Assignment 2 - Temporal Difference Learning

COMP 596 - Brain-Inspired Artificial Intelligence

LE, Nhat Hung
McGill ID: 260793376
Date: April 15, 2020
Due date: April 22, 2020

Prof. Blake Richards
Winter 2020

Part 1 - Understanding the Foster, Morris and Dayan paper (30 marks total)

Question 1 (5 marks)

(a) In your own words, describe the distal reward problem. (2 marks)

Only place cells representing locations near a goal would fire as a reward. Therefore, if a rat starts out too far from hidden food, it wouldn't know where to go – no place cells would activate, so any reward guiding that rat's actions would be too far away, hence "distal".

(b) What type of navigation learning model suffers from the distal reward problem, and why is it a problem for them? (3 marks)

Models that assume that place cells provide the ideal representation for reward-based learning, i.e. place fields evenly span across the environment, suffer from the distal reward problem. This is because most locations within an environment are often very far from the goal location, compared to the breadth of a place field. In such an environment, the agent would have no direct information as to the direction in which to move if it starts at a location far from the goal where none of the place cells associated with appropriate actions is active.

Question 2 (5 marks)

(a) In your own words, describe the global consistency problem. (2 marks)

A coordinate system to aid navigation would rely on self-motion information. However, this information is egocentric, relative to the agent. Therefore, if the agent is tested with different starting locations on each trial, it will acquire inconsistent coordinates over the environment as a whole.

(b) What type of navigation learning model suffers from the global consistency problem, and why is it a problem for them? (3 marks)

Models that assume place cells become associated with metric coordinates for locations within the environment suffers from global consistency. This is because of their dependence on local, egocentric self-motion information to learn these coordinates. For each laboratory task trial, an animal is always picked up from the goal location and moved to the starting location, and this starting location can sometimes be completely novel. If the animal path-integrates from each new starting position, it will learn inconsistent coordinates.

Question 3 (10 marks)

(a) What is the role of the place cells in the Foster et al. (2000) model? Specifically, what do they encode? (2 marks)

Place cells map to place fields which are evenly distributed in the environment, thereby encoding the position of the agent. They serve to learn the correct reward expectation and action for each location of the environment.

(b) What sorts of information can place cells not encode that a navigation system might want? (2 marks)

By itself, a place cell merely fires when the agent is located within its place field. Therefore, place cells cannot inherently encode the action the agent ought to take to reach its goal.

(c) What is the role of the critic in the Foster et al. (2000) model? Specifically, what is the critic learning to encode? (2 marks)

Consider a value function over location, *V(p)*, an ideal evaluation of the actions specified by the actor at location *p*. The critic learns to encode *V(p)*, and as learning progresses, the critic's output function *C(p)* converges to *V(p)*.

(d) What is the role of the actor in the Foster et al. (2000) model? Specifically, what is the actor learning to encode? (2 marks)

The actor learns a policy function to output the optimal action at each location *p*, specifically, a probability distribution over all actions for each location in the environment.

(e) What is the error signal used in TD learning? Give the equation for it, and explain all the terms/variables contained in it.

The error signal used in TD learning is the *prediction error*, computed as follows

$$\begin{gathered} \delta_t = R_t + \gamma C(p_{t+1}) - C(p_t) \\ \delta_t \text{ prediction error} \\ R_t \text{ reward at time } t \\ \gamma \text{ discount factor} \\ C(p_{t+1}) \text{ critic output at } t+1 \\ C(p_t) \text{ critic output at } t \end{gathered}$$

As we want *C(p)* to converge to *V(p)*, we want to minimize the prediction error, effectively enforcing this equality

$$\begin{gathered} V(p_t) = \langle R_t \rangle + \gamma V(p_{t+1}) \\ \langle \cdot \rangle \text{ mean over all trials} \end{gathered}$$

which is the Bellman derivation of the *V(p)* equation.

Question 4 (10 marks)

(a) Does TD learning with an actor-critic system solve the distal reward problem? Explain why or why not. (5 marks)

Conventional error-driven learning rules only adjust their value estimates once the outcome is known, which is challenging due to the distal reward problem. However, though the reward information is only available at the goal, note that the expected reward should increase smoothly along a path to the goal. Therefore, there exists a temporal difference between reward expectations at successive locations. TD learning uses this temporal difference to learn while navigating the environment. Using TD learning, the actor-critic system can, therefore, learn reward expectation and action policy on-the-go, effectively sidestepping the distal reward problem.

(b) According to Foster et al. (2000), what information must be available to a rat if it is going to solve the global consistency problem with TD learning? Do you think that it is realistic that a rat would possess this information? (5 marks)

According to Foster et al. (2000), in order to develop a general coordinate system, a rat would require self-motion estimates of its current position and memory of the goal coordinate. Rats very likely do possess this information as they are able to not only learn changing platform locations on successive days, effectively avoiding interference, but also generalize from experience on early days to improve performance on later days. This is shown in Figure 1 of the paper, and demonstrates their ability to learn consistent global coordinates.

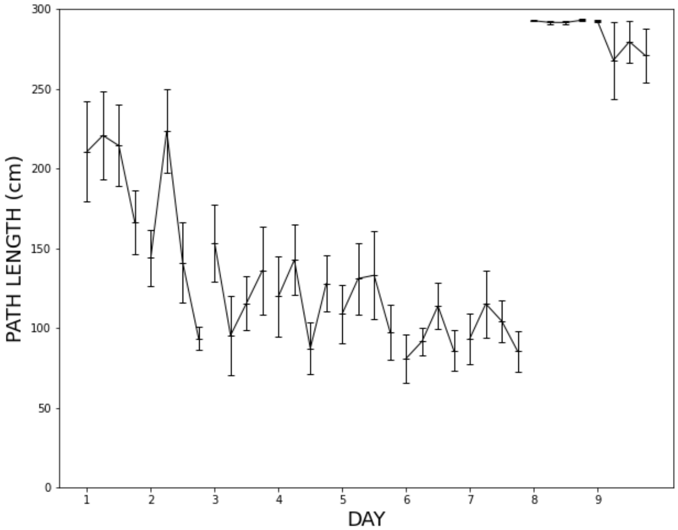
Part 2 - Implementing the pure TD algorithm (20 marks total)

Question 2 (5 marks)

With a single platform location, does the model successfully reduce the latency over trials? Get your code to generate the plots, and also include them in your pdf file. (Note: you should have a flag in your code that lets the TA set it to a single-platform or multi-platform case.)

For all plots below, each data point is the mean path length from 10 simulated episodes, each error bar representing the standard error.

The pure TD learning actor-critic model successfully reduces latency, measured by path length, across 7 "days" of simulation with a single platform location. However, on day 8, the platform is moved to the opposite quadrant, resulting in a sudden spike in latency. This data is congruent with Figure 4.a. of the paper.

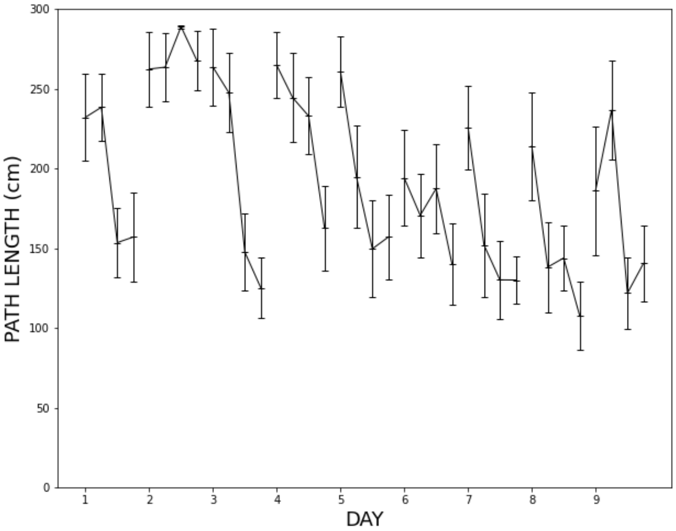


Question 3 (5 marks)

What happens to the performance of the model when the flag is toggled to a multi-platform case (provide plots of the escape latency)? With reference to your answers in Part 1, explain why this happened?

When the platform occupies a novel location every day, the model first captures acquisition on day 1, similar to the single-platform performance, but fails to generalize to new goal locations. Overall high latency, especially on day 2, shows that the model also suffers from interference from the previous days' goal locations.

The failure to generalize is due to the model's lack of a learned coordinate system. This system will be integrated for part 3.

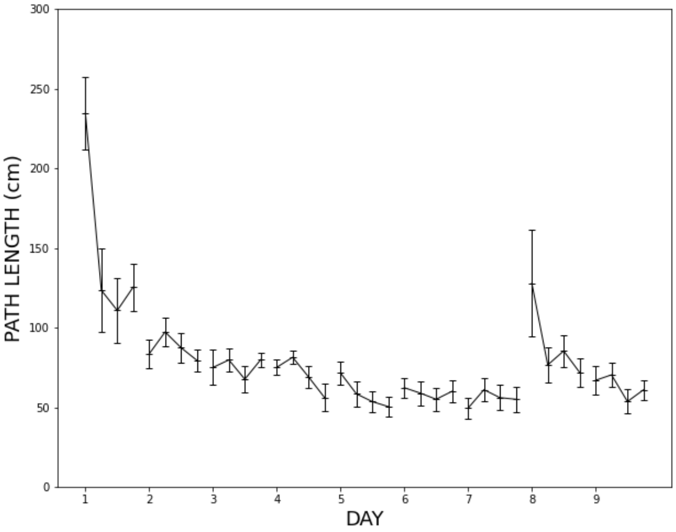


Part 3 - Implementing the combined coordinate and TD algorithm (20 marks total)

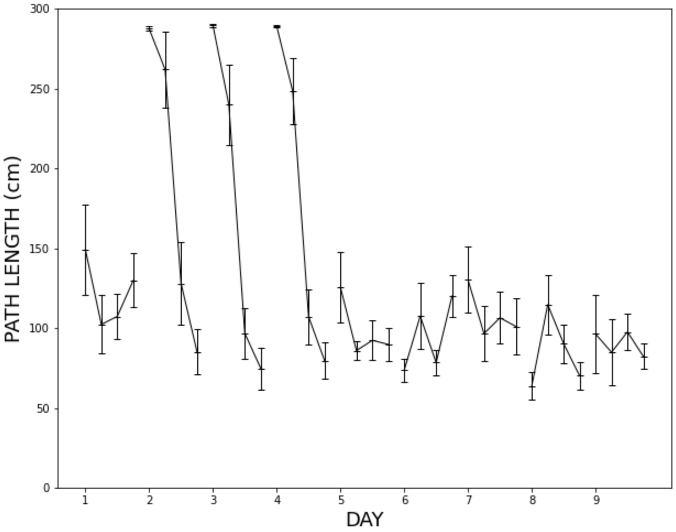
Question 2 (5 marks)

With multiple platform location, does the model successfully reduce the latency over trials? Get your code to generate the plots, and also include them in your pdf file. (Note: you should have a flag in your code that lets the TA set it to a single-platform or multi-platform case.)

The first plot shows the model not only reduces latency in the single-platform task for the first 7 days, but also successfully generalizes for when the platform is moved on day 8 - a significant improvement from the pure actor-critic model.



This generalization is clearly demonstrated in the multiplatform task. The improvement within each day is first gradual, but then becomes immediate starting day 6.



Question 3 (5 marks)

What is your opinion of the coordinate model? Does it seem like a good strategy for an AI? Why or why not? Justify your answer.

In the water maze task, the coordinate model was the missing link to build a goal-independent representation of the environment, enabling navigation to arbitrary goals. It tackled the problem of global consistency, one that affects all navigating systems which use self-motion information to build map-like representations. The basis of the model is the phenomenon of dead reckoning, exhibited by many animals, which suggests that a neural coordinate representation does exist. The coordinate model is therefore probably a good AI strategy, especially in the brain-inspired domain.

However, the model is unlikely to be a panacea to navigation tasks due to its simplicity. It depends solely on metric information but remains mostly unaware of the environment's topological structure. Its reliance on a remembered goal coordinate might, therefore, prove ineffective when obstacles e.g. barriers are introduced, or when the environment is 3-dimensional and topologically rich e.g. tunnels, shortcuts. These cases would present many choices of paths to the goal, among which metric navigation might have difficulty taking the optimal action. Indeed, we are not yet sure if animals like rats use topologically richer representations of space to aid navigation.

The coordinate model is, therefore, not the best AI strategy for navigation. However, it can be a step in the right direction, a lower level fallback strategy to a more sophisticated model that is aware of topological information, much like how pure actor-critic was a lower level strategy before the coordinate model was integrated. Until then, the coordinate model is a suitable strategy for navigation tasks with a simple environment e.g. 2-dimensional, minimal to no obstacles.