COMP 596: Programming Assignment 2

Hong, Joon Hwan
260832806
Winter 2021

Part 1 – Understanding the Foster, Morris and Dayan Paper

Question 1.1

a) The global consistency problem is the issue when learning the global environment (coordinates) given that only relative motion of oneself is available. The issue arises when the initial position is not consistent – a different starting coordinate position per experiment/attempt of learning – leading to inconsistencies.
b) Navigation learning models that assume place cells become associated with metric coordinates for positions in the environment suffers from the global consistency problem; the locations are learned from an agent's motion, which will only give relative information. This is an issue given that the starting position is changed over different trials, integration over self motion from new locations leads to the model learning inconsistent coordinates over the global environment.

Question 1.2

a) The distal reward problem is the issue where the place cells representing locations near an objective would activate, resulting in a situation where no place cells activate (or very little) if the model initializes far from the objective; there is no direct information as for where to move.
b) Navigation learning models that assume that place cells are the ideal method/ representation for a reward-based learning environment. This issue occurs as only the place cells in close proximity to the objective would be associated with the motion/direction model should take. Place cells too far from the goal never obtains a learning signal; thus, no direction can be decided due to no information from the place cell(s).

Question 1.3

a) The actor learns the optimal action for the agent throughout the environment. The actor encodes how frequently each action should be taken at positions in the space, given a probability distribution over all actions (the direction j to take at location p).
b) The critic learns the optimal value function over location V(p), which is the evaluation – the expected discounted future reward the agent will receive – of the actions by the actor at location p. The critic learns to encode V(p).
c) The TD error (δ_t) is calculated as the following: $\delta_t = R_t + \gamma C(p_{t+1}) - C(p_t)$
    a. Where: R_t is the agent's reward at time t; γ is the discount factor, in the range of [0,1]; $C(p_t)$ is the critic output at time t.

d) For the place cells' role: they are mapped to place fields in the model. The place cells encode the location as its Gaussian activation is dependent on the location *p* within the environment.
e) The place cells just encode for the location of the agent as they fire when the agent is located within their corresponding place fields. A navigation system might want additional navigational information such as the distance or direction from the goal.

Question 1.4

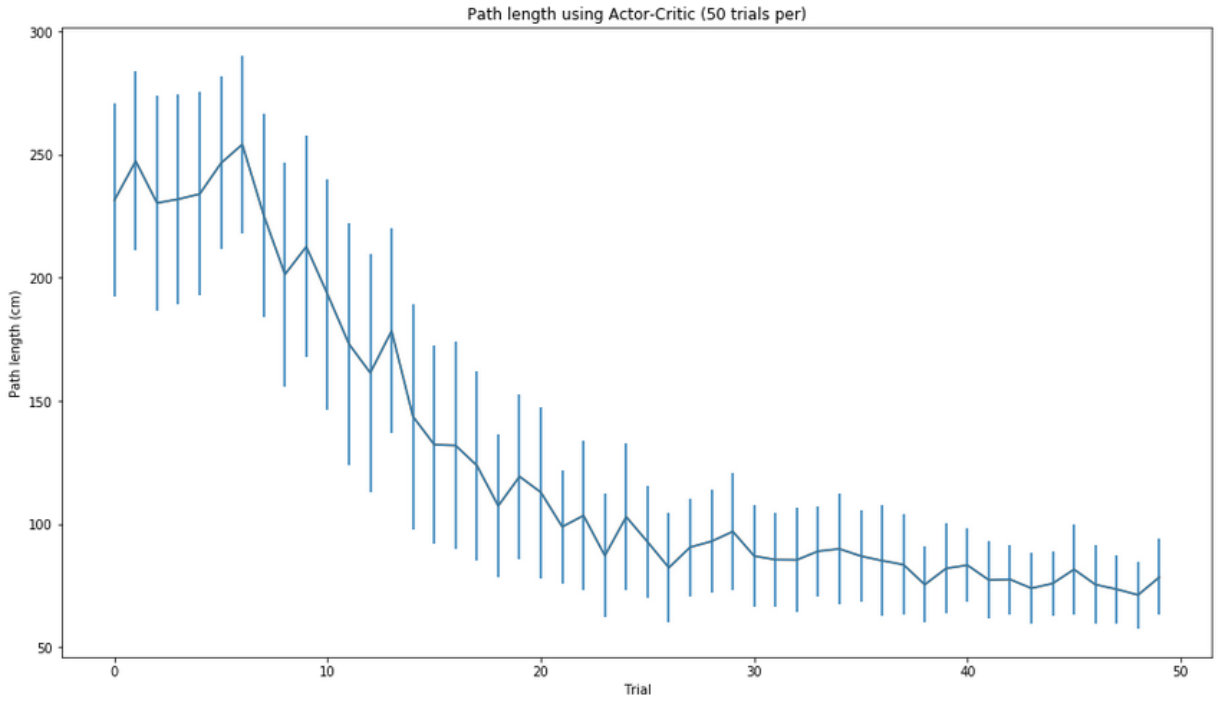
a) Yes, the TD learning with the actor-critic system does solve the distal reward problem mentioned above. The critic values coordinates closer to the goal more than the coordinates farther away, resulting in non-zero temporal difference between reward expectations at consecutive locations. This enables learning at every timestep instead of only when near the goal, the case in the distal reward problem.
    a. In the end, as *C(p)* is bound to converge to the value function *V(p)* according to the paper, meaning the temporal difference error allows the actor to successfully reinforce positive actions (moving closer to the goal) while inhibiting negative actions (moving away from the goal) in any location *p*.
b) The authors discuss that in order for the animal to deal with the global consistency problem with TD learning, it needs to obtain estimates of its self-motion and its current position and memory/information of the objective location. This is to obtain an allocentric environment representation independent of the origin position.
    a. A rat would possess the necessary information: shown in *Figure 1*, rats are able to learn the changing platform locations on different days and generalize their experience from previous days to improve later trials. Thus, the paper illustrates the rat's ability to learn global coordinates.

Part 2 – Implementing the pure TD algorithm

Question 2.2

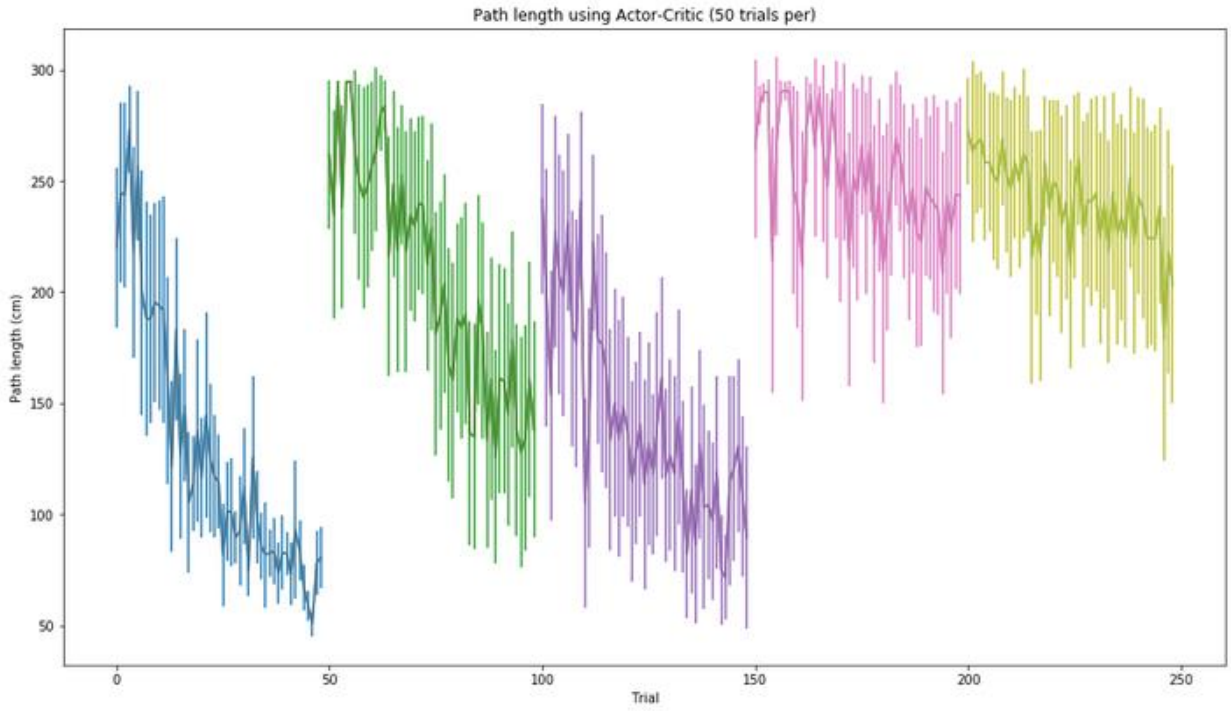
With a single platform location, the model successfully reduces the latency over trials. The code is set to just have 5 arbitrary platform coordinates for debugging purposes, but the option for random integer coordinates is available for use after de-commenting it. The result from a single platform coordinate using the actor-critic model, a figure is shown below. The navigation model successfully reduces the latency, measured in proxy by the path length over trials.

No extensive hyperparameter search was conducted; the current values were arbitrarily chosen and they appeared to work. The figure was generated with the following values averaged over 50 runs with: eta_actor=0.1; eta_critic=0.01; discount_factor=0.9; arbitrary platform coordinates=[15,15]. Standard deviations are represented as error bars in the plot.



Question 2.3

With the multi-platform setting/flag, the results are as follows:



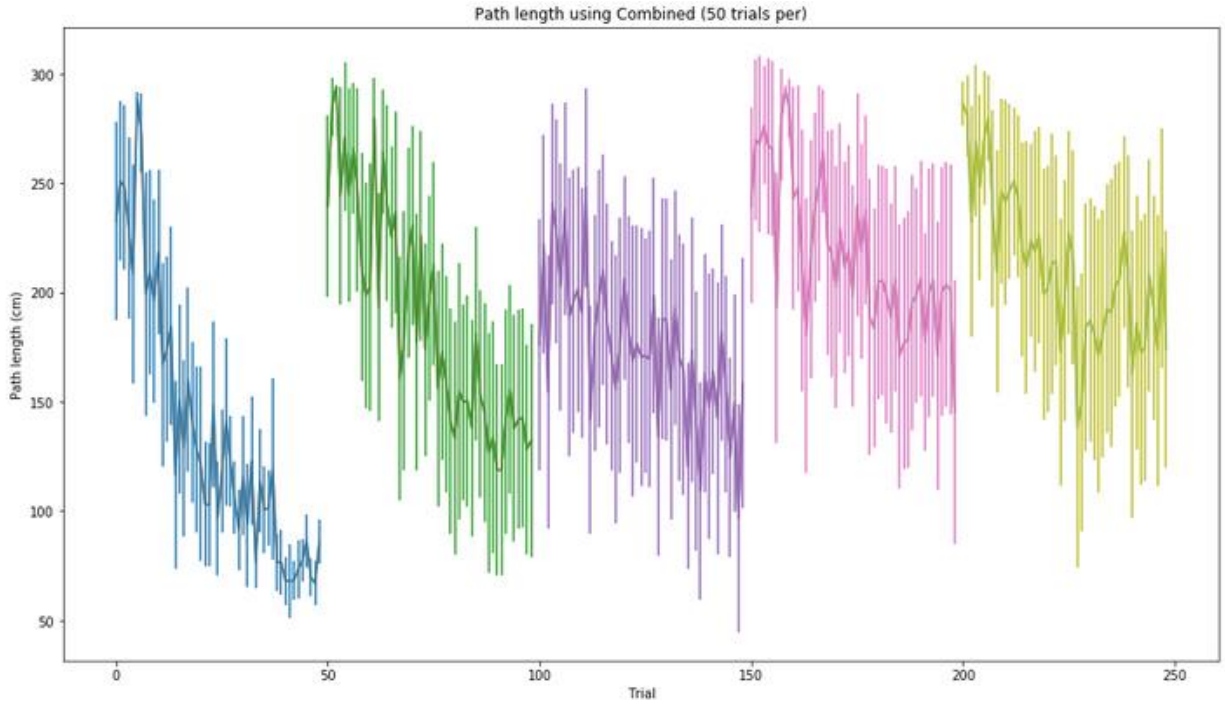
Evidently, the first, second, and third cases are able to reduce the distance to a sufficient level, while the fourth and fifth platform locations fail to reduce their path length in a meaningful manner given 50 trials. The results will vary each time the simulation is ran.

The simulation was conducted with the same parameters as the single platform, with the addition of new fixed platform locations for the purpose of this exploration: [15,15], [-15,-15], [30,10], [10, -30], [-40,20].

Given the two plots/results above, the TD learning model is able to solve the distal reward problem. But, when the platform is changed (the goal coordinates change) too drastically, the model initially struggles to learn, seen with the slow rate of decrease in the second and third platform durations from trials 150-250. This should be due to the fact that the previously learned value function and policy function now is actually hindering the algorithm from acclimatizing to the new goal coordinates. The issue would be theoretically alleviated by the combined coordinate and TD algorithm, where the coordinate model tries to learn globally consistent coordinate approximations using place cell activation functions; it should not be as strongly affected by the sudden change in goal location as it learns the global positions of the simulation.

Part 3 – Implementing the combined coordinate and TD algorithm

Question 3.2



The combined coordinate and TD algorithm is able to visibly decrease the path length, and thus the escape latency for all platform durations, unlike the multi-platform example seen in Part 2. This is seen in the fourth and fifth platforms, where previously the model was unable to learn and effectively decrease the path length, while now it is able to learn more effectively. The specific figure above was obtained with a changed parameter of discount_factor=0.8; other parameters remained the same from Part 2.

Question 3.3

In this specific circumstance, of a 2D discrete coordinate system, the combined model did resolve the issues mentioned in Part 2.3, as it theoretically learns a goal-independent representation of the simulation/environment. Thus, the combined model should resolve the global consistency & distal reward problem that was discussed prior in the assignment report.

In addition, the paper has shown that it was biologically feasible and the type of learning is exhibited in many animals, showing validity in particular as a "brain-inspired" AI. However, this is given the assumption of a "flat" 2D discrete coordinate system without consideration of real-life situations & circumstances. For instance, the model has no information of the environment's topological features, such as obstacles, barriers, and other potential hazards that should be considered for a navigational AI in practice (would a navigation system tell one to go over a mountain because it is less distance in the X-dimension compared to going around? Etc.). In these circumstances, strictly reducing the pathlength/objective may not be the most ideal.

Consequently, the combined coordinate TD model is not the best AI strategy for navigation. There are far too many abstractions required to utilize this beyond a flat 2-dimensional environment. A model that can integrate more topological information would be preferred.