# Smart Cab

**Implement a basic driving agent**

Q1: In your report, mention what you see in the agent’s behavior. Does it eventually make it to the target location?

**Identify and update state**

Q2: Justify why you picked these set of states, and how they model the agent and its environment.

**Implementing Q-Learning**

*Q3:* What changes do you notice in the agent’s behavior?

**Enhance the driving agent**

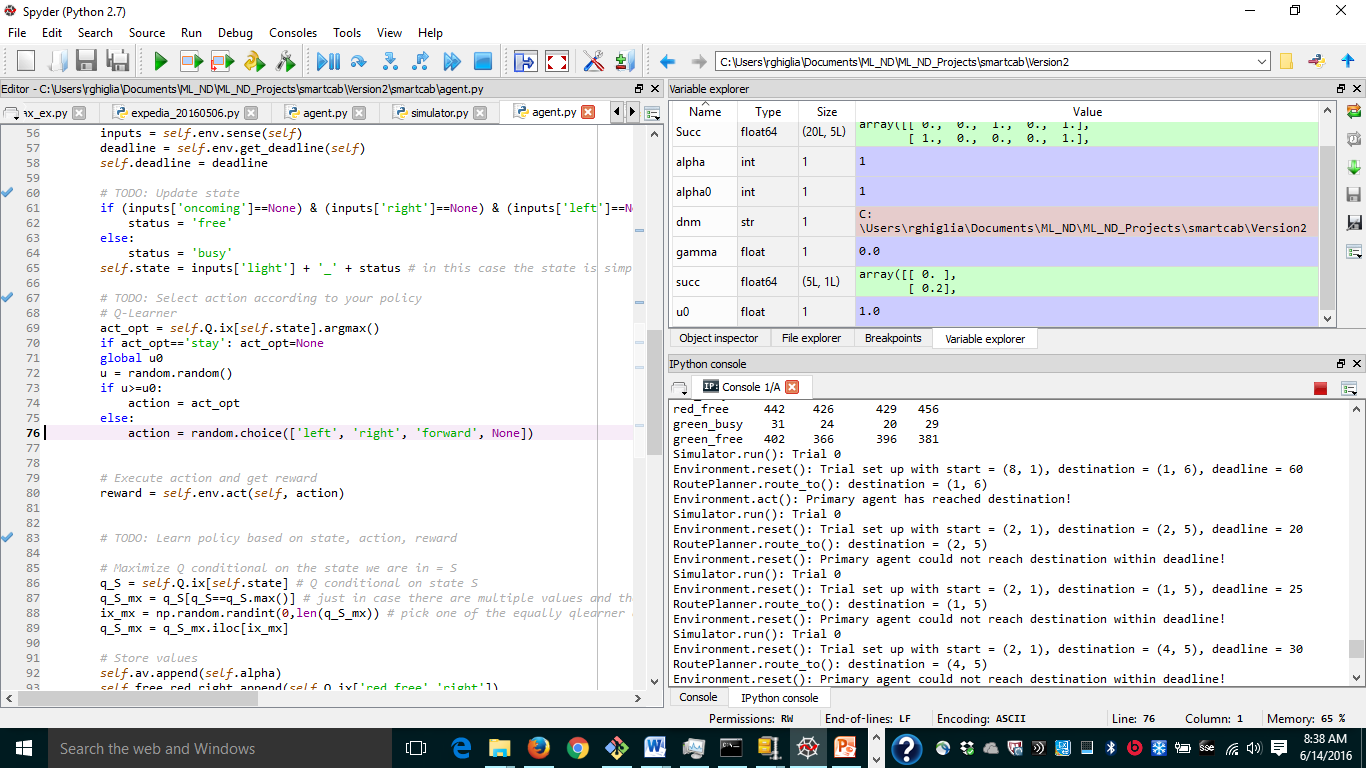
Q4a: Report what changes you made to your basic implementation of Q-Learning to achieve the final version of the agent. How well does it perform?

Q4b: Does your agent get close to finding an optimal policy, i.e. reach the destination in the minimum possible time, and not incur any penalties?

Given the feedback of first review I will try to keep the code as unchanged as possible and refer to previous work, now in the Appendix. Here I will address the issues pointed out in the review. For the most part this discussion should be self-contained with occasional reference to previous version in the appendix.

**Implement a basic driving agent**

Several driving strategies had been tried previously: unconditional left (‘stubborn’), random, guided (= planner’s action), and optimal (Q-learner) (see Appendix - Implement a basic driving agent). For reasons to be explained later in this version I implement only one strategy:

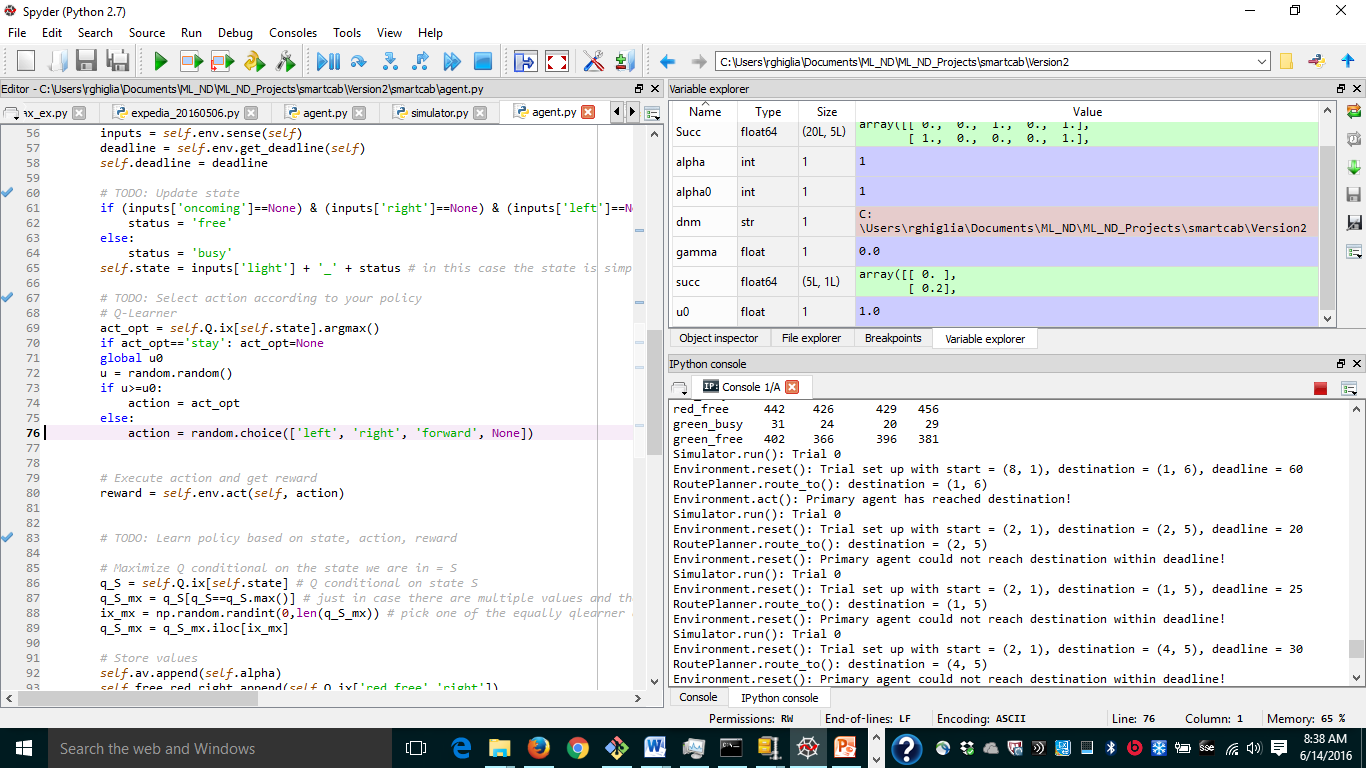


The strategy is a combination of optimal policy under q-learning and a random action. The parameter u0 controls the relative proportion of the two policies. With u0=0 the policy is only optimal, with u0=1 the policy is always random and with u0 some values in [0, 1) say 0.1 the policy will be random 10% of the time. This allows the learner to randomly explore the state-action space with a certain probability = u0.

Behavioral differences in the different strategies have been discussed in the Appendix - Implement a basic driving agent.

**Identify and update state**

The state space has been enlarge to include both the light and the busyness of the intersection



**Implementing Q-Learning**

The Q-learning policy has been implemented as follows:

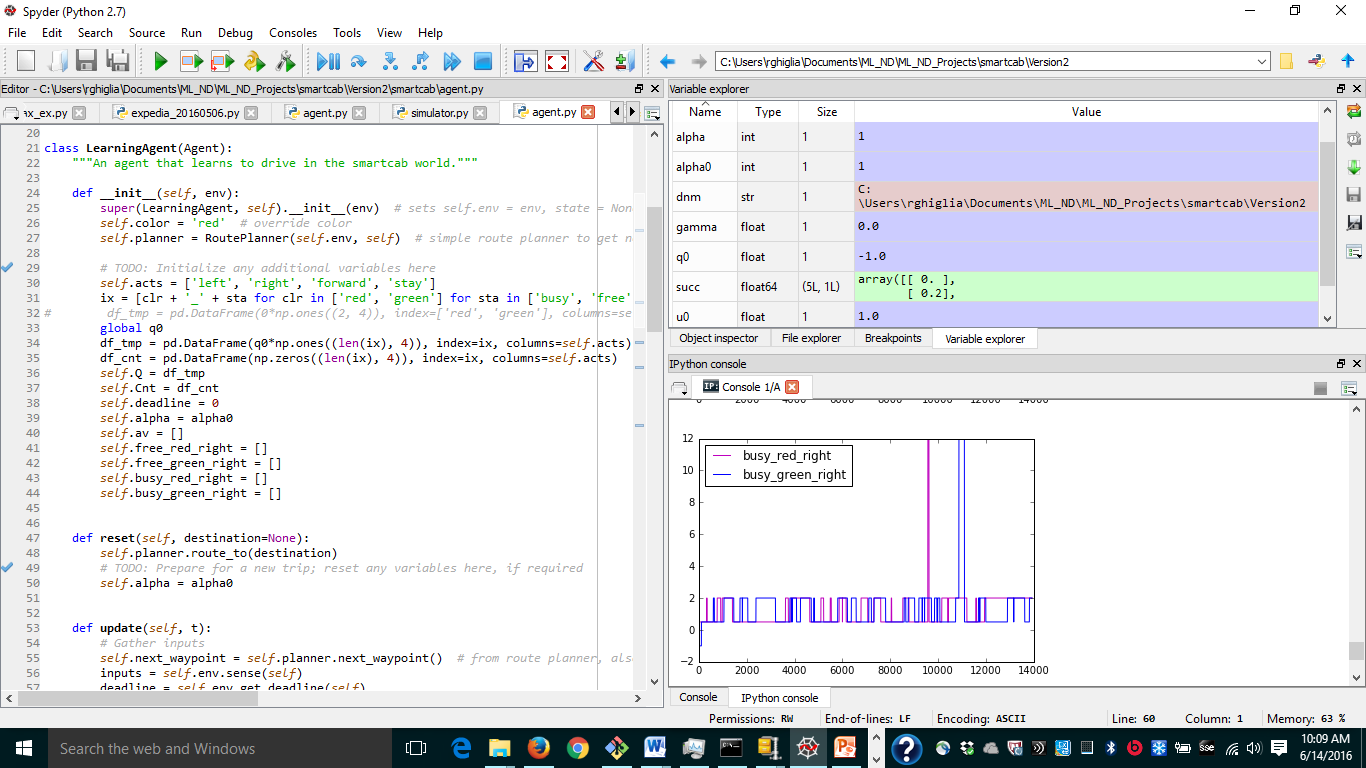
he learning rate α determines the ‘speed’ at which new information is incorporated into the Q function. When α is zero, there is no learning and the Q function remains equal to the initialized value of . I implemented an initialization so that each entry in the matrix Q is equal to the same number at inception. Finally, the parameter γ determines the balance between the immediate reward and the expected reward from future state-actions while behaving optimally. When γ is zero, the Q function is updated just according to the immediate reward. Conversely, when γ is large the Q function depends mostly on the future expected rewards. However, a very large γ would prevent much learning because the update would depend mostly on the initial estimate . Unlike α that ranges between 0 and 1 there is no ‘natural’ value for γ because it will depend on the numerical scale of the reward. In general if one initializes Q with values in the ballpark of the reward a γ of 1 would (roughly) equally weigh the immediate reward with expected future rewards.

In summary, the problem has been parametrized as follows:

* : initial value of all entries in
* : learning rate
* : decay value of the learning rate
* : immediate vs. future reward trade-off
* : probability of a random action

Side note:

In the code review it was suggested that my previous implementation reset the initial estimate of Q at the beginning of a run. Actually it set it to a value that was defined globally which held memory of the estimate over previous trials. In any case I simplified the code so it should be more evident that the Q matrix is initialized only once and then it is not reset between trials:



**Enhance the driving agent**

As part of the enhancement I augmented the state-action space to include the busyness of the intersection. The following will discuss in more depth how I used the above parameters to fine-tune the learner.

As part of exploratory analysis I observed:

1. The value of alpha decayed quite fast
2. Some action-states were not really sampled

To address the first point, first note that the decay is necessary to guarantee convergence to the mean of Q. That said, from a practical point of view, if alpha becomes really small there is no updating involved any more. Instead of setting an alpha very (very) close to 1 I implemented the following:

* Split the learning into 2 parts
  + A batch of trials (nB in the code)
  + A set of epochs (nE in the code) where I run nB trials per epoch
* Reset alpha after each trial

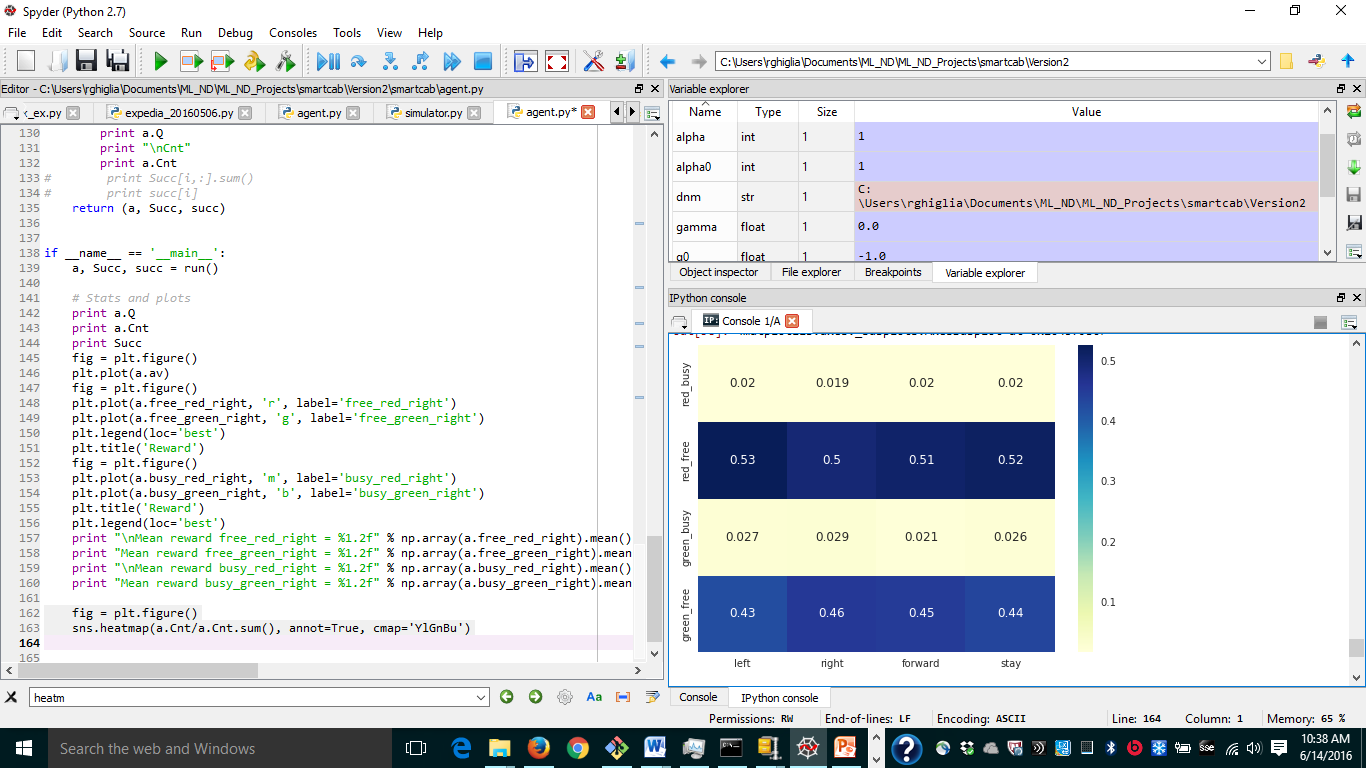
This is implemented in code line 50 (figure above). A typical profile of how alpha changes over time is shown below (next discussion).

To address the second point I added the random action with probability u0.

Before being more specific I investigated the reward function in isolation. This can be achieved by the following setting which turns the agent into a random actioner who only cares about immediate reward:

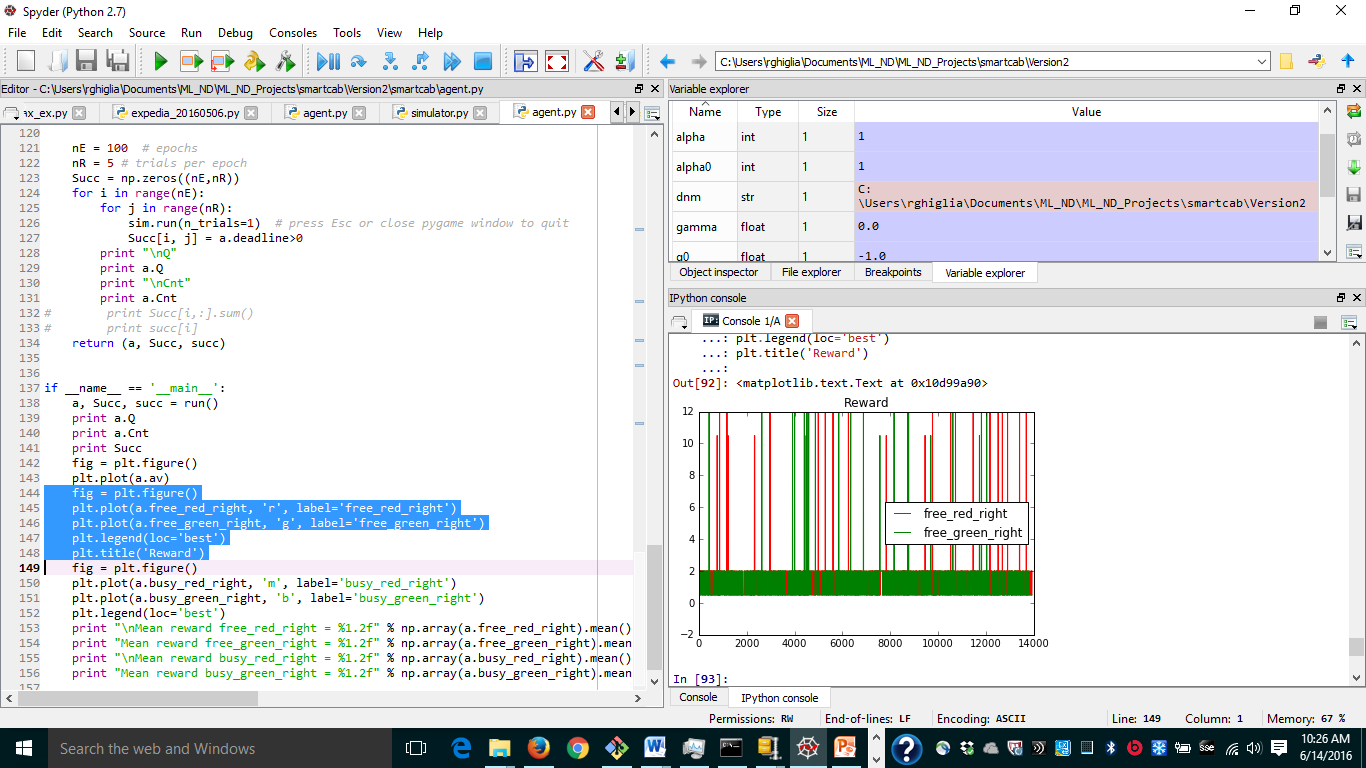
alpha0 = 1, gamma = 0.0, u0 = 1.0.

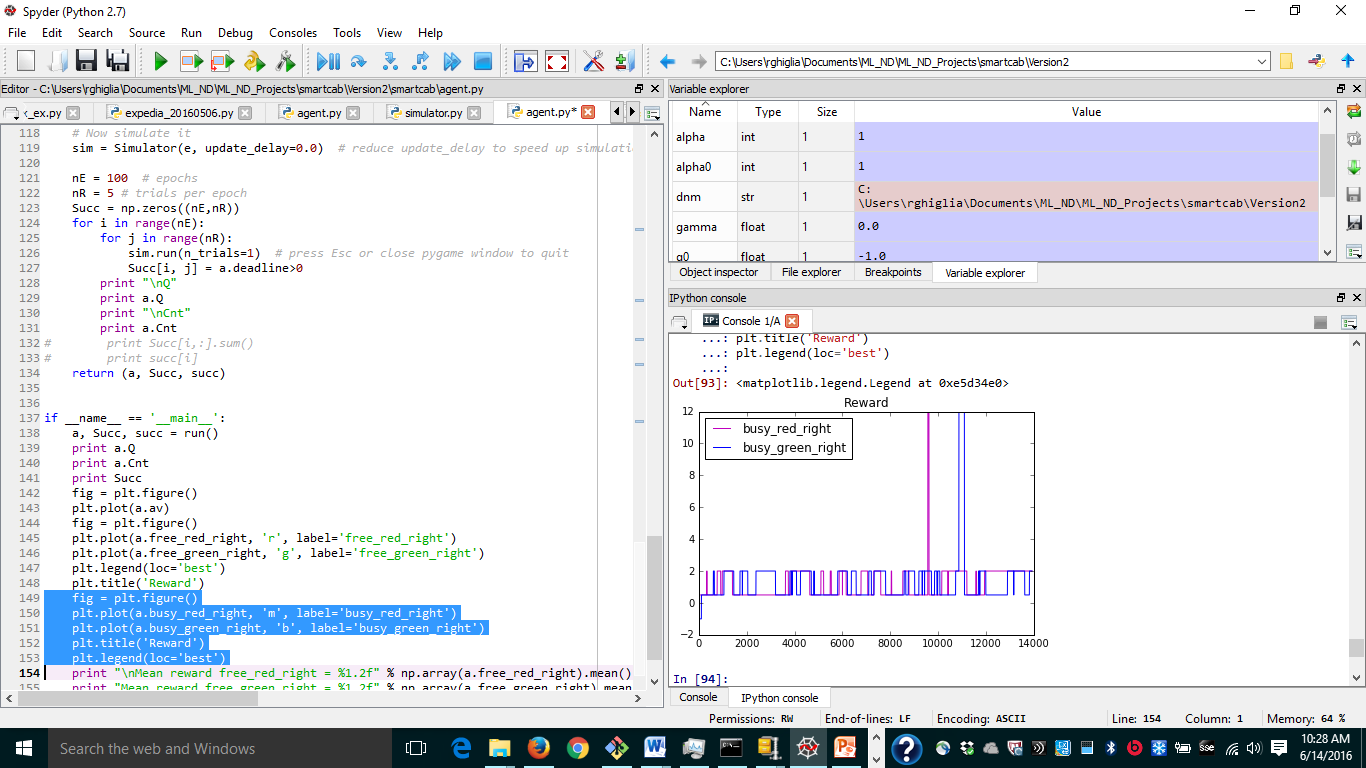
This makes the Q function essentially sampling the state-action space. The following figures shows the % of time each state-action is visited.



As we can see, the intersections are mostly free, actions are taken in equal proportion (values across columns are roughly) the slight biases, e.g. the light being red or green if the intersection is free, should be sampling error.

The following figure shows the reward for 4 state-actions (100 epochs with 5 trials each):





I would have expected to see a slight bias in reward for right turn on red, but if there is one it is not markedly obvious (each of the following runs have 100 epochs with 5 trials each).

Run #1:

Mean reward free\_red\_right = 1.33

Mean reward free\_green\_right = 1.36

Mean reward busy\_red\_right = 0.99

Mean reward busy\_green\_right = 1.39

Run #2:

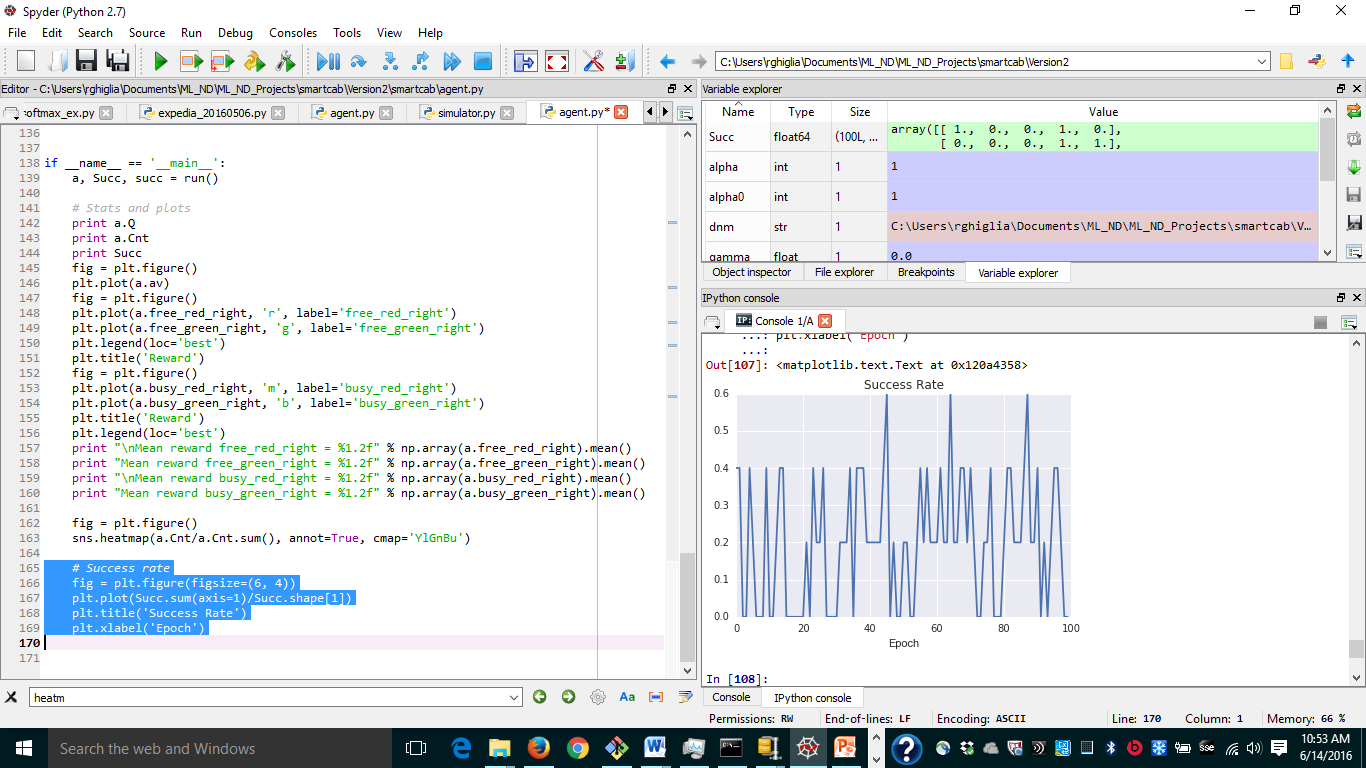
Mean reward free\_red\_right = 1.35

Mean reward free\_green\_right = 1.37

Mean reward busy\_red\_right = 1.28

Mean reward busy\_green\_right = 1.12

The figure below shows the success rate of the random agent across epochs. This will constitute the benchmark.



# Parameter set #1

alpha0 = 0.2

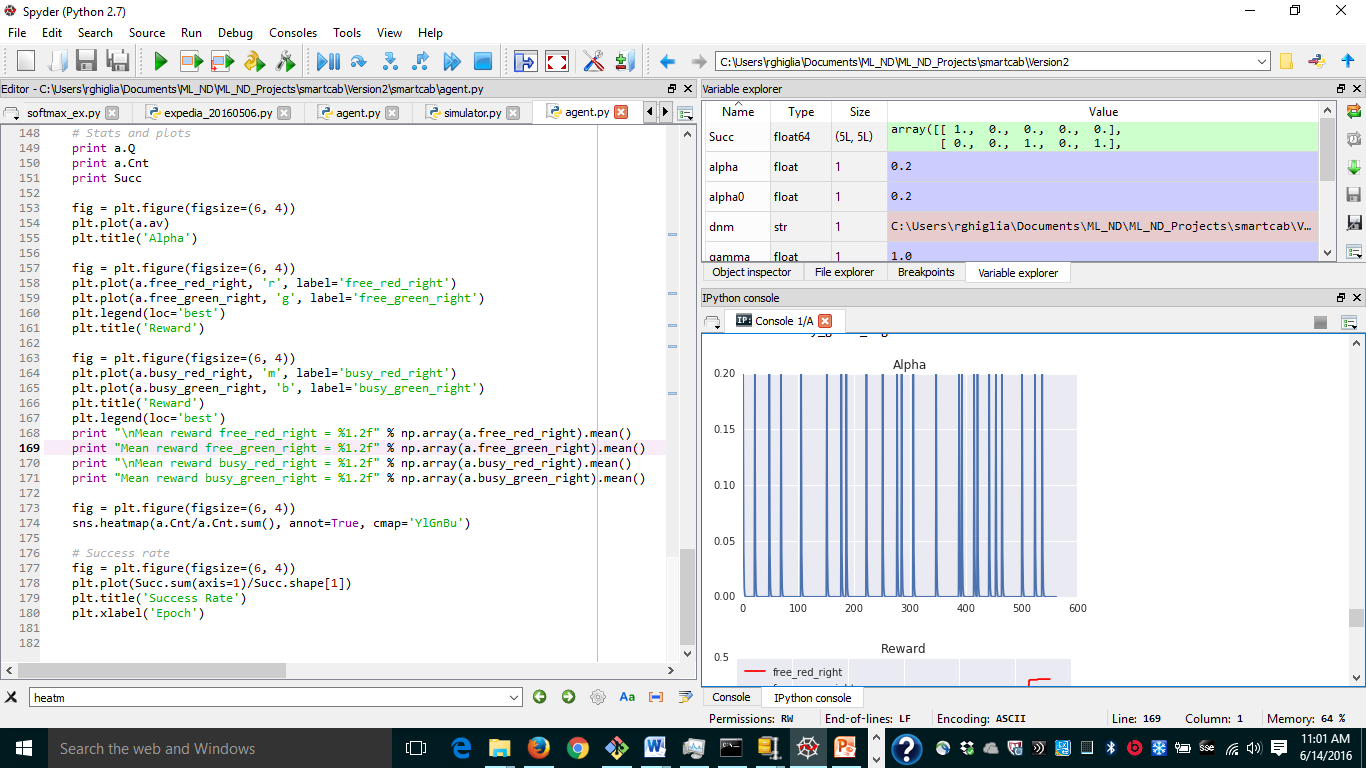
alpha = alpha0

gamma = 1.0

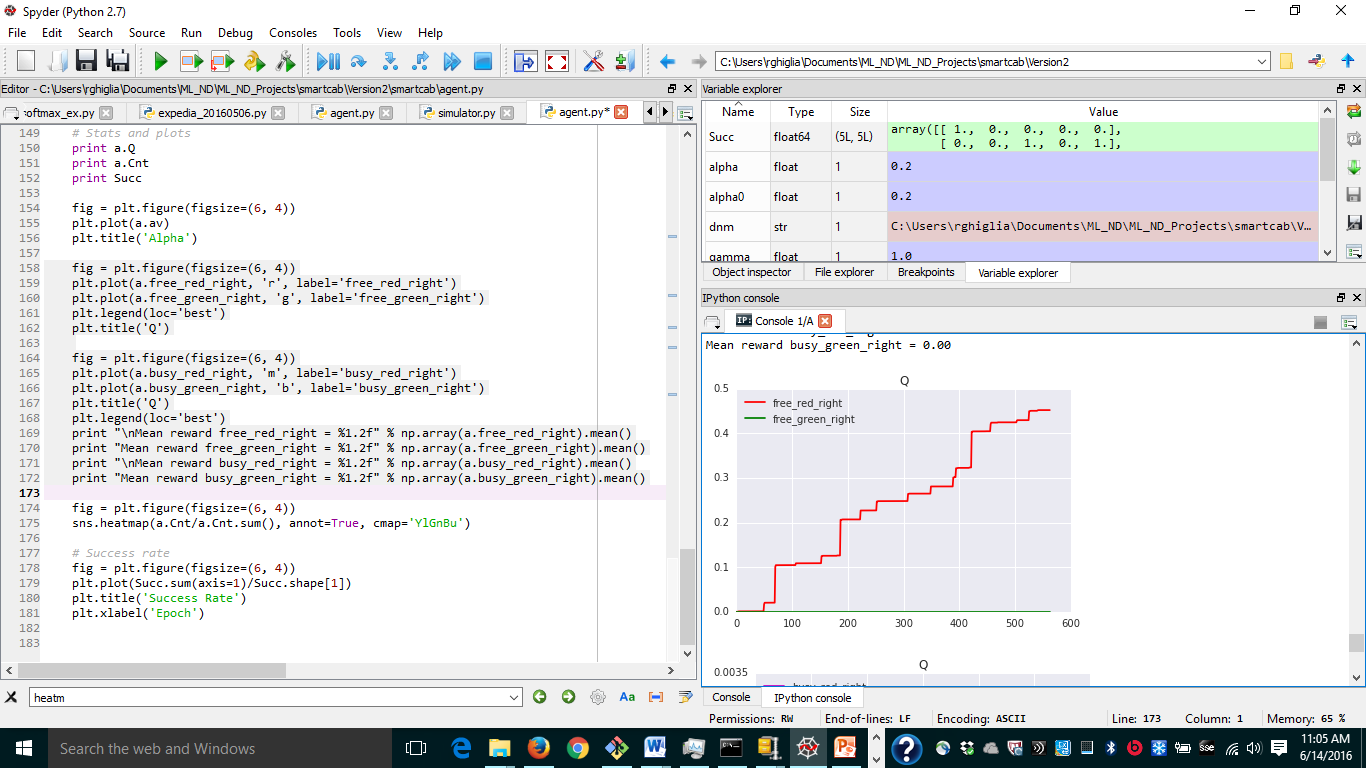
u0 = 0

q0 = 0.0

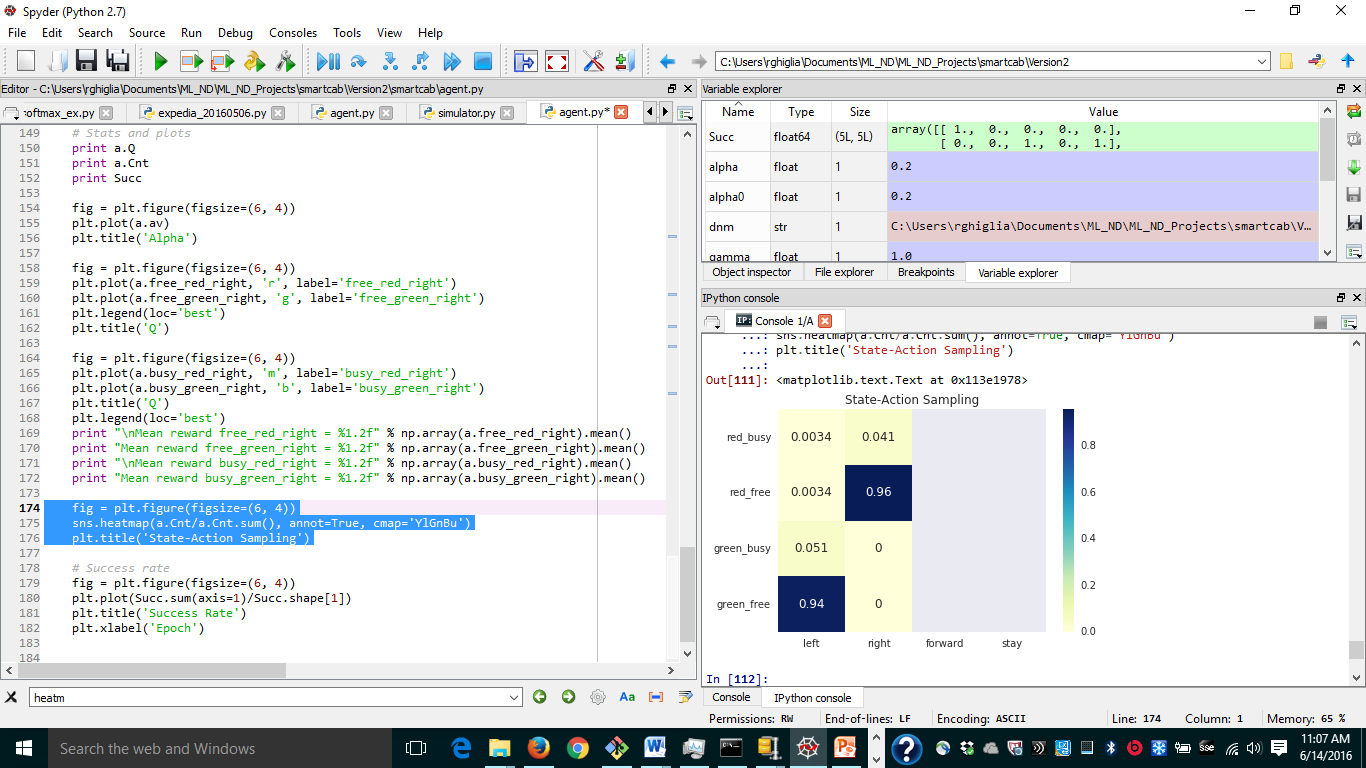
The following figure shows how alpha changes over time. Note that the total time is random since it depends on how long each trial lasts. It might be hard to see but these are actually short spikes indicating that alpha decays to zero really fast and learning stops.



This is also evident from the following figure that shows that the Q function increases for a bit then remains constant.



Finally note that because of the choice of the parameters only a subset of the state-action space is sampled.



We have a few ways to improve on that. Within this parametrization the best way is to introduce some random actions (changing the value of the initial Q did not improve unless we differentiate between state-actions with different initial values).

# Parameter set #1b

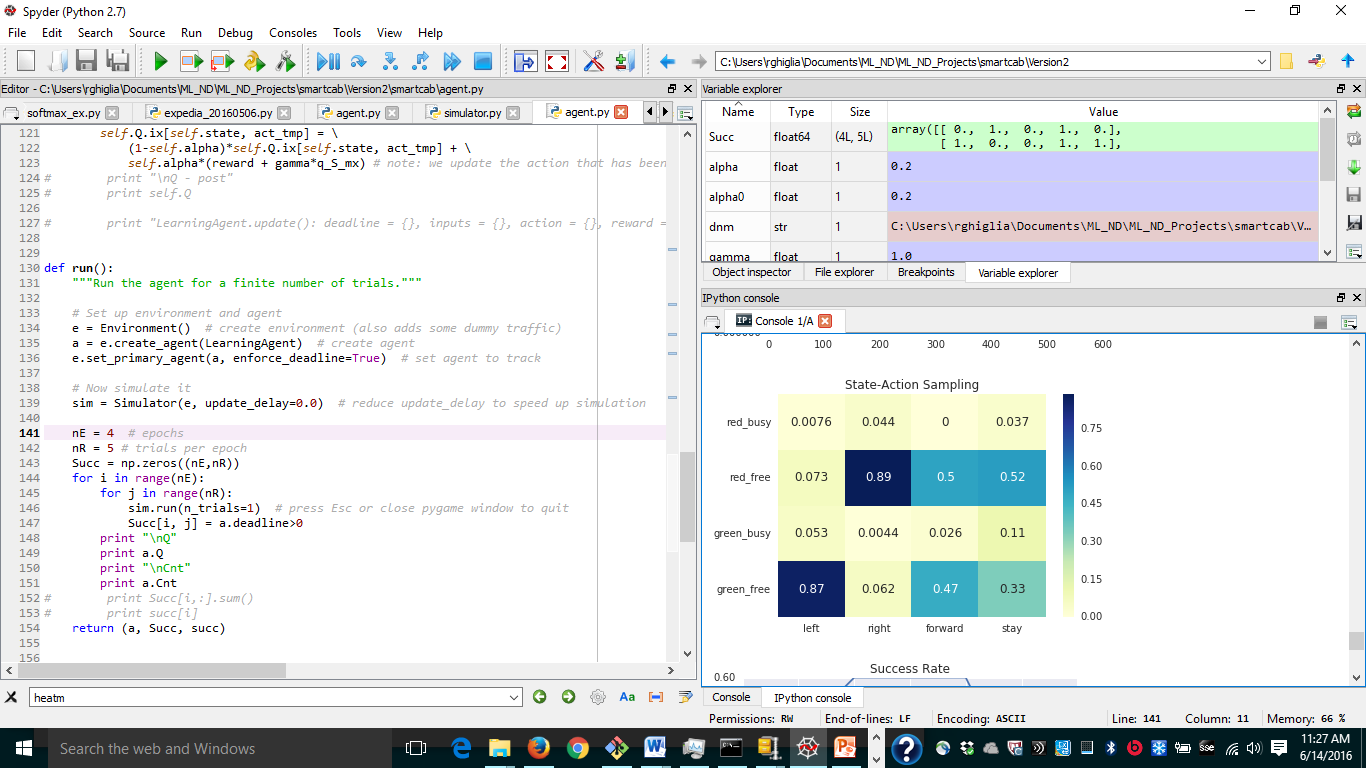
alpha0 = 0.2

alpha = alpha0

gamma = 1.0

u0 = 0.2

q0 = 0



Before fine-tuning further it is worth commenting on γ. As we can see above, when γ is equal to 1 the value of the Q function grows over time. Despite the discussion during the lecture about α needing to decay for convergence to the mean I suspect that there is a restriction on γ itself or possibly an interplay between α and γ. This is shown in the next figure where it seems more clear that the estimate does not grow unbounded (a larger # of epochs would show that, but results not shown here).

# Parameter set #1c

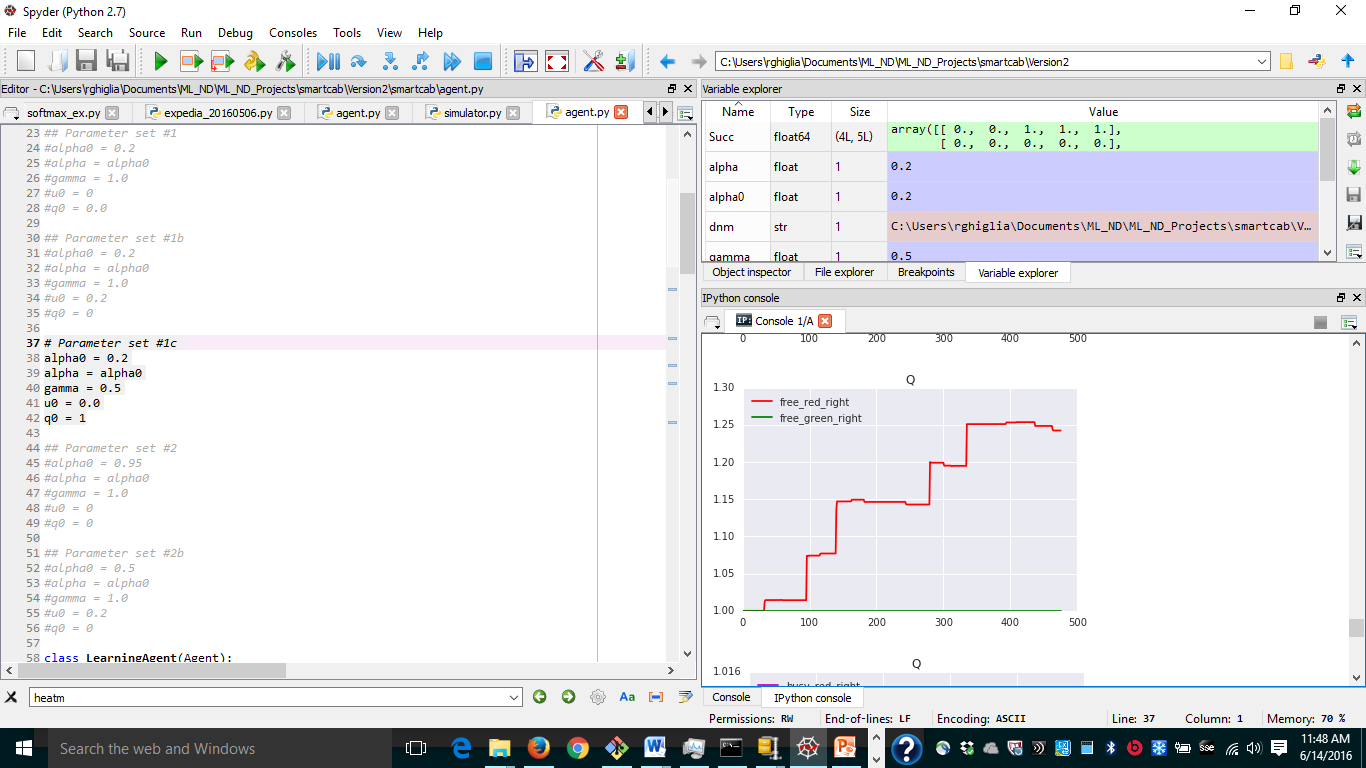
alpha0 = 0.2

alpha = alpha0

gamma = 0.5

u0 = 0.0

q0 = 1



# Appendix

**Implement a basic driving agent**

The driving agent needs to make a decision about where to go. I will consider 4 different ‘strategies’:

* stubborn: agent always makes a left turn
* random: agent takes a random action out of ‘left’, ‘right’, ‘forward’, or ‘stay’ (no action taken)
* guided: agent always follows the planner if light is green
* qlearner: agent uses reinforcement learning (Q-learning) to find the ‘best’ action to take

Figure 1 shows paths followed by an agent following each of the strategy, whereas Figure 2 shows the cumulative reward, respectively. I fixed the code so that they all start from the same location and have the same destination.

Figure 1

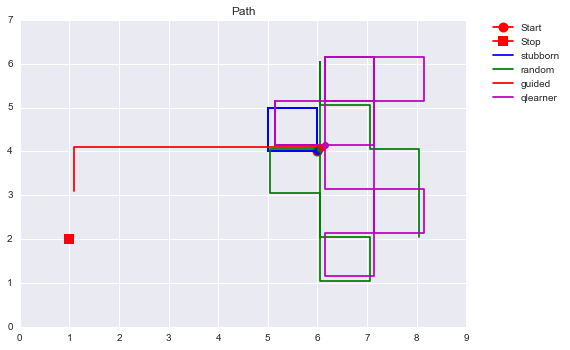
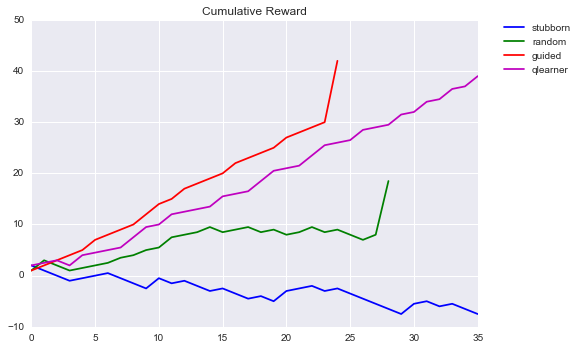


Figure 2



As we can see the stubborn driver keeps looping around without going anywhere. The random driver was actually lucky in this case and ended up at the destination. As a preview, the Q-learner is exploring the space but is not reaching its destination. More comments on this are given below.

For this section I will consider the strategy ‘guided’, which essentially follows the planner:

# Guided

action= None

if inputs['light']=='green':

action = self.next\_waypoint

Running several trials I observe that it normally reaches the destination in time before the deadline.

**Identify and update state**

The state should be a function of what type of inputs we are sensing from the environment. While distance to the destination, etc. could be useful information I assume these remain unknown and we can only sense the traffic light and oncoming traffic. The simplest definition of state (I can think of) is the color of the traffic light. This makes the ‘state space’ limited to 2 states.

One could also add a state that represents the intersection, generically:

* free: when there is no incoming traffic
* busy: if there is any car, left, right, or incoming

This would make the state space be 2-dimensional: red/green x free/busy. If one wanted to use all the available information the state space would be 4-dimensional.

To start, I picked the simplest description of the environment: red or green. This will help building intuition of the algorithm.

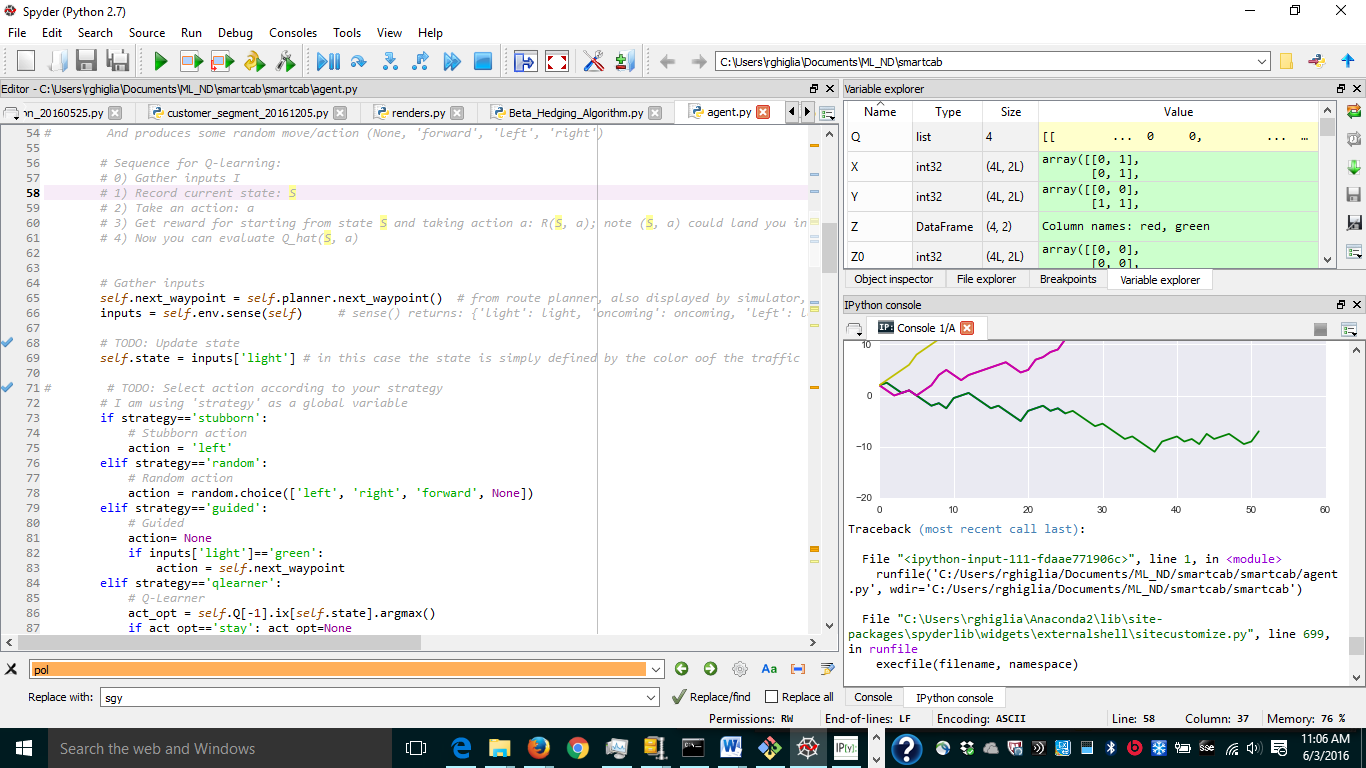
The questions asks about the ‘agent and its environment’; I not sure if I should interpret it as those two in isolation or in combination. The explanation below tries to address both interpretations.

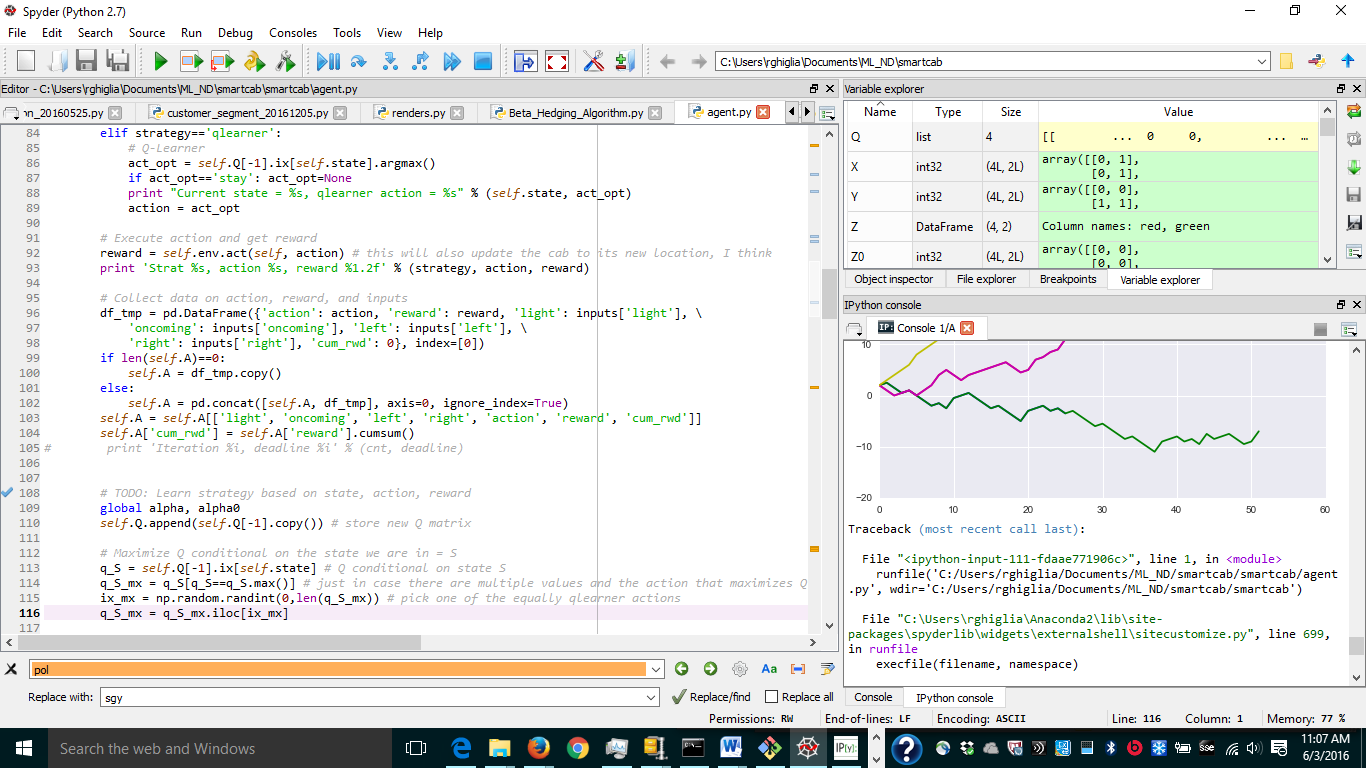
These inputs do not describe the agent though. It seems to me none of the available inputs are descriptive of the agent himself. Examples of that would be the distance to the target. Although, strictly speaking, the latter would also not be describing the agent independently but rather the relation of the agent to the environment. A state that would be strictly relative to the agent could be something like the agent’s health affecting his ability to drive or something.

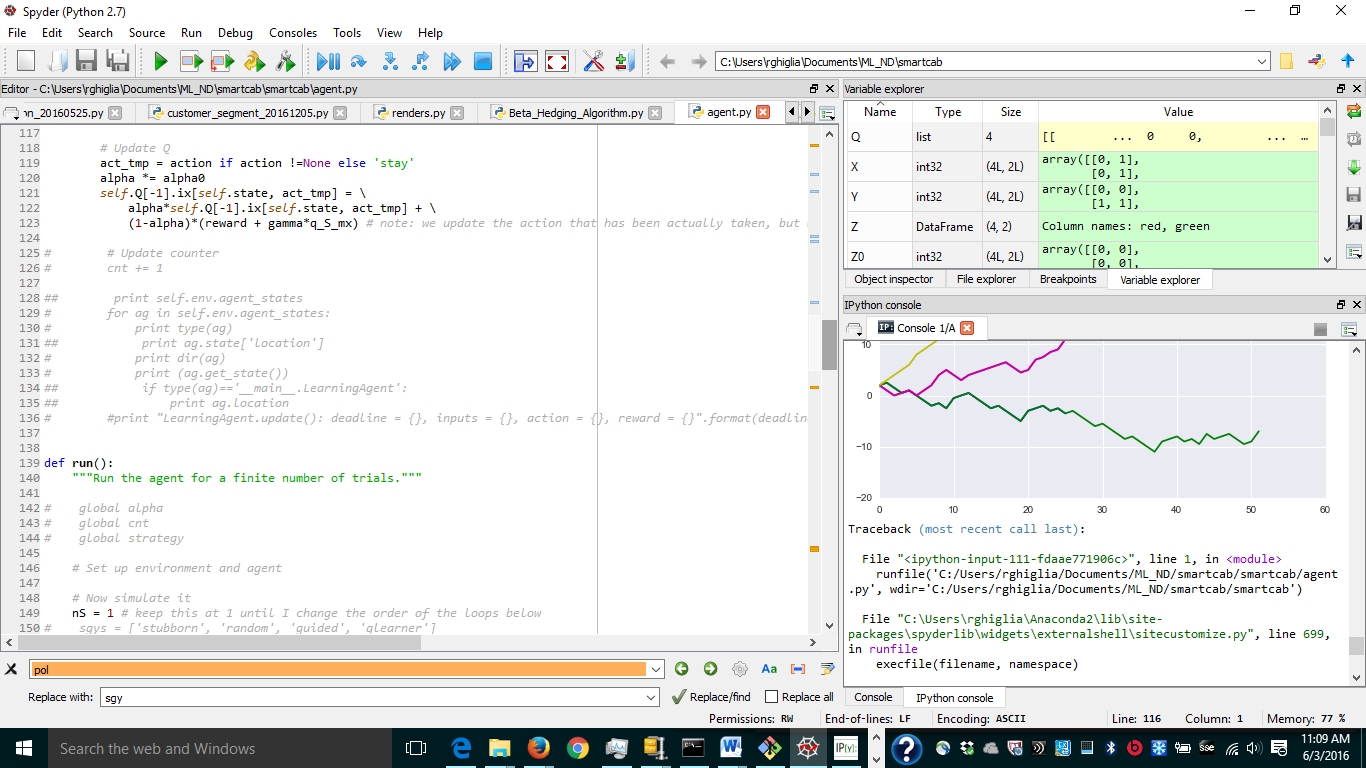
**Implementing Q-Learning**

The ‘guided’ strategy in sense is a bit like cheating because the planner has likely more global (i.e. extended) information than the agent and can therefore be more effective. The ‘Q-learning’ strategy estimates an optimal combination of state and action by trial and error. It takes actions, gets a reward and incorporates (updates) that information into an estimate of how best to proceed.

Figure 3







As such, the agent wanders around a lot more and the success rate is a lot lower than the ‘guided’ strategy. It is hard to describe the actual behavioral changes without collecting more data. From visual inspection, it seems that in the ‘guided’ strategy the agent tries to go on more straight-line paths.

I ran 5 simulations and kept track of the paths for each strategy (Figure 4, start = black circle, destination= black square)[[1]](#footnote-1). The results confirm that observation. The guided strategy gets there in a most linear way, whereas the q-learner wanders around a lot more. Occasionally, the Q-learner does better than the guided driver, however that is probably mostly by chance[[2]](#footnote-2). The purely random driver never reached the destination.

Figure 4

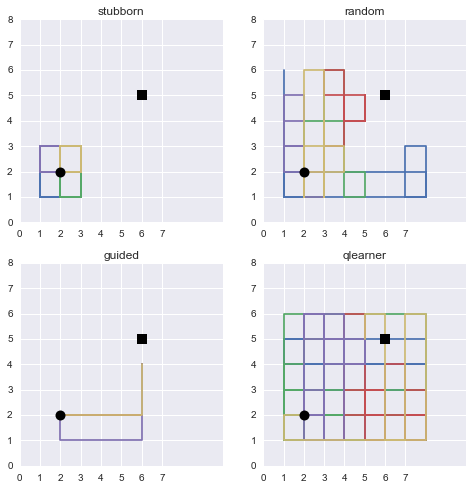
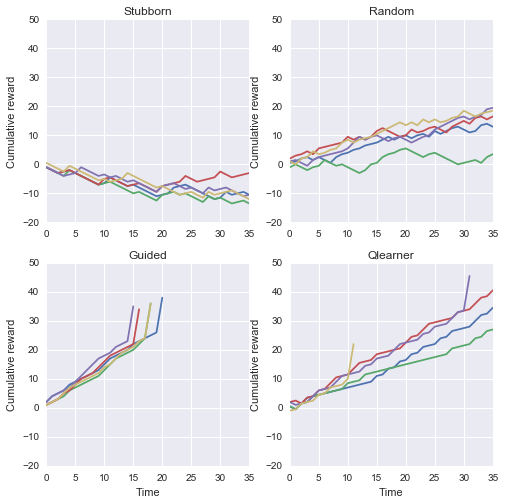


Figure 5



In this small sample the success rate is 100% for the guided learner, 40% for the Q-learner and 0 for the others.

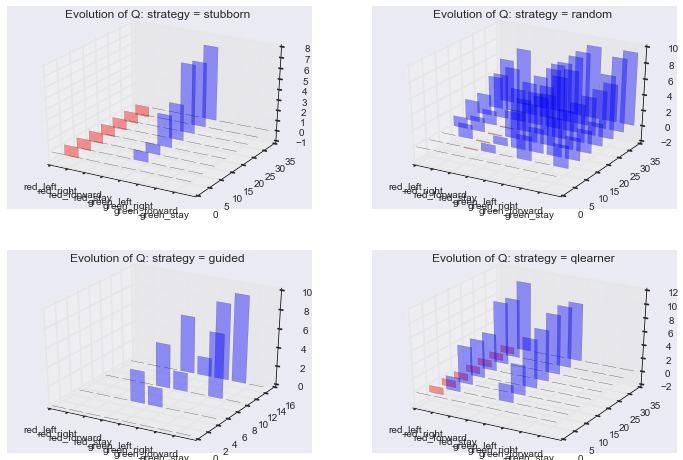
**Enhance the driving agent**

We can visualize the estimated value function Q for all strategies as shown in Figure 6. Q is really a function of 3 variables: state, action, and the updating step. I flattened the state-action variables so that we can see all 8 components of the Q matrix and their evolution over time.

As we might expect, the stubborn strategy only visits 2 state-action combinations: red-left and green-left. The red-left is consistently a bad choice and is maxed by other alternatives that have at least the value of zero as Q was initialized to 0.

The random strategy visits a lot of state-action combinations, whereas the guided and q-learning strategies only a few. However, it is noticeable that q-learning strategy switches from preferring green\_left to red\_right a couple of times.

Figure 6



**Enhance the driving agent**

In this section I will focus on the q-learning agent. As a test case, I ran 100 simulations and obtained the following success rate:

guided 0.99

qlearner 0.26

Alpha is set to 0.5 and decays at 0.5^t.

self.Q[-1].ix[self.state, act\_tmp] = \

alpha\*self.Q[-1].ix[self.state, act\_tmp] + \

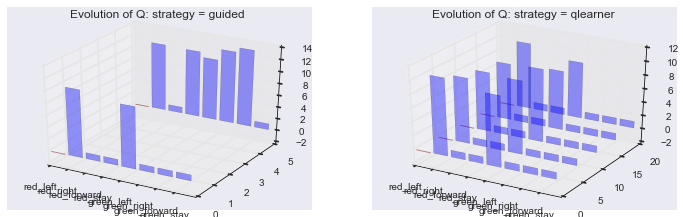
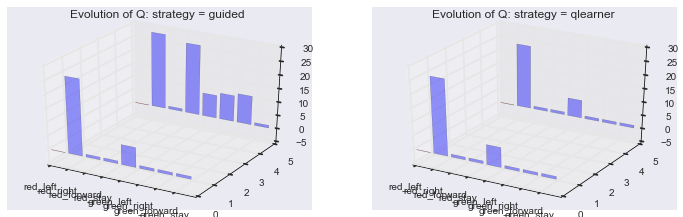
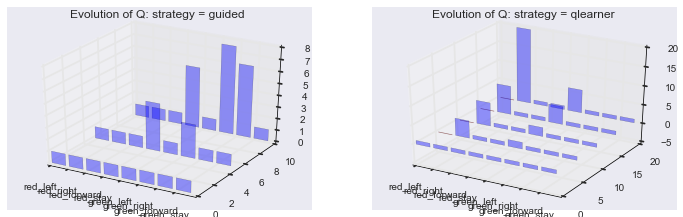
(1-alpha)\*(reward + gamma\*q\_S\_mx)

A larger alpha creates more inertia, i.e. the function Q incorporates new information at a slower rate. Gamma controls the contribution of the future expected utility compared to the immediate reward and is set to 0.9. Q has been initialized to zero.

The first improvement is to initialize Q to a larger number so that more of the space gets sampled. With Q = 10, the success rate went down to 0.02, whereas with Q=1 it went up to 0.7.

I then implemented a cross-trip learning whereby the best estimate of Q is transferred from one trip to the next, i.e. the Q function that is learned during a trip becomes the starting estimate of the Q function at the start of the next iteration. Figure 7 shows the evolution of the Q function estimate in the first trip (top left panel), the second (top right panel) and the last trip (bottom left panel). The success rate was 0.5.

Figure 7



As far as reaching the destination in the minimum time, I collected some statistics that measure the efficiency of reaching the destination. An efficiency of 1 would indicate that the target has been reached in 1 step, in general efficiency = 1 – time spent / deadline

Efficiency:

guided qlearner

0 0.520000 0.24

1 0.600000 0.36

2 0.700000 0.75

3 0.550000 0.00

4 0.657143 0.00

5 0.133333 0.90

6 0.520000 0.68

7 0.800000 0.00

8 0.650000 0.00

9 0.750000 0.00

The table shows that at times the q-learner reaches the destination faster than the guided solution. It does not reach the destination all the time though.

The following table shows all the actions and rewards taken during the last iteration and it shows that no action was taken that resulted in a negative reward.

# light oncoming left right action reward cum\_rwd

0 red None None None right 0.5 0.5

1 green None None None left 2.0 2.5

2 red None None None right 0.5 3.0

3 red None None None right 0.5 3.5

4 green None None None left 0.5 4.0

5 red None None None right 0.5 4.5

6 green None None None left 0.5 5.0

7 green None None None left 0.5 5.5

8 red None None None right 0.5 6.0

9 green None None None left 2.0 8.0

10 green None None None left 0.5 8.5

11 green None None None left 0.5 9.0

12 green None None left left 0.5 9.5

13 red None None None right 0.5 10.0

14 green None None None left 0.5 10.5

15 red None None None right 2.0 12.5

16 red None None None right 0.5 13.0

17 red None None None right 0.5 13.5

18 red None None None right 2.0 15.5

19 green None None None left 0.5 16.0

20 red None None None right 0.5 16.5

1. Note that the paths do not show the very last step because the code exists before recording the very last step. [↑](#footnote-ref-1)
2. One can identify reaching the target by the big jump in the reward function. [↑](#footnote-ref-2)