P4: TRAIN A SMARTCAB TO DRIVE

ANDY PEREZ

1. Implementation of a basic driving agent

Since we are initially setting the move/action to be entirely random (using python's random number generation to choose one of the four possible actions), it's not surpising that the car moves rather randomly and erratically. Turns are as likely as moving forward, so it tends to snake around quite a lot. It can take a considerable amount of time, but the car does eventually reach the target location, although it never reached it on time during my runs.

2. Identify and update state

The set of states I chose to model the environment are as follows, along with a justifaction for each.

- Orientation of traffic light at intersection (N-S or E-W)

  When at an intersection and making a decision as to whether proceed or wait, it's vitally important to know whether the light is red (and hence you have to wait else incur a negative reward) or green (in which case you can proceed forward). There are no other variables besides inputs['light'] that can convey this information, thus it's vital to include it.

- Whether there is oncoming traffic, traffic to the left, or traffic to the right

  These are all related so instead of considering them separately I've put them all under one item. It is important to know the position of other cars at an intersection before proceeding wth an action. According to the rules of the road, one needs to be aware of other cars at the intersection before making a decision so that they can be given the right of way if necessary, particularly when one is attempting to make a turn.

- The next waypoint

  There needs to be some way of including in the status variables some sort of guidance as to which direction the car is to go, or else it'll have no reason to move towards the destination square as opposed to some random square. This is the direction we want to go and that we want the car to go unless there's a good reason not to, so hence it should be included.

- Whether the deadline is imminent

  This is another factor that affects how one should drive. Since the reward of reaching the desired end state is larger than the possible positive or negative rewards in absolute value at any intersection, it may be the case that when a car is near the goal but almost out of time it should be willing to incur short-term negative rewards for a better shot at the large reward of reaching the goal. Since the negative reward is -1, I chose to use whether

the deadline is within 6 or not, for the environment isn't very large, and there's little gain to be had to start rushing and incurring negative rewards before, lest the negative rewards incurred outweigh the positive end reward.

The environment is modeled as an ever-changing intersection with the use of these variables. Cars appear, lights change, and waypoints change as given by the actual full environment outside these variables. However, this myopic worldview is sufficient, since the car does not need to see past the string of intersections or know it's position on the map in order to make its decisions except for figuring out how to get to the destination. However, that is already taken care of by the planner, which takes into consideration the bearing and the relative positions of the car and the destination. There isn't any other data that we can use unless we have control over the planner too and use reinforcement learning on the planner as well.

3. Implement Q-Learning

The agent is much more effective at reaching the target after implmenting Q-learning, at least after going through a large amount of rounds of learning. Initially, it's still going to be mostly random, with the relatively high Q initialization values causing it to be very exploratory. It still doesn't reach the target a lot of the time, but that's still a massive improvement. It seems to go in circles sometimes for no particular reasion, making four right turns in a row sometimes.

How the Q learning progresses is itself quite interesting at well. It takes rather little time for the car to figure out how to proceed in a vacuum with no other cars in the vicinity. In this case, there are only two light states, four actions, and three waypoint possibilities for a total of $2 * 3 * 4 = 24$ total state-action pairs. Thus, it takes only a few rounds (less than six) for the car to figure out that it should just go in the direction of the waypoint if the light is in it's favor. However, anthromorphisizing a bit, the car seems to get flustered whenever there are other cars at an intersection, at least to start. These rules take a long time to learn correctly since the existence of other cars is responsible for a large amount of the size of the state space. In fact, it is the nuances of the interaction with other cars and lights and right of way that requires the use of a long training period.

4. Enhancing the Driving Agent

First of all, the *deadline_approaching* variable was removed from the list of states that Q learning operates on. The reason being that there were already too many states, and having this extra state doubled the number of possible states. It would be worth using it, possibly, if there was a way to couple states where *deadline_approaching* was True/False so that learning Q in one affected Q in the other. If it's a good idea to normally make a right turn in a certain situation, it's at least a decent idea to do so when a deadline is approaching as well. The car arrived at its destination quite before the target time almost always though, so there wasn't much learning on the proper Q values for states with this variable set to True, so I just removed it.

Several rounds of parameter testing were used to find optimal values of the other parameters. The tables used to figure them out can be seen at the end of this document.

Furthermore, I reduced the initialized default values for Q from 10 to 1.25. It turned out that setting them to 10 slowed down the convergence of Q to it's actual values by too much. Because the learning rate went as $\alpha = 1/t$, by the time some of the rarer state-action pairs were being accessed, the learning rate was too slow to reduce the 10s to anywhere near their actual values. However, when Q is nearer the actual expected values of states, which tend to be around 2-3, then we create a model that is still very exploratory, but not too much so to the point of not learning the good choices in time. The value 1.25 was chosen after testing (tables below). I reduced ϵ , the frequency with which random actions are taken, from .1 to .03, although this didn't have much effect except to make it so that there are less random actions in the model once it's been trained. Having epsilon go even lower but start higher by allowing it to vary with t would have been appropriate as well, but just having it set to a constant sufficed for our purposes.

Rare state-action pairs were, well, rare. And because of this, since α was a globally set variable, sometimes α would decrease too quickly to let us meaningfully learn how to act on rare states. This was especially a problem when Q was initialized to 10, and although this problem was mitigated when Q was initialized to a much lower value, it is still something that was able to be improved. Thus, I made $\alpha = \alpha(S, A)$, the learning rate was made to be a function of the state-action pair, where it went as $1/5t$, except where t here was the amount of times the state-action pair (S, A) was visited as opposed to how many times states were visited in general. It makes no sense not to learn how to handle a novel state from the first occurance of a novel state just because it happened to occur 2000 t steps in. I multiplied t by a constant, however, so that for common states there wasn't too much of a big difference from how it was before α was made to vary with the state and action. This constant was initially chosen to be set to .2, and this value turned out to be near optimal.

The discount factor was initially set to $\gamma = 1/2$, but after some testing a slightly higher discount factor of $\gamma = 1/2.25$ was found to work better. A value of $1/2$ is very commonly used, but there's no specific reason why we should expect it to be the optimal value for solving our problem, and it turned out increasing it slightly improved our results.

My agent definitely gets quite close to finding an optimal policy! After the 100 trials of learning, it gets to the destination very efficiently (at least as efficiently it can reach the destination using the built-in planner) and rarely incurs penalties. From many runs of the final code, it has never failed to reach the destination in time with positive reward. Typically, in the 10 test runs after the 100 trials, it only makes a suboptimal stop maybe four to five times, occasionally because of the constant ϵ value resulting in a random wrong action, on other occasions because of a novel state that not much learning has occured on. There are 1536 state-action pairs, and most of them are uncommon (involving multiple cars at an intersection), so we can't expect the agent to learn every one of them, mostly just the more common ones.

4.1. **Parameter Selection.** Below are several tables produced by testing runs. A model was trained on 100 trips and then tested on 10 trips for each value of the four hyperparameters under consideration. ϵ is the random action rate, γ is the discount rate, k is a multiplier to the learning rate, such that $\alpha(S, A) = k/t$, and Q is the value at which the Q function is initialized. "Steps" is the number of steps,

on average, required to reach the goal in the 10 testing runs, and "P Steps" is the number of steps, on average, with positive reward.

4.1.1. *First Run.*

ϵ	γ	k	Q	Steps	P Steps
0.05	0.5	0.2	1	11.7	11.6
0.05	0.5	0.2	0.5	14.4	14.3
0.05	0.5	0.1	0.5	16	15.9
0.05	0.5	0.1	1	19.1	18.7
0.02	0.5	0.2	1	19.5	19.5
0.02	0.5	0.2	0.5	21.4	21.3
0.02	0.5	0.1	0.5	24.9	24.6
0.02	0.5	0.1	1	29.1	28.7

4.1.2. *Second Run.* A very large number and range of parameters were tried, so only the top 50 performing choices are listed here for the sake of brevity.

ϵ	γ	k	Q	Steps	P Steps
0.15	0.5	1	2	10.2	9.7
0.1	0.5	0.2	2	12.2	12.2
0.05	0.25	5	2	12.2	12
0.1	0.75	1	8	12.5	11.3
0.15	0.5	5	1	12.6	12.1
0.05	0.25	1	3	12.7	12.5
0.15	0.25	0.2	2	12.9	12.1
0.15	0.5	1	3	13.3	12.7
0.15	0.75	1	6	13.5	13.1
0.15	0.75	0.2	2	13.7	13.1
0.05	0.5	1	3	13.8	13.3
0.05	0.5	5	3	13.8	12.8
0.1	0.25	1	1	13.9	13.3
0.1	0.5	0.2	3	13.9	13.3
0.1	0.25	1	3	13.9	13.3
0.1	0.75	5	6	13.9	13.3
0.15	0.5	0.2	1	14	13
0.1	0.75	0.2	6	14	13.3
0.1	0.75	5	1	14.1	13.3
0.05	0.25	1	2	14.1	13.9
0.1	0.25	0.2	3	14.2	13
0.05	0.25	0.2	2	14.3	14
0.15	0.25	5	2	14.3	13.4
0.15	0.5	0.2	2	14.4	14
0.1	0.25	5	2	14.4	13.8
0.1	0.5	1	2	14.6	14
0.1	0.25	0.2	2	14.6	13.9
0.05	0.75	1	6	14.7	14.2
0.15	0.25	0.2	1	14.8	14.4
0.05	0.25	1	1	15.1	14.5

0.05	0.5	0.2	2	15.1	14.7
0.1	0.5	1	1	15.2	15
0.1	0.25	5	3	15.6	14.9
0.05	0.5	0.2	1	15.7	15.7
0.1	0.5	1	3	15.7	14.9
0.1	0.75	1	6	15.7	14.8
0.05	0.25	0.2	3	15.8	15.3
0.05	0.5	5	1	15.9	15.6
0.05	0.75	1	1	15.9	14.9
0.1	0.5	1	8	15.9	15.3
0.1	0.5	0.2	1	16	15.7
0.15	0.5	5	2	16	15
0.15	0.25	1	2	16	15.4
0.05	0.5	1	1	16.2	15.8
0.05	0.75	0.2	8	16.2	15.2
0.15	0.25	5	3	16.8	15.6
0.15	0.5	0.2	3	16.9	15.9
0.05	0.25	0.2	1	17	16.3
0.1	0.5	5	2	17.1	16.7
0.05	0.25	1	6	17.2	16.2

4.1.3. *Third Run.* Since a large amount of values were tested here, small ranges of values were used, and since we could expect some variability, instead of choosing the parameter values that give the lowest number of steps, I instead assumed the relationships between the parameters and the number of steps is linear in the range that I tested and found the most likely optimal values under this assumption. The only difference between the parameter choices used adn the top performing one according to this table was a choice of $\epsilon = 0.03$ instead of $\epsilon = 0.01$.

ϵ	γ	k	Q	Steps	P Steps
0.01	0.444	0.2	1.25	11.4	11.4
0.02	0.444	0.1	1	12.4	12.3
0.03	0.444	0.1	1.25	12.9	12.9
0.03	0.5	0.1	1.25	14.5	14.4
0.03	0.5	0.2	1.25	15.2	14.8
0.01	0.5	0.1	1	15.6	15.6
0.03	0.444	0.2	1.25	15.8	15.2
0.02	0.444	0.1	1.25	16.1	16.1
0.01	0.571	0.1	1	16.2	16.2
0.02	0.5	0.2	1.25	16.2	16.2
0.03	0.5	0.2	1	16.7	16.5
0.03	0.444	0.2	1	16.7	16.6
0.03	0.571	0.2	1	16.8	16.6
0.02	0.571	0.2	1	17.9	17.7
0.02	0.571	0.1	1.25	17.9	17.8

0.01	0.5	0.1	0.75	18.3	18.3
0.02	0.444	0.2	1.25	18.4	18.3
0.02	0.444	0.2	1	18.7	18.5
0.01	0.5	0.2	1	19.1	19.1
0.03	0.5	0.1	0.75	19.5	19.3
0.03	0.444	0.1	1	19.7	19.5
0.01	0.571	0.2	0.75	19.7	19.6
0.03	0.571	0.2	1.25	20	19.9
0.02	0.444	0.1	0.75	20.2	20
0.02	0.5	0.1	0.75	20.8	20.7
0.03	0.571	0.1	1.25	22	21.5
0.02	0.444	0.2	0.75	22.7	22.7
0.01	0.444	0.2	1	23.3	23.1
0.01	0.444	0.2	0.75	24.1	24.1
0.02	0.5	0.2	1	24.7	24.5
0.03	0.5	0.2	0.75	25.7	25.6
0.01	0.571	0.2	1	25.9	25.7
0.02	0.571	0.2	1.25	26	25.7
0.03	0.444	0.1	0.75	26.2	26
0.01	0.571	0.1	0.75	26.2	26.1
0.01	0.571	0.2	1.25	26.7	26.5
0.01	0.5	0.1	1.25	27.6	27.4
0.02	0.5	0.1	1	27.7	27.7
0.02	0.571	0.1	0.75	27.7	27.7
0.01	0.5	0.2	1.25	28	27.7
0.01	0.5	0.2	0.75	28.5	28.4
0.01	0.444	0.1	1.25	28.5	28.4
0.01	0.571	0.1	1.25	28.5	28.1
0.02	0.571	0.2	0.75	28.7	28.5
0.03	0.444	0.2	0.75	28.9	28.7
0.03	0.571	0.1	0.75	28.9	28.6
0.03	0.5	0.1	1	29	28.7
0.01	0.444	0.1	0.75	29.1	28.9
0.02	0.5	0.2	0.75	29.2	29
0.02	0.5	0.1	1.25	29.2	29
0.02	0.571	0.1	1	29.7	29.4
0.01	0.444	0.1	1	30.5	30.2
0.03	0.571	0.1	1	30.5	30.4
0.03	0.571	0.2	0.75	30.5	30.2