A Computational Theory of Learning Flexible Reward-Seeking Behavior with Place Cells

Yuanxiang Gao
University of Electronic Science and Technology of China
gaoyuanxiang123@alu.uestc.edu.cn

Abstract

An important open question in computational neuroscience is how various spatially tuned neurons, such as place cells, are used to support the learning of reward-seeking behavior of an animal. Existing computational models either lack biological plausibility or fall short of behavioral flexibility when environments change. In this paper, we propose a computational theory that achieves behavioral flexibility with better biological plausibility. We first train a mixture of Gaussian distributions to model the ensemble of firing fields of place cells. Then we propose a Hebbian-like rule to learn the synaptic strength matrix among place cells. This matrix is interpreted as the transition rate matrix of a continuous time Markov chain to generate the sequential replay of place cells. During replay, the synaptic strengths from place cells to medium spiny neurons (MSN) are learned by a temporal-difference like rule to store place-reward associations. After replay, the activation of MSN will ramp up when an animal approaches the rewarding place, so the animal can move along the direction where the MSN activation is increasing to find the rewarding place. We implement our theory into a high-fidelity virtual rat in the MuJoCo physics simulator. In a complex maze, the rat shows significantly better learning efficiency and behavioral flexibility than a rat that implements a neuroscience-inspired reinforcement learning algorithm, deep Q-network.

1 Introduction

During the past decades, neuroscientists have discovered a variety of spatially tuned neurons, including place cells, head-direction cells, border cells, grid cells, etc [5, 41]. Among them a representative cell type, the place cell, activates when an animal is in a particular location of the environment, serving as the representation of an animal's position. An important open question is how the activation patterns of place cells are used by an animal to support the learning of reward-seeking behavior.

Some computational models [14, 23, 4] proposed that the activation patterns of place cells are the set of basis functions for representing the value function in reinforcement learning (RL). However, reward-seeking behavior learned by RL can not flexibly adapt to the change of the rewarding location or the spatial layout. Another computational theory [42] proposed that place cells and synapses among them can be modeled as a graph. To navigate to a rewarding place, an animal can follow the shortest path found in the graph from an initial place to the rewarding place. However, this model assumes that brains can implement the Dijkstra's shortest path algorithm [12]. There are no evidences that brains can perform such complex path search computations.

In this paper, we propose a computational theory that achieves behavioral flexibility with better biological plausibility. First of all, we use a mixture of Gaussian distributions to model the ensemble of firing fields of place cells. The optimal firing field centers and dispersities are found by maximizing the total responsiveness of place cells to spatial positions. Then we propose a Hebbian-like rule to learn the strengths of synapses among place cells. We find that the synaptic strength between a

Preprint.

pair of place cells not only encodes the spatial correlation of their firing fields but also determines the speed of activation propagation from one cell to another cell. As a result, the synaptic strength matrix among place cells is interpreted as the transition rate matrix of a continuous time Markov chain (CTMC), so the sequential replay of place cells during immobility [45, 55] is modeled as state transitions in the CTMC.

During replay, the strengths of synapses from place cells (PC) to medium spiny neurons (MSN) are learned by a dopamine-modulated temporal-difference like rule to store place-reward associations. We find that the PC-MSN synaptic strength encodes the spatial proximity between the firing field of each place cell and the rewarding place. As a result, during active movement, the activation of the MSN population will ramp up when an animal approaches the rewarding place. In addition, the rate of change of the MSN activation along a direction encodes the motivation of the animal to move along that direction. Accordingly, we design a behavior policy that chooses a direction to move with a probability proportional to the derivative of the MSN activation along that direction.

We implement our theory into a high-fidelity virtual rat in the MuJoCo physics simulator [37, 58]. In a complex maze environment, we compare the rat with a rat that implements a neuroscience-inspired RL algorithm, deep Q-network (DQN) [39, 24]. The rat with our theory learns reward-seeking behavior over 50$\times$ faster than the DQN-based rat. When the rewarding location changes, the DQN-based rat requires over 10000 trials for relearning to navigate to the new location. In contrast, the rat with our theory can flexibly navigate to different rewarding locations without relearning costs. When the maze layout changes, the rat with our theory flexibly adapts its behavior to achieving rewards under the new layout, over 100$\times$ faster than the DQN-based rat.

2 Theory

2.1 Learning firing fields of place cells

Let $f(x|c_k)$ denote the normalized firing rate of the k-th place cell for an animal located at the position x. $f(x|c_k)$ is modeled by a 2D isotropic Gaussian density function denoted as $\mathcal{N}(x|\mu_k, \sigma_k^2)$, where the mean vector μ_k is the firing field center and the standard deviation σ_k characterizes the dispersity of the firing field [14, 23, 4, 42]. For a population of K place cells, the average of their firing rates at position x is denoted by $\bar{f}(x) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(x|\mu_k, \sigma_k^2)$. $\bar{f}(x)$ quantifies the responsiveness of the population of place cells to the position x. For a dataset of positions $X = \{x_1, x_2, ..., x_N\}$ visited by a virtual rat exploring an novel environment, the optimal firing field centers and dispersities are found by maximizing the log-sum of average firing rate at each position as follows,

$$\max_{\{\mu_k, \sigma_k^2\}} \sum_{n=1}^{N} \log[\frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(x_n|\mu_k, \sigma_k^2)]. \tag{1}$$

The objective in Eq. (1) maximizes the total responsiveness of place cells to visited positions in the environment. The logarithmic transformations ensure that the responsiveness of place cells is distributed across all visited positions rather than focusing only on several particular positions.

Solving the optimization problem in Eq. (1) is equivalent to finding a mixture of Gaussians that fits the dataset X by the maximum likelihood principle [7]. We use the expectation-maximization (EM) algorithm to iteratively solve the optimal firing field centers and dispersities from arbitrary initial μ_k and σ_k^2. The E step updates the contribution ratio of the k-th place cell to the total firing rate at x_n,

$$\gamma_{nk}^{(t+1)} = \mathcal{N}(x_n|\mu_k^{(t)}, (\sigma_k^2)^{(t)}) / \sum_j \mathcal{N}(x_n|\mu_j^{(t)}, (\sigma_j^2)^{(t)}), \forall n, k. \tag{2}$$

The M step updates firing field centers and dispersities as follows,

$$\mu_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}^{(t+1)} x_n, \forall k, \tag{3}$$

and

$$(\sigma_k^2)^{(t+1)} = \frac{1}{2N_k}\sum_{n=1}^{N}\gamma_{nk}^{(t+1)}(x_n - \mu_k^{(t+1)})^{\mathrm{T}}(x_n - \mu_k^{(t+1)}), \forall k, \tag{4}$$

where $N_k = \sum_{n=1}^{N}\gamma_{nk}^{(t+1)}$. After each iteration, the total average firing rate in Eq. (1) is guaranteed to increase until convergence [7].

2.2 Learning synaptic strengths among place cells

During random exploration, a pair of place cells with nearby firing fields will often fire in close temporal order. The pair of recurrent synapses between this pair of place cells will undergo symmetric long-term potentiation (LTP) [27, 38], whereas synapses between place cells with faraway firing fields will undergo no change in strength. These experimental observations are modeled as follows.

The strength of the synapse from the i-th place cell to the j-th place cell is denoted by the ij-th element of the $K \times K$ synaptic strength matrix W. For a sequence of positions $X' = \{x_{N+1}, x_{N+2}, ..., x_{N+T}\}$ visited by a virtual rat after learning place cell firing fields, the synaptic strength matrix is symmetrically updated by the following Hebbian-like rule [42, 25],

$$W^{(t+1)} = W^{(t)} + \alpha(f(x_{N+t})^{\mathrm{T}} f(x_{N+t}) - W^{(t)}), \tag{5}$$

where $f(x_{N+t}) = \{f(x_{N+t}|c_k), k = 1, ..., K\}$ is the row vector of place cell firing rates at position x_{N+t}. α is the learning rate. $W^{(0)}$ is chosen as a zero matrix.

The learning rule in Eq. (5) is a stochastic approximation process [43] that solves the following equation as $t \to \infty$,

$$W^* = \mathbb{E}[f(x)^{\mathrm{T}} f(x)], \tag{6}$$

where the expectation is taken with respect to possible positions. The above equation is written elementwisely as follows,

$$W^*_{ij} = \mathbb{E}[f(x|c_i) f(x|c_j)], \forall i, j. \tag{7}$$

As shown by Eq. (7), the synaptic strength between a pair of place cells encodes the correlation of their firing fields, which is determined by the proximity of their firing fields. Thus, the synaptic strength between a pair of place cells also characterizes the proximity of their firing fields.

2.3 The sequential replay of place cells

During wakeful immobility or sleep periods, place cells are spontaneously and sequentially reactivated, similar to the pattern of activations during previous movement periods [45]. This process is called hippocampal replay. Recent studies further show that place cells are activated following a random walk like pattern during replay [55, 47]. These experimental observations inspired us to model the replay as follows.

During replay, the activation is started from a randomly chosen place cell i. According to the quantal release of neurotransmitters [9, 17, 35], each spike of place cell i induces the firing of a connected place cell j with a certain probability proportional to the synaptic strength W_{ij}. Thus, the time for the activation of place cell i to fire place cell j approximately follows an exponential distribution T_{ij} with mean $1/W_{ij}$. During these exponentially distributed waiting times, the fastest fired place cell will inhibit the activations of other place cells through fast inhibitory interneurons following a winner-take-all like mechanism [2, 11]. As a result, only the activation of the winning place cell is maintained and the activation is now transferred from place cell i to the winning place cell. Similarly, the activation of the winning place cell will be transferred to another place cell following the waiting and competition process. This process is formally a continuous time Markov chain (CTMC) [44] with the transition rate matrix given by the synaptic strength matrix W.

According to properties of exponential distributions [44], the minimum of the set of waiting times $H_i = \min_j T_{ij}, \forall i$ is still exponentially distributed with mean $1/\sum_{j \neq i} W_{ij}$. In addition, the probability that the waiting time of the place cell k achieves the minimum is $P_{ik} = W_{ik}/\sum_{j \neq i} W_{ij}$. Accordingly, an equivalent approach to generate activation sequences is as follows. The activation is started from a randomly chosen place cell i. After an exponentially distributed holding time H_i, the activation is transferred to another place cell k with probability P_{ik}. Similarly, the activation of the place cell k is stochastically transferred to another place cell after a holding time H_k.

2.4 Learning PC-MSN synaptic strengths

The major functional role of replay is to learn place-reward associations [31, 53, 49, 46, 22]. Such associations are stored in the synaptic strengths from hippocampal place cells (PC) to medium spiny neurons (MSN) in the striatum [60, 32, 59]. During replay, the activation of a place cell leaves a chemical trace (e.g., Ca^{2+} concentration increase) at the synapse from the place cell to the population

of MSN [28, 62, 1]. The Ca^{2+} is gradually taken up by organelles hence the chemical trace decays exponentially at the synapse [62, 1]. Before the Ca^{2+} is fully absorbed, if the dopamine concentration increases from a base level at the synapse, the synapse is strengthened proportional to both the chemical trace and the increase of the dopamine concentration [28, 62, 21]. These experimental observations are modeled as follows.

Let Δt denote the refractory period [20] between successive spikes of a place cell. Let $V_i^{(n\Delta t)}$ denote the synaptic strength from the i-th place cell to the MSN population at the n-th period. We call the place cell with maximal responsiveness to a given goal location x_g as the goal cell [18]. The index of the goal cell is determined by $\arg\max_k f(x_g|c_k)$. The activation of the goal cell at the n-th period is denoted by $g^{(n\Delta t)}$, which equals 1 if the activated place cell at this period is the goal cell, otherwise it equals 0. Let $e_i^{(n\Delta t)}$ denote the chemical trace at the synapse from the i-th place cell to the MSN population at the n-th period. Let $d^{(n\Delta t)}$ denote the deviation of dopamine concentration upon a base level at PC-MSN synapses at the n-th period. During replay, the PC-MSN synaptic strengths are updated by the following temporal-difference (TD) like rule [57, 21],

$$\begin{cases} V_i^{[(n+1)\Delta t]} = V_i^{(n\Delta t)} + \alpha e_i^{(n\Delta t)} d^{(n\Delta t)}, \forall i, \\ e_i^{[(n+1)\Delta t]} = e_i^{(n\Delta t)} \gamma^{\Delta t} + I_{c[(n+1)\Delta t]i}, \forall i, \\ d^{(n\Delta t)} = g^{(n\Delta t)} + \gamma^{\Delta t} V_{c[(n+1)\Delta t]}^{(n\Delta t)} - V_{c(n\Delta t)}^{(n\Delta t)}, \end{cases} \tag{8}$$

where $c(n\Delta t)$ and $c[(n+1)\Delta t]$ denote the index of the place cell activated at time $n\Delta t$ and $(n+1)\Delta t$, respectively. $I_{c[(n+1)\Delta t]i}$ is the indicator function that equals 1 if $c[(n+1)\Delta t]$ is i and 0 otherwise. α is the learning rate and γ is a discount factor close to 1 (e.g., 0.99). In Eq. (8), the chemical trace at each synapse decays geometrically with rate γ. In addition, if the place cell activated at time $(n+1)\Delta t$ is the i-th place cell, the chemical trace at the i-th PC-MSN synapse increases by 1. In Eq. (8), the deviation of dopamine concentration characterizes the activation of VTA dopamine neurons, which is a summation of following three signals, the signal on the disinhibitory pathway from the goal cell [34], the signal on the disinhibitory direct pathway from the MSN population and the signal on the inhibitory indirect pathway from the MSN population [29, 40].

As derived in Appendix A.1, the learning rule in Eq. (8) is equivalent to a stochastic approximation process [43] that solves the following equation as $n \to \infty$,

$$V_i^* = \mathbb{E}[\sum_{n=0}^{\infty} \gamma^{n\Delta t} g^{(n\Delta t)}], \;\; c(0) = i, \tag{9}$$

where the expectation $\mathbb{E}[\cdot]$ is taken with respect to all possible activation transitions among place cells during replay. According to Eq. (9), V_i^* is the expectation of the total discounted activation of the goal cell during replay started from the i-th place cell. For a place cell with its firing field geodesically close to the firing field of the goal cell, replays started from this place cell are more probable to frequently activate the goal cell in a near future. Thus, the synaptic strength from this place cell to the MSN population is stronger and vice versa. Therefore, V_i^* encodes the geodesic proximity between the firing field of the i-th place cell and the firing field of the goal cell.

2.5 The generation of reward-seeking behavior

Let $V(x)$ denote the activation of the MSN population when the animal is located at x during active movement. At the i-th place cell's firing field center μ_i, where the activation of other place cells is typically weak, the activation of the MSN population depends only on the i-th PC-MSN synapse strength V_i, so we have $V(\mu_i) \approx V_i, i = 1, ..., K$. Since $V(x)$ is a continuous function [60], $V(x)$ can be estimated by linearly interpolating the activation at three nearby firing field centers μ_i, μ_j and μ_k as follows,

$$V(x) = a_1 V(\mu_i) + a_2 V(\mu_j) + a_3 V(\mu_k), \tag{10}$$

where nonnegative coefficients $a_1 + a_2 + a_3 = 1$ and μ_i, μ_j, μ_k are determined by first finding the Delaunay triangulation [6] of the set of firing field centers then choosing the three vertices of the triangle enclosing x as μ_i, μ_j and μ_k, respectively. a_1 equals the area ratio between the triangle $\Delta x\mu_i\mu_j$ and the triangle $\Delta\mu_i\mu_j\mu_k$ [48, 10]. a_2 equals the area ratio between $\Delta x\mu_j\mu_k$ and $\Delta\mu_i\mu_j\mu_k$. a_3 equals the area ratio between $\Delta x\mu_k\mu_i$ and $\Delta\mu_i\mu_j\mu_k$.

Let $\partial_\theta V(x)$ denote the directional derivative of $V(x)$ along the direction θ (the angle with x-axis) at location x. $\partial_\theta V(x)$ characterizes the dopamine release upon a base level while moving along the θ direction at location x [30, 19], which determines the motivation for moving along that direction [50, 26]. Normalizing the motivation along each direction by a softmax function defines a probability distribution with respect to θ as follows,

$$P_x(\theta) = \frac{\exp(\beta \partial_\theta V(x))}{\int_0^{2\pi} \exp(\beta \partial_\theta V(x)) \mathrm{d}\theta}, \tag{11}$$

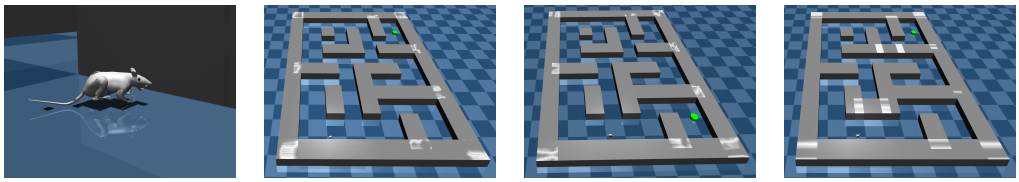
where β characterizes the greediness level of behavior. When β varies from zero to infinity, the behavior changes from purely random exploration to dopamine-maximization behavior. Since it's difficult to sample the above continuous probability distribution, we use a discrete probability distribution by considering M uniformly spaced directions $\{\theta_m, m = 1, ..., M\}$ as follows,

$$P_x(\theta_m) = \frac{\exp(\beta \partial_{\theta_m} V(x))}{\sum_{m=1}^{M} \exp(\beta \partial_{\theta_m} V(x))}. \tag{12}$$

Dopamine release is a gating signal to initiate a period of movement rather than continuously modulating ongoing movements [50, 26]. Accordingly, the locomotion of the virtual rat is divided into periods of movements. During each movement period, the virtual rat first samples the discrete distribution in Eq. (12) and then keeps moving along the chosen direction until the end of this movement period.

3 Experiments

3.1 Experimental setup



(a) (b) (c) (d)

Figure 1: The virtual rat and maze environments. (a) The virtual rat used in experiments. (b) The maze environment used in the goal-fixed experiment. (c) The maze environment used in the goal-changing experiment. (d) The maze environment used in the maze-changing experiment.

Subject. The subject used is a virtual rat that reproduces the anatomical features of seven Long-Evans rats in the MuJoCo physics simulator [37, 58], as shown in Figure 1(a). At each simulation step, the proprioceptive input to the virtual rat is an 148-dimension vector that includes the position, velocity and angular velocity of each joint. The motor output from the virtual rat is a 38-dimension vector that contains the torque applied to each joint. We use the TD3 deep RL algorithm [16] to pretrain the rat in an open field to master several basic locomotion policies, including running forward along the current body direction, turning the body by a certain angle then running forward. The trained turning angles are 45°, 90 °, 135°, 180 °, -45 °, -90 ° and -135 °. Each basic locomotion policy is represented by a feedforward neural network that maps the proprioceptive input into the correct motor output to generate the desired locomotion. The pretraining details see Appendix A.2. For moving along a direction sampled from Eq. (12), the rat compares its current body direction with the sampled direction and invokes the closest locomotion policy to approximately run along the sampled direction for a period.

For testing the proposed computational theory, we build a large scale 9×11m maze in MuJoCo as shown in Figure 1(b). We conduct following three experiments in the maze environment.

Goal-fixed experiment. In the goal-fixed experiment, a location at the top right corner is set as the goal location, as shown in Figure 1(b). The rat first randomly explores the maze for 100 epochs

(trials) starting from random initial positions in the maze. Every epoch contains 4000 simulation steps. Before every 100 simulation steps, the rat randomly chooses a turning angle and runs along this direction in next 100 simulation steps. The positions visited during the first 50 epochs are used to learn the firing fields of 120 place cells by the EM iterations given in Eq. (2)-Eq. (4). After that, the positions visited during the last 50 epochs are used to learn the synaptic strength matrix of place cells by the Hebbian-like rule given in Eq. (5). Then the rat reactivates place cells to learn the PC-MSN synaptic strengths by the TD-like rule given in Eq. (8). After that, the rat performs a series of test epochs with every epoch containing 4000 simulation steps. Before every 100 simulation steps, the rat samples a new direction from Eq. (12) and runs along this direction for next 100 simulation steps. Test epochs with the rat ending up within a 1×1m square enclosing the goal location are treated as successful epochs. The success rate of the most recent 50 epochs is recorded as the performance measure of the rat.

Goal-changing experiment. In the goal-changing experiment, the exploration and learning phase is same as the goal-fixed experiment. However, after some test epochs, the goal location is changed from the top right corner to the bottom right corner as shown in Figure 1(c). After goal changing, the rat reperforms replay to update the PC-MSN synaptic strengths with a new goal cell. Then the rat continues to perform test epochs.

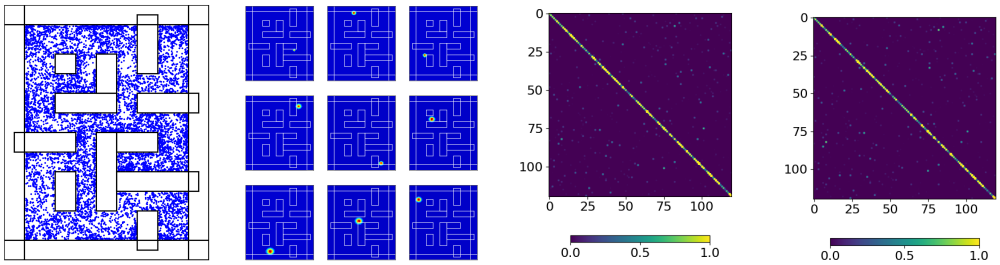
Maze-changing experiment. In the maze-changing experiment, the exploration and learning phase is also same as the goal-fixed experiment. However, after some test epochs, two passages in the maze are closed as shown in Figure 1(d). After spatial layout changing, the rat randomly re-explores the maze for 100 epochs. The positions visited during the first 50 and last 50 epochs are used to update place cells' firing fields and synaptic strength matrix, respectively. After that, the rat reactivates place cells to update the PC-MSN synaptic strengths and then continues to perform test epochs.

We compare the rat equipped with our theory against rats that implement a neuroscience-inspired deep RL algorithm, deep Q-network (DQN) [39, 24]. We design two variations of DQN algorithms, PC+DQN and GC+DQN.

PC+DQN. This baseline uses the experienced positions and head directions during the exploration phase to learn the firing fields of 120 place cells and 120 head direction cells. Then a series of training epochs is performed. Before every 100 simulation steps during each training epoch, the firing rates of place cells and head direction cells are fed into a feedforward neural network to predict the Q-values of the eight basic locomotion policies. During next 100 simulation steps, the policy with the maximal Q-value is used with probability $1-\epsilon$ or a randomly chosen policy is used with probability ϵ. The reward is one if the rat is located within a 1×1m square enclosing the goal location, otherwise the reward is zero. The Q-value network is trained by the DQN algorithm.

GC+DQN. This baseline uses firing rates of grid cells (GC) and head direction cells as input of the Q-value network. The model used to compute firing rates of grid cells is provided in Appendix A.3. Except for the input, the GC+DQN baseline uses the same setup as the PC+DQN baseline.

Hyperparameters and their values used in experiments are summarized in Appendix A.4.

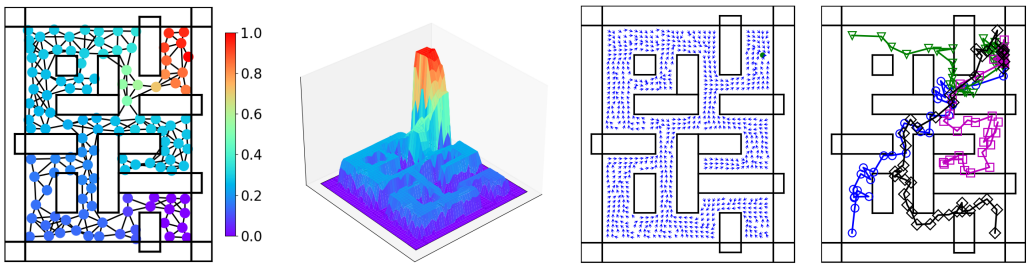


(a) (b) (c) (d)

Figure 2: Goal-fixed experiment. (a) Spatial positions visited during exploration. (b) Exemplary place cell firing fields. (c) The spatial correlation matrix of firing fields. (d) The synaptic strength matrix of place cells.

3.2 Experimental results

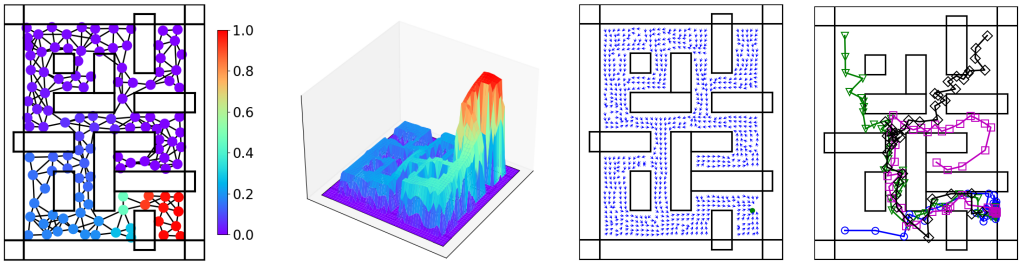
Goal-fixed experiment. Figure 2(a) shows the positions visited by the rat during exploration epochs. After the 100 exploration epochs, the maze is fully explored and the spatial experiences cover the whole maze. Figure 2(b) shows the firing fields of several place cells. Each place cell is selectively responsive to a restricted area of the maze. The firing fields of all place cells can cover the whole maze. Figure 2(c) shows the correlation between each pair of firing fields computed by Eq. (7). A firing field is typically highly correlated with several nearby firing fields. Figure 2(d) shows the synaptic strength matrix of place cells learned by Hebbian-like learning. The synaptic strength matrix well characterizes the firing field correlations among place cells.



(a) (b) (c) (d)

Figure 3: Goal-fixed experiment. (a) The graph representation of place cells and synapses. (b) The activation function of the MSN population. (c) The gradient field of the MSN activation function. (d) Exemplary movement trajectories all converge to the goal location.

Goal-fixed experiment. Figure 3(a) is the graph representation of place cells and their synapses. In the graph, each node represents a place cell with firing field center located at the position of the node. Each edge represents a pair of recurrent synapses with strength larger than a predefined threshold (e.g., 0.0001). The synaptic strengths among place cells well reproduces the complex layout of the maze. The color on each node characterizes the synaptic strength from this place cell to the population of MSN cells after replay. As our theory predicts, the PC-MSN synaptic strength encodes the geodesic proximity between the firing field of each place cell and the firing field of the goal cell. Figure 3(b) shows the MSN activation as a function of the rat's location in the maze. The activation will ramp up if the rat is approaching the goal location from any initial locations, which conforms with experimental observations about MSN cells [60, 33]. Figure 3(c) shows the gradient field of the MSN activation function. The gradient field has a clear tendency to approach the goal location. A rat that greedily chooses the gradient direction to move can navigate to the goal location from any initial locations. Figure 3(d) shows several exemplary movement trajectories during test epochs.



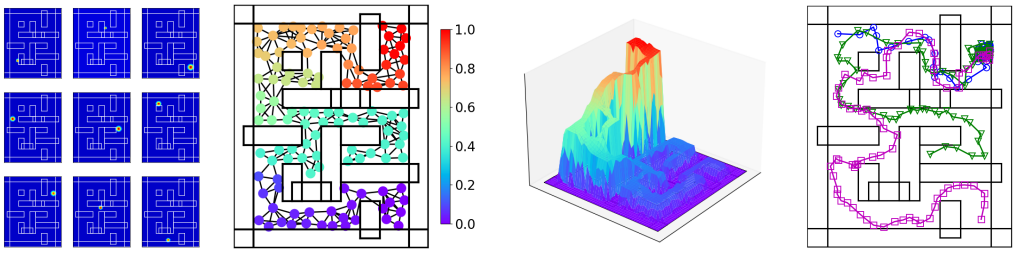
(a) (b) (c) (d)

Figure 4: Goal-changing experiment. (a) The graph representation after goal change. (b) The MSN activation function after goal change. (c) The gradient field of the MSN activation function. (d) Exemplary movement trajectories all converge to the new goal location.

Goal-changing experiment. Figure 4(a) shows the graph representation of place cells and their synapses after reperforming replay. The topology of the graph representation before goal changing is preserved after goal changing. However, the PC-MSN synaptic strengths are updated during replay

to re-encode the geodesic proximity between each place cell to the new goal cell. As Figure 4(b) shows, the MSN activation function now ramps up along paths to the new goal location. Accordingly, the gradient field of the MSN activation function is re-directed toward the new goal location as shown in Figure 4(c). Figure 4(d) shows several exemplary movement trajectories during test epochs. Starting from arbitrary initial locations, the rat can flexibly navigate to the new goal location without re-exploring the maze.

Maze-changing experiment. Figure 5(a) shows several updated place cell firing fields after re-exploration. Due to spatial layout change, place cells will no longer respond to the blocked areas. As Figure 5(b) shows, the updated graph representation well characterizes the new spatial layout. Accordingly, the PC-MSN synaptic strengths reflect the changed geodesic proximity from each place cell to the goal cell. Similarly, the MSN activation function in Figure 5(c) ramps up along new paths to the goal location. Figure 5(d) shows several exemplary movement trajectories in the new maze. Starting from initial positions far away from the goal location, the rat can find a detour to the goal location. Such detour behavior resembles that observed in mammals [3].

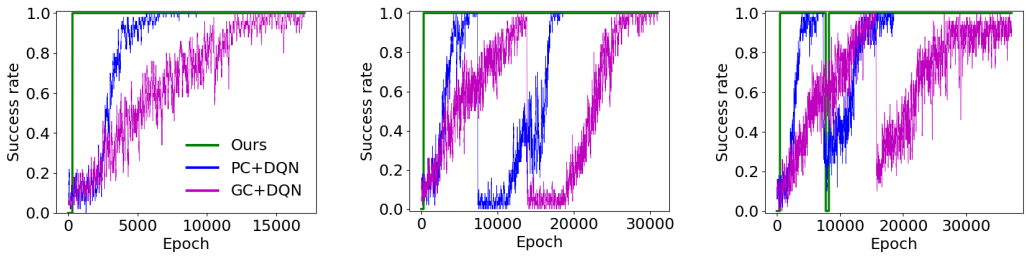


(a) (b) (c) (d)

Figure 5: Maze-changing experiment. (a) Exemplary place cell firing fields after re-exploration. (b) The graph representation after maze change. (c) The MSN activation function after maze change. (d) Exemplary movement trajectories show detour behavior.

Comparison to baselines. Figure 6(a) shows the performance of baselines in the goal-fixed experiment. The success rate of the rat using our theory keeps at 1 after exploration epochs. In contrast, DQN-based rats require a large amount of epochs to incrementally improve performance by trial and errors. The GC+DQN shows slower progress than PC+DQN probably because it's more difficult to estimate the animal's position from grid cells' firing rates than from place cells' firing rates [61].

Figure 6(b) shows the performance of baselines in the goal-changing experiment. The success rate of the rat using our theory keeps at 1 despite the change of the goal location after 500 epochs. In contrast, the DQN-based rats suffer from a significant degradation of performance after goal changing at the 7750-th epoch and the 15500-th epoch for PC+DQN and GC+DQN, respectively. After goal changing, the trained state-policy mapping of DQN-based rats is unable to reach the new goal location. However, relearning a new mapping needs a large amount of trial and error epochs to gradually overwrite the learned mapping.



(a) (b) (c)

Figure 6: Comparison to baselines. (a) Performance in the goal-fixed experiment. (b) Performance in the goal-changing experiment. (c) Performance in the maze-changing experiment.

Figure 6(c) shows the performance of baselines in the maze-changing experiment. The success rate of the rat using our theory recovers to 1 after 100 re-exploration epochs at the 7750-th epoch. In contrast, the performance of DQN-based rats is significantly disrupted after the change of spatial layout at the 7750-th epoch and the 15500-th epoch for PC+DQN and GC+DQN, respectively. After spatial layout change, the trained state-policy mapping of DQN-based rats still tries to go through the blocked passages. However, relearning a correct mapping adapted to the new spatial layout still requires a large amount of trial and error epochs. Differently, the rat using our theory adapts to the new maze by updating the synaptic strengths among place cells to match the new spatial layout. With updated inter-PC synaptic strengths, PC-MSN synaptic strengths and the MSN activation function are immediately updated by replaying place cells. As a result, the rat can fast learn to reach arbitrary goal locations in the new maze by stochastically climbing the updated MSN activation function.

4 Discussion

We have drawn inspiration from many neuroscience experiments and provided an integrative theory on how various information processing mechanisms in brains support the efficient learning of flexible reward-seeking behavior. Such a systematic view about brains is beneficial to the design of neuroscience-inspired artificial intelligence. We have shown that incorporating neuroscientific understandings about learning, memory and motivation into the design loop of an artificial agent is important for building artificial intelligence that matches the behavioral performance of animals.

Several works [52, 13, 4] trained recurrent neural networks (RNN) to perform path integration tasks in square environments. After the RNN correctly predicts positions from velocities, the firing fields of the hidden units of the RNN show grid-like patterns. The formation of grid patterns in these works typically requires some conditions, such as square environments without obstacles, place cells with firing fields given by the difference of Gaussians, special regularizations, etc. For a complex maze environment without these conditions satisfied, there is no guarantee that grid patterns will emerge. We leave the study of how place cell firing fields in a complex maze are formed from grid cell activations for our future work. The work [4] further fed the activations of RNN hidden units into another RNN for representing the policy and value function in actor-critic based RL. Similar to the DQN-based baselines, learning reward-seeking behavior with the actor-critic RL is inefficient and the learned behavior is inflexible to change of the rewarding location or the spatial layout.

The work [14] proposed a linear actor-critic RL model and trained a simulated rat to perform Morris water maze tasks. Another study [23] found that using place cells with firing fields decaying with the geodesic distance from the field center can improve the efficiency of reinforcement learning. However, computing the geodesic firing fields of place cells requires complicated coordinate transformations of the original space. How such computations can be performed in brains remains unclear. Another study [54] interpreted the firing field of a place cell as encoding the future expected number of visits to the firing field peak (center) from an initial location. Such predictive perspective on place cell firing fields can explain some asymmetric firing fields observed in previous experiments [36]. Our theory adopts the Gaussian model of place cell firing fields [14, 23, 4, 52, 42]. It's an interesting direction to extend our theory based on other models of place cell firing fields. One study [15] used spiking neural networks to implement the actor-critic reinforcement learning for navigation tasks. It's a promising future direction to implement our theory with spiking neural networks.

The work [42] interpreted the hippocampus as a graph that connects place cells by synapses. This work used the biologically unrealistic Dijkstra's algorithm to compute the path with the smallest total synaptic resistance (reciprocal synaptic strength) between two place cells, which is further used to find the shortest path between two spatial positions. Compare to such path search algorithms, we show that the learning of PC-MSN synaptic strengths during hippocampal replay is a more biologically plausible model to explain how synaptic information stored among place cells is used to generate reward-seeking behavior.

References

[1] T. Abrams and E. Kandel. Is contiguity detection in classical conditioning a system or a cellular property? learning in aplysia suggests a possible molecular site. *Trends in Neurosciences*, 11(4):128–135, 1988.

[2] L. Almeida, M. Idiart, and J. Lisman. A second function of gamma frequency oscillations: An E%-max winner-take-all mechanism selects which cells fire. *The Journal of Neuroscience*, 29(23):7497–7503, 2009.

[3] A. Alvernhe, E. Save, and B. Poucet. Local remapping of place cell firing in the Tolman detour task. *European Journal of Neuroscience*, 33:1696–1705, 2011.

[4] A. Banino, C. Barry, and B. Uria et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557:429–433, 2018.

[5] T. Behrens, T. Muller, and J. Whittington et al. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

[6] M. Berg, O. Cheong, M. Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications*. Springer, 3rd edition, 2008.

[7] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[8] H. Blair, A. Welday, and K. Zhang. Scale-invariant memory representations emerge from Moire interference between grid fields that produce theta oscillations: A computational model. *The Journal of Neuroscience*, 27(12):3211–3229, 2007.

[9] T. Branco and K. Staras. The probability of neurotransmitter release: variability and feedback control at single synapses. *Nature Reviews Neuroscience*, 10:373–383, 2009.

[10] S. Buss. *3D Computer Graphics: A Mathematical Introduction with OpenGL*. Cambridge University Press, 2003.

[11] L. Colgin. Rhythms of the hippocampal network. *Nature Reviews Neuroscience*, 17:239–249, 2016.

[12] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 3rd edition, 2009.

[13] C. Cueva and X. Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *International Conference on Learning Representations (ICLR)*, 2018.

[14] D. Foster, R. Morris, and P. Dayan. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, 10(1):1–16, 2000.

[15] N. Fremaux, H. Sprekeler, and W. Gerstner. Reinforcement learning using a continuous time actor-critic framework with spiking neurons. *PLoS Computational Biology*, 9(4):1–21, 2013.

[16] S. Fujimoto, H. Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning (ICML)*, 2018.

[17] F. Gabbiani and S. Cox. *Mathematics for Neuroscientists*. Elsevier, 2007.

[18] J. Gauthier and D. Tank. A dedicated population for reward coding in the hippocampus. *Neuron*, 99(1):179–193, 2018.

[19] S. Gershman. Dopamine ramps are a consequence of reward prediction errors. *Neural Computation*, 26(3):467–471, 2014.

[20] W. Gerstner. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014.

[21] W. Gerstners, M. Lehmann, V. Liakoni, D. Corneil, and J. Brea. Eligibility traces and plasticity on behavioral time scales: Experimental support of neohebbian three-factor learning rules. *Frontiers in Neural Circuits*, 12(53):1–16, 2014.

[22] S. Gomperts, F. Kloosterman, and M. Wilson. VTA neurons coordinate with the hippocampal reactivation of spatial experience. *eLife*, 4:1–22, 2015.

[23] N. Gustafson and N. Daw. Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS Computational Biology*, 7(10):1–14, 2011.

[24] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[25] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1998.

[26] M. Howe and D. Dombeck. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature*, 535:505–510, 2016.

[27] J. Isaac, K. Buchanan, R. Muller, and J. Mellor. Hippocampal place cell firing patterns can induce long-term synaptic plasticity in vitro. *The Journal of Neuroscience*, 29(21):6840 – 6850, 2009.

[28] H. Kasai, N. Ziv, and H. Okazaki et al. Spine dynamics in the brain, mental disorders and artificial neural networks. *Nature Reviews Neuroscience*, 22(7):407–422, 2021.

[29] R. Keiflin and P. Janak. Dopamine prediction errors in reward learning and addiction: From theory to neural circuitry. *Neuron*, 88(2):247–263, 2015.

[30] H. Kim, A. Malik, and J. Mikhael et al. A unified framework for dopamine signals across timescales. *Cell*, 183(6):1600–1616, 2020.

[31] C. Lansink, P. Goltstein, J. Lankelma, B. McNaughton, and C. Pennartz. Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biology*, 7(8):1–11, 2009.

[32] T. LeGates, M. Kvarta, and J. Tooley et al. Reward behaviour is regulated by the strength of hippocampus–nucleus accumbens synapses. *Nature*, 564:258–262, 2018.

[33] T. London, J. Licholai, and I. Szczot. Coordinated ramping of dorsal striatal pathways preceding food approach and consumption. *Journal of Neuroscience*, 38(14):3547–3558, 2018.

[34] A. Luo, P. Tahsili-Fahadan, and R. Wise et al. Linking context with reward: A functional circuit from hippocampal CA3 to ventral tegmental area. *Science*, 333(15):353–357, 2011.

[35] W. Maass and A. Zador. Dynamic stochastic synapses as computational units. *Neural Computation*, 11(4):903–917, 1999.

[36] M. Mehta, M. Quirk, and M. Wilson. Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, 25(3):707–715, 2000.

[37] J. Merel, D. Aldarondo, and J. Marshall et al. Deep neuroethology of a virtual rodent. In *International Conference on Learning Representations (ICLR)*, 2020.

[38] R. Mishra, S. Kim, S. Guzman, and P. Jonas. Symmetric spike timing-dependent plasticity at CA3–CA3 synapses optimizes storage and recall in autoassociative networks. *Nature Communications*, 7:1–11, 2016.

[39] V. Mnih, K. Kavukcuoglu, and D. Silver et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[40] K. Morita, M. Morishima, and K. Sakai et al. Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways. *Trends in Neurosciences*, 35(8):457–467, 2012.

[41] E. Moser, E. Kropff, and M. Moser et al. Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, 31:69–89, 2008.

[42] R. Muller, M. Stead, and J. Pach. The hippocampus as a cognitive graph. *Journal of General Physiology*, 107(6):663–694, 1996.

[43] M. Nevelson and R. Hasminskii. *Stochastic Approximation and Recursive Estimation*. American Mathematical Society, 1976.

[44] J. Norris. *Markov Chains*. Cambridge University Press, 1997.

[45] H. Olafsdottir, D. Bush, and C. Barry. The role of hippocampal replay in memory and planning. *Current Biology*, 28(1):37–50, 2017.

[46] C. Pennartz, E. Lee, and J. Verheul et al. The ventral striatum in off-line processing: Ensemble reactivation during sleep and modulation by hippocampal ripples. *The Journal of Neuroscience*, 24(29):6446–6456, 2004.

[47] B. Pfeiffer and D. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497:74–79, 2013.

[48] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd edition, 2007.

[49] J. Sadowski, M. Jones, and J. Mellor et al. Sharp-wave ripples orchestrate the induction of synaptic plasticity during reactivation of place cell firing patterns in the hippocampus. *Cell Reports*, 14(8):1916–1929, 2016.

[50] J. Silva, F. Tecuapetla, V. Paixao, and R. Costa. Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature*, 554:244–248, 2018.

[51] T. Solstad, E. Moser, and G. Einevoll. From grid cells to place cells: A mathematical model. *Hippocampus*, 16(12):1026–1031, 2006.

[52] B. Sorscher, G. Mel, S. Ganguli, and S. Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[53] M. Sosa, H. Joo, and L. Frank. Dorsal and ventral hippocampal sharp-wave ripples activate distinct nucleus accumbens networks. *Neuron*, 105(4):725–741, 2020.

[54] K. Stachenfeld, M. Botvinick, and S. Gershman et al. The hippocampus as a predictive map. *Nature Neuroscience*, 20:1643–1653, 2017.

[55] F. Stella, P. Baracskay, J. O'Neill, and J. Csicsvari. Hippocampal reactivation of random trajectories resembling Brownian diffusion. *Neuron*, 102(2):450–461, 2019.

[56] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[57] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[58] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[59] S. Trouche, V. Koren, and N. Doig et al. A hippocampus-accumbens tripartite neuronal motif guides appetitive memory in space. *Cell*, 176(6):1393–1406, 2019.

[60] M. van der Meer and A. Redish. Theta phase precession in rat ventral striatum links place and reward information. *The Journal of Neuroscience*, 31(8):2843–2854, 2011.

[61] X. Wei, J. Prentice, and V. Balasubramanian. A principle of economy predicts the functional architecture of grid cells. *eLife*, 4:1–29, 2015.

[62] S. Yagishita, A. Hayashi-Takagi, and G. Ellis-Davies et al. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345:1616–1620, 2014.

A Appendix

A.1 Derivations in Section 2.4

The chemical trace of the i-th PC-MSN synapse at time $n\Delta t$ is given by,

$$e_i^{(n\Delta t)} = \sum_{k=0}^{n} \gamma^{(n-k)\Delta t} I_{c(k\Delta t)i}. \tag{13}$$

Accordingly, the accumulated update to the i-th PC-MSN synapse when $n \to \infty$ can be written as follows,

$$\begin{aligned}
\sum_{n=0}^{\infty} \Delta V_i^{(n\Delta t)} &= \alpha \sum_{n=0}^{\infty} d^{(n\Delta t)} \sum_{k=0}^{n} \gamma^{(n-k)\Delta t} I_{c(k\Delta t)i}, \\
&= \alpha \sum_{k=0}^{\infty} d^{(k\Delta t)} \sum_{n=0}^{k} \gamma^{(k-n)\Delta t} I_{c(n\Delta t)i}, \\
&= \alpha \sum_{n=0}^{\infty} I_{c(n\Delta t)i} \sum_{k=n}^{\infty} \gamma^{(k-n)\Delta t} d^{(k\Delta t)}, \\
&= \alpha \sum_{n=0}^{\infty} I_{c(n\Delta t)i} \sum_{k=n}^{\infty} \gamma^{(k-n)\Delta t} [g^{(k\Delta t)} + \gamma^{\Delta t} V_{c[(k+1)\Delta t]}^{(k\Delta t)} - V_{c(k\Delta t)}^{(k\Delta t)}], \\
&= \alpha \sum_{n=0}^{\infty} I_{c(n\Delta t)i} (\sum_{k=n}^{\infty} \gamma^{(k-n)\Delta t} g^{(k\Delta t)} - V_{c(n\Delta t)}^{(n\Delta t)}), \\
&= \alpha \sum_{n=0}^{\infty} I_{c(n\Delta t)i} (\sum_{k=n}^{\infty} \gamma^{(k-n)\Delta t} g^{(k\Delta t)} - V_i^{(n\Delta t)}).
\end{aligned} \tag{14}$$

In Eq. (14), the second line is obtained by changing the role of k and n in the first line. The third line is obtained by interchanging the order of summation in the second line. The fourth line is obtained by plugging $d^{(k\Delta t)}$ in Eq. (8) into the third line. The fifth line is obtained by canceling out intermediate terms in the fourth line. The last line holds due to the indicator function $I_{c(n\Delta t)i}$. This derivation is inspired from the equivalence of forward and backward views of the TD(λ) algorithm [56].

The accumulated update in Eq. (14) equals the accumulated update of the following update rule,

$$V_i^{[(n+1)\Delta t]} = V_i^{(n\Delta t)} + \alpha I_{c(n\Delta t)i}(\sum_{k=n}^{\infty} \gamma^{(k-n)\Delta t} g^{(k\Delta t)} - V_i^{(n\Delta t)}), \forall i. \tag{15}$$

Following Eq. (15), the i-th PC-MSN synapse is updated only when the i-th place cell is activated. The change of synaptic strength is proportional to the difference between the future total discounted activation of the goal cell and the current synaptic strength. When n is large, the synaptic strengths in Eq. (8) approximate $V_i^{(n\Delta t)}$ in Eq. (15). Accordingly, the update rule in Eq. (8) is equivalent to a stochastic approximation process [43] that solves the following equation as $n \to \infty$,

$$V_i^* = \mathbb{E}[\sum_{n=0}^{\infty} \gamma^{n\Delta t} g^{(n\Delta t)}], \;\; c(0) = i, \tag{16}$$

where the expectation $\mathbb{E}[\cdot]$ is taken with respect to all possible activation transitions among place cells during replay.

A.2 Pretraining details

In order to train the policy that turns θ° then runs forward, we use the following reward function,

$$r = vel_g + 0.1\vec{h}_g \cdot \vec{h}_r + 0.1\vec{n}_z \cdot \vec{n}_r. \tag{17}$$

Here vel_g is the velocity of the rat along the targeted direction. $\vec{h}_g$ is the unit vector along the targeted direction. $\vec{h}_r$ is the unit vector along the head direction of the rat. $\vec{n}_z$ is the unit vector along the z-axis in the global coordinate frame. $\vec{n}_r$ is the unit vector along the z-axis in the local coordinate frame of the rat. To obtain high rewards, the rat requires to turn into the targeted direction and run as fast as possible while keeping $\vec{n}_r$ upward to avoid falling down. The TD3 algorithm trains each policy network from scratch to maximize the accumulated rewards collected from a behavior trajectory. Training each policy takes several hours on a GPU server.

A.3 The model for computing grid cell firing rates

When the rat is located at position x, the firing rate of a grid cell is modeled by the sum of three 2-dimension cosine waves as follows [8, 51],

$$f_{gc}(x) = \sum_{k=1}^{3} \cos(\omega_k \cdot (x - p)), \tag{18}$$

where p is the spatial phase of the grid cell and ω_k is the wave vector of the k-th cosine wave. The directions of the wave vectors determine the angular orientation of the grid pattern. To produce a grid pattern with a given angular orientation θ, the directions of the three vectors should take $\theta + \pi/2$, $\theta - \pi/6$ and $\theta + \pi/6$, respectively. The norms of the three wave vectors are the same and they determine the spatial periodicity of grid patterns. To produce a collection of grid cells [23], the values of θ, $||\omega_k||$ and p are taken from the set $[0, \pi/12, \pi/6, \pi/4]$, the set $[1, 2, 3, 4, 5]$ and the Cartesian product $[0, 0.5, 1] \times [0, 0.5, 1]$, respectively.

A.4 Hyperparameters

Table 1: Hyperparameters and their values

Hyperparameter	Value
Number of place cells	120
Learning rate for Hebbian learning	0.00001
Discount factor	0.999
Refractory period	0.01
Learning rate for TD learning	0.00001
Greediness level of behavior	25
Number of sample directions	20
Replay buffer size of DQN	100000
Learning rate of DQN	0.0003
Initial exploration coefficient of DQN	0.6
Final exploration coefficient of DQN	0.005
Discount factor of DQN	0.99
Mini-batch size of DQN	32
Number of layers of DQN	5
Target net update rate of DQN	0.01
Num of parallel workers of DQN	32
Replay buffer size of TD3	100000
Learning rate of TD3	0.0001
Exploration coefficient of TD3	0.2
Discount factor of TD3	0.98
Mini-batch size of TD3	256
Number of layers of TD3	5
Target net update rate of TD3	0.01
Noise standard deviation of TD3	0.2
Noise clip coefficient of TD3	0.5