Group 10

CS-534 Artificial Intelligence
Assignment-3 Report

Part-3

Q1. Explain your state representation, and why you chose it.

Ans. The main components of our state representation are:

- **Earliest Waiting Time**
- **Number Of Packages**

Our state representation in tabular form is:

Policy Table

Earliest Wait Time	0	1	2	3	4	5	6	7	8	9
60	?	W	?	?	?	?	?	?	?	?
59	?	W	?	?	?	?	?	?	?	?
58	?	W	W	?	?	?	?	?	?	?
57	?	W	W	W	?	?	?	?	?	?
56	?	W	W	W	W	?	?	?	?	?
55	?	W	D	W	?	?	?	?	?	?
54	?	W	W	W	W	W	?	?	?	?
53	?	D	W	W	?	W	?	?	?	?
52	?	W	W	W	W	W	?	W	?	?
51	?	W	W	W	W	W	?	D	?	?
50	?	W	W	D	D	W	W	D	?	?
49	?	W	W	W	D	W	D	D	?	?
48	?	D	W	D	W	W	D	D	W	W
47	?	W	D	W	D	W	D	D	W	W
46	?	W	W	W	D	D	W	W	W	W
45	?	W	D	D	D	D	D	W	W	?
44	?	D	D	D	D	D	D	D	W	?
43	?	W	W	D	D	W	D	W	D	?
42	?	W	W	D	D	W	W	W	?	?
41	?	W	W	D	D	D	W	W	?	?
40	?	W	D	D	D	W	D	D	W	?
39	?	D	D	D	D	D	W	D	W	?
38	?	W	D	D	D	W	D	D	W	?
37	?	W	D	D	D	W	W	D	?	?
36	?	W	D	D	D	D	D	W	?	?
35	?	D	D	D	D	D	D	W	?	?
34	?	D	D	W	D	D	W	W	?	?
33	?	D	D	D	D	D	W	?	W	?
32	?	W	D	D	D	D	W	W	D	?
31	?	W	D	D	D	W	W	?	D	?
30	?	W	D	D	D	D	W	?	?	?
29	?	D	D	D	D	W	D	W	?	?
28	?	D	D	D	W	W	D	W	?	?
27	?	W	D	D	W	W	D	W	?	?
26	?	W	D	D	W	W	D	D	?	?
25	?	W	D	D	D	D	D	W	?	?
24	?	D	D	W	D	W	W	?	?	?
23	?	W	D	D	D	D	W	W	?	?
22	?	D	D	W	D	W	W	?	?	?
21	?	D	D	W	D	W	W	?	?	?
20	?	W	D	D	W	W	W	W	?	?
19	?	W	D	D	W	W	?	?	?	?
18	?	D	D	D	D	W	W	?	?	?
17	?	W	W	D	W	W	W	?	?	?
16	?	W	D	W	W	W	W	?	?	?
15	?	W	D	W	D	W	W	?	?	?
14	?	D	W	W	W	W	?	?	?	?
13	?	D	D	W	W	W	?	?	?	?
12	?	D	D	W	D	W	?	?	?	?
11	?	D	W	W	W	W	?	?	?	?
10	?	W	D	D	W	W	?	?	?	?
9	?	W	W	D	W	D	?	?	?	?
8	?	D	W	W	W	W	?	?	?	?
7	?	D	W	W	W	W	?	?	?	?
6	?	D	W	W	W	W	?	?	?	?
5	?	D	W	W	W	?	?	?	?	?
4	?	D	W	W	W	?	?	?	?	?
3	?	W	W	W	W	?	?	?	?	?
2	?	W	W	W	?	?	?	?	?	?
1	?	W	W	?	?	?	?	?	?	?
0	W	W	?	?	?	?	?	?	?	?

Number Of Packages

1. **Earliest Waiting Time:** Earliest waiting time corresponds to the time associated with package which was created at the earliest in the package queue. It is a key component as due to which the driver will get the incentive to opt for delivery after a large waiting time. This is because a large waiting time is associated with a large negative reward.
2. **Number of Packages:** Number of packages is very important component because the positive reward on delivering packages is proportional to the number of packages being delivered.

We also tried to represent our states with some other parameters like:

- Percentage of truck filled
- Maximum destination to deliver

But we preferred the above mentioned combination(Earliest Waiting Time and Number of Packages) due to their larger state repeatability. This in-turn helped in training the agent in a much better and efficient way.

The above presented Policy Table has been generated with input parameters:
Truck Capacity: 30, Length of Road: 25, Penalty to start a delivery: -250, Number of clocks: 50,000.
Our observations for the policy generated are as follows:

- The truck successfully learns to avoid delivering when there is no package
- The chance to deliver the packages increases with the increase in number of packages
- The chance to deliver the packages increases with the increase in earliest waiting time
- There may be some unexpected decisions at some states (mostly far away). This is due to the fact that they have not been visited that often and hence their Q-value was not appropriately updated.

There has been another table generated which shows up how many times were the states visited during training. Due to size concerns, it has not been presented in the report but has been included in the submission file

Q2. How will you account for future summed rewards that would go to infinity during infinite time for simulation?

Ans. The future summed rewards can very much expected to go to infinity if we run simulation to infinity. But, this will not happen because with time, the step-size parameter (**alpha**) decays to a very small value close to zero. This causes the agent to follow the

learned policy without affecting the q-values by any appreciable amount owing to our Q-Update equation:

current_state.q[action]+=self.alpha * (self.reward + self.gamma * (np.max(next_state.q)) - current_state.q[action])

Hence, With Alpha decaying to almost 0 with time, the update to the Q value will nearly be zero leading to a constant Q value. Hence, our future summed up rewards will not end up to infinity.

Q3.1: Explain the policy your simulation learns for a road length of 25, a truck capacity of 30, and a penalty of starting the truck on a deliver action of -250.

- Below is the policy chart learned after 50,000 ticks for a road length of 25, a truck capacity of 30, and a penalty of starting the truck on a deliver action of -250.

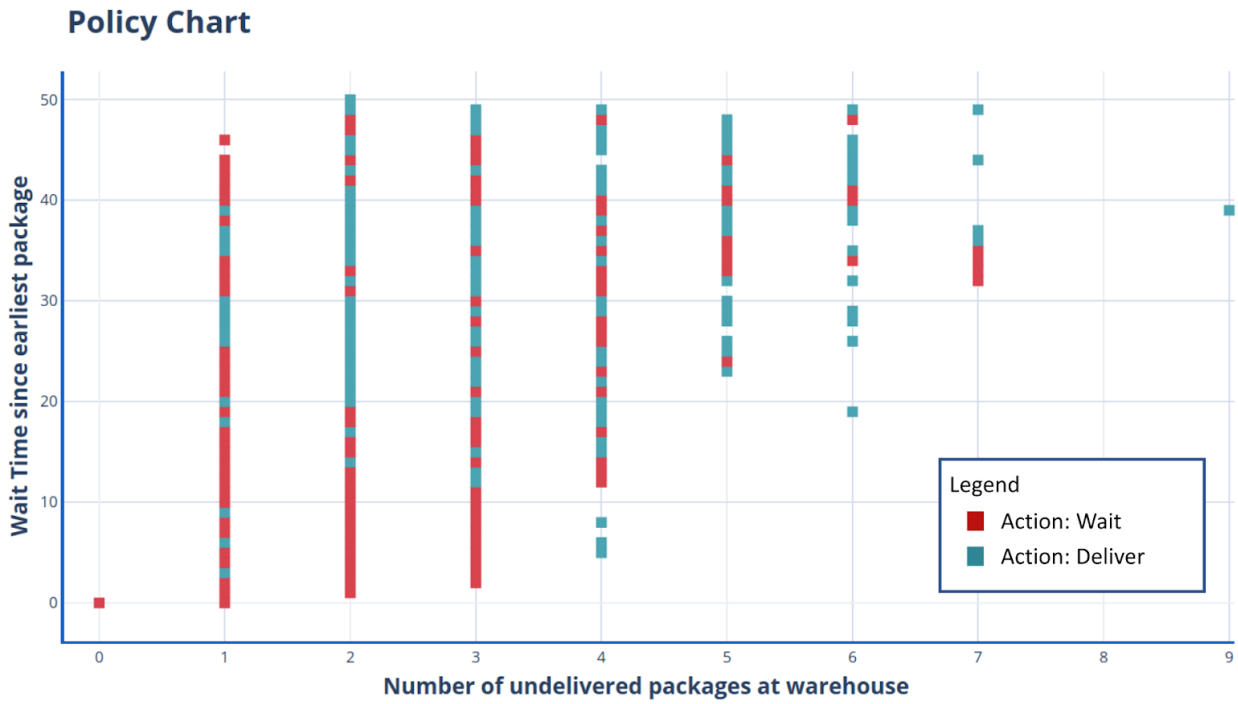

Fig 1: Policy Chart

As discussed in earlier questions, our state is defined by two factors: Number of undelivered packages at the warehouse (X-axis in the graph) and Time elapsed since the generation of the earliest undelivered package (Y-Axis). The red points show that the learned policy is to wait, whereas the blue points denote the learned policy of delivering. We can study the chart in three steps:

- ***Empty truck:*** At any point in time, if there is no undelivered package at the warehouse, both of our variables will be 0. We see in the policy chart, that the agent

has learned not to start delivery if there are no packages at the warehouse. This can be seen at the red point at the co-ordinate (0,0)

- ***A small number of packages (1-3):*** If the undelivered packages are small (1-3), we see that at the initial stage, the agent wants to wait for the next package. Hence for the 1-3 undelivered packages, we see that the agent prefers to wait. But if the wait time increases, the preferred action can be seen as delivery.

- ***A large number of packages (4-9):*** For more packages(4-9) we see that the agent prefers to deliver rather than waiting. This means, that the agent has learned that if there are more packages at hand, it is better to deliver them and get more reward, than waiting for more and losing the reward due to waiting.

Q 3.2: How much reward does your agent receive on average after it is trained?

- The training agent was run for 10 times, for a road length of 25, a truck capacity of 30, and a penalty of starting the truck on a deliver action of -250 and 50,000 ticks. Below is the reward agent had accumulated at the end of the training. The Mean of the rewards can be seen as 84811.8

Trial Number	**The reward at the end of Training**
Trial 1	72562
Trial 2	125658
Trial 3	125170
Trial 4	78369
Trial 5	3119
Trial 6	112678
Trial 7	86566
Trial 8	21706
Trial 9	122926
Trial 10	99364
Average	**84811.8**

Q 3.3: After your agent is finished training, run it for enough iterations ticks to get a large enough sample for average reward.

- At the end of the training, we test the agent for 10,000 ticks each for 10 times. After the testing is done, we display the average of rewards. Below is the table containing the average rewards after 10 iterations each. Hence the 'Average' shown below is actually the average of 100 iterations of 10,000 ticks. As we see, the performance of the agent is very consistent after it has trained.

Trial Number	**The average reward of 10 iterations**
Trial 1	48293.5
Trial 2	52232.8
Trial 3	56826.4
Trial 4	54811.1
Trial 5	48369.0
Trial 6	56566.1
Trial 7	54704.9
Trial 8	56407.0
Trial 9	53200.0
Trial 10	52003.3
Average	**53341.41**