Advanced Computational Neuroscience - Week 5

—

Reinforcement Learning

Sam Suidman

April 13, 2022

1 Introduction

Finding your way in an environment you have never been before is hard, because you do not have any orientation. It seems to be that place cells in the Hippocampus play a crucial role in this. The model in the '*Foster et al.*' paper uses these place cells together with a critic and an actor to improve the orientation of a rodent inside a water maze. The critic criticizes if actions that the rodent takes are correct. The actor takes actions based on the information from the critic. This is an example of reinforcement learning. The water maze is in fact a round barrel filled with water. In it is a small platform, which the rodent can not see, but need to find.

2 Theory

2.1 Place cells

The model is in the first place based on the Hippocampus place cells. The firing rate of a place cell is modelled as a 2D Gaussian function with the maximum firing rate at the place where the place cell is. In equation 1 the firing rate at some point $\vec{p}$ caused by a place cell i with coordinates $\vec{s_i}$ is shown. The width of the Gaussian is given by σ.

$$f_i(\vec{p}) = e^{-\frac{\|\vec{p}-\vec{s_i}\|}{2\sigma^2}} \tag{1}$$

2.2 Critic

The critic is the weighted sum of the firing rates of all place cells. It is giving by equation 2 for some point $\vec{p}$.

$$C(\vec{p}) = \sum_{i=1}^{N} w_i f_i(\vec{p}) \tag{2}$$

In the beginning the weights w_i are initialized randomly and therefore $C(\vec{p})$ is not correct at all. The goal for the critic is to find the correct values of w_i such that if $C(\vec{p})$ has a high value the rodent is in a good position to find the platform. In the paper is described how the values of w_i can be optimized. $C(\vec{p})$ should reflect a reward function $V(\vec{p})$ that depends on the current reward R_t and future rewards $R_{t+1}, R_{t+2}, ...$ and a factor $0 < \gamma < 1$ as given in equation 3. The reward R_t is 1 if the rodent is on the platform en 0 otherwise. If the rodent is on the platform (and $R_t = 1$) then $V(p_{t+1}) \approx C(p_{t+1}) = 0$.

$$V(\vec{p_t}) = \langle R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + ... \rangle \tag{3}$$

Which can be rewritten as equation 4.

$$V(\vec{p_t}) = \langle R_t \rangle + \gamma V(\vec{p}_{t+1}) \tag{4}$$

In the end the critic $C(\vec{p})$ should be the same as the reward function $V(\vec{p})$ such that equation 5 holds.

$$C(\vec{p}_t) = \langle R_t \rangle + \gamma C(\vec{p}_{t+1}) \tag{5}$$

The value δ_t is the prediction error and can be used to update w_i. It is defined in equation 6, where $\langle R_t \rangle$ is not available in real life and replaced by R_t.

$$\delta_t = R_t + \gamma C(p_{t+1}) - C(p_t) \tag{6}$$

Now w_i is updated via equation 7, where η_w is the learning rate.

$$\Delta w_i = \eta_w \delta_t f_i(p_t) \tag{7}$$

2.3 Actor

The actor in this model is a construction of 8 cells that each represent a direction: N (North), NE (North East), E, SE, S, SW, W, NW. When the cell SE is the only one active, the rodent moves to this direction, hence down and to the right. The actor cells are the weighted sum of the firing rates just as for the critic, only now with 8 direction. This is shown in equation 8.

$$a_j(\vec{p}) = \sum_{i=1}^{N} z_{ji} f_i(\vec{p}) \tag{8}$$

If the rodent is at a location $\vec{p}$ and wants to move it uses a_j to do this. It picks the direction via the probability distribution given in equation 9.

$$P_j = \frac{\exp(2a_j)}{\sum_{k=1}^{8} \exp(2a_k)} \tag{9}$$

The actor weights z_{ji} are also updated via δ_t, which contains information from the critic. This is shown in equation 10, where $g_i(t)$ is 1 for the direction that has been chosen and 0 for the rest and η_z is the learning rate.

$$\Delta z_{ji} = \eta_z \delta_t g_j(t) f_i(\vec{p}_t) \tag{10}$$

3 Results

In the simulation the rodent starts a the right of the barrel and the $N = 484$ place cells are equally spaced in x,y-direction. The space grid where rodent can move in has 952 points. The barrel has a radius of $R = 2$ and the platform is centered at $(x, y) = (-0.5, 0.5)$ and has a radius $r = 0.2$. The learning rates for w and z are both $\eta_w = \eta_z = 0.1$. For the firing rate has been chosen to take $\sigma = 0.2$ and δ_t uses $\gamma = 0.9$. w_i and z_{ji} are both initialized as 0.

In Figure 1 is $C(\vec{p})$ shown at $t = 0$, which is the starting point of the simulation. Figures 2, 3 and 4 show respectively the trials 20,50 and 100 with $C(\vec{p})$ shown as a contour plot and $a_j(\vec{p})$ as vector field, where a vector is the sum of all directions at that point. In Figure 5 is $C(\vec{p})$ shown after 160 trials.

4 Conclusion

In the paper from '*Foster et al.*' it is shown that the reinforcement learning model of actor-critic together with the place cells can be used to create a sense of orientation. This result could not not be established in the model I created. There are certainly high values of $C(\vec{p})$ in the middle. However, it is a vertical area and not a round around the platform. This can be due to wrong parameter settings or a mistake in the code.

5 Appendix

The code to run this can be found in the attached Jupyter Notebook and python file.

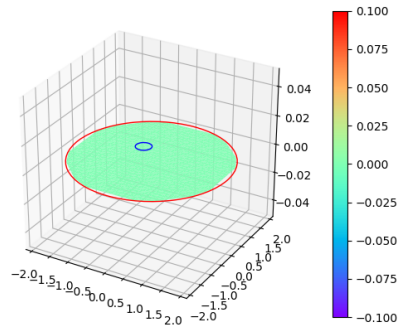

Figure 1: 3D plot of $C(\vec{p})$ for the Barrel at the start $t = 0$.

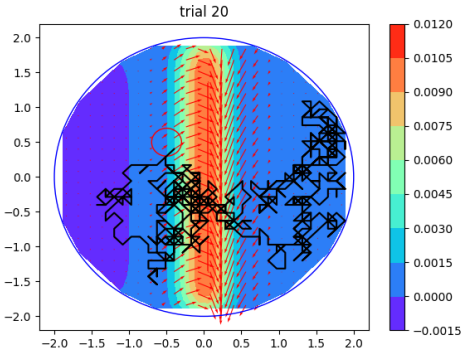

Figure 2: Trial 20 with $C(\vec{p})$ given as contour plot and $a_j(\vec{p})$ as vector field.

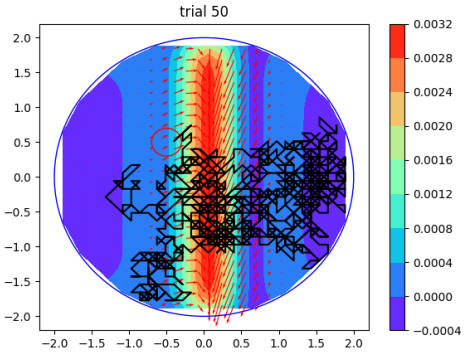

Figure 3: Trial 50 with $C(\vec{p})$ given as contour plot and $a_j(\vec{p})$ as vector field.

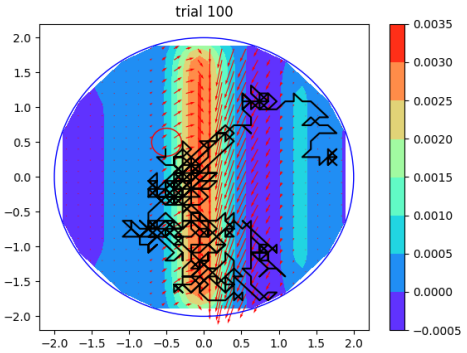

Figure 4: Trial 100 with $C(\vec{p})$ given as contour plot and $a_j(\vec{p})$ as vector field.

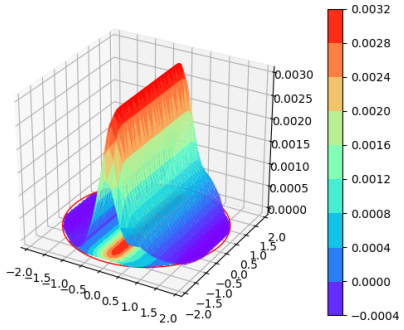

Figure 5: 3D plot of $C(\vec{p})$ for the Barrel at the end.