, i.e., it does not know Table **??**, but can learn it.

Communication System A assigns rewards to its actions based on the transmission of successful packets, but assigns a penalty for collisions with System B. Specifically, assume that System A observes the state s' takes action a. This results in next state s and reward R:

$$R(s,a) = 10 \times sum(a) - 50 \times sum(a. * s) \tag{1}$$

The term $sum(a)$ replects the number of frequency bands that System A attempts to use. It receives positive reward for each band it attempts to use. However, the term $sum(a.*s)$ reflects the number of collisions it causes which incurs a penalty. As an example, assume System A observes the state $s' = [01]$ and takes action $a = [11]$ resulting in the next state [11]. The accompanying reward R is

$$R([11],[11]) = 10 * 2 - 50 * 2 = -80 \tag{2}$$

3 Required Validation

Create a simulation with $N = 2$ frequency bands and a primary System B that follows the behavior shown in Table **??**.

- Use 10,000 training samples (i.e., random actions by System A) and an initial state of 01 to estimate the state transition probability matrix and the reward matrix. Show that this matches the behavior in Table **??** and equation **??**.
- Use policy iteration to determine the optimal policy of System A.

Table 1: Behavior of System B (Next State) Based on the Action of System A (Action) and the Previous State

State	Action	Next State	
00	00	00	
	01	10	
	10	01	
	11	11	$p = 0.8$
		00	$p = 0.2$
01	00	01	
	01	10	$p = 0.8$
		11	$p = 0.2$
	10	01	$p = 0.8$
		11	$p = 0.2$
	11	11	$p = 0.8$
		00	$p = 0.2$
10	00	10	
	01	10	$p = 0.8$
		11	$p = 0.2$
	10	01	$p = 0.8$
		11	$p = 0.2$
	11	11	$p = 0.8$
		00	$p = 0.2$
11	any	11	$p = 0.8$
		00	$p = 0.2$

- Assuming an initial state of 10, and using the optimal policy determined in the previous step, plot the number of collisions over 10,000 steps. Also plot the rewards obtained over those same time steps.
- Compare the collisions and rewards with those obtained by a reactive system that attempts to use frequency bands that are open based on the observed state. In other words, observing 00 results in action 11, observing 01 results in the action 10, etc.