Biological Reinforcement Learning via Predictive Spacetime Encoding

Minsu Abel Yang
KAIST
Department of Bio and Brain Engineering
Department of Electrical Engineering
minsuyang@kaist.ac.kr

Jee Hang Lee
Sangmyung University
Department of Human-Centered AI
jeehang@smu.ac.kr

Sang Wan Lee
KAIST
Department of Bio and Brain Engineering
sangwan@kaist.ac.kr

Abstract

Recent advances in reinforcement learning (RL) have successfully addressed several challenges, such as performance, scalability, or sample efficiency associated with the use of this technology. Although RL algorithms bear relevance to psychology and neuroscience in a broader context, they lack biological plausibility. Motivated by recent neural findings demonstrating the capacity of the hippocampus and prefrontal cortex to gather space and time information from the environment, this study presents a novel RL model, called spacetime Q-Network (STQN), that exploits predictive spatiotemporal encoding to reliably learn highly uncertain environment. The proposed method consists of two primary components. The first component is the successor representation with theta phase precession implements hippocampal spacetime encoding, acting as a rollout prediction. The second component, called Q switch ensemble, implements prefrontal population coding for reliable reward prediction. We also implement a single learning rule to accommodate both hippocampal-prefrontal replay and synaptic homeostasis, which subserves confidence-based metacognitive learning. To demonstrate the capacity of our model, we design a task array simulating various levels of environmental uncertainty and complexity. Results show that our model significantly outperforms a few state-of-the-art RL models. In the subsequent ablation study, we showed unique contributions of each component to resolving task uncertainty and complexity. Our study has two important implications. First, it provides the theoretical groundwork for closely linking unique characteristics of the distinct brain regions in the context of RL. Second, our implementation is performed in a simple matrix form that accommodates expansion into biologically-plausible, highly-scalable, and generalizable neural architectures.

1 Introduction

One of the key challenges of reinforcement learning (RL) is to resolve environmental uncertainty and complexity with limited resources. In spite of the recent developments in machine learning to deal with complex problems [1–9], RL algorithms often have a hard time learning simple tasks that efficiently learnt by animals [10].

Preprint. Under review.

One common criticism is that a majority of learning algorithms based on deep learning have little biological relevance, though one can find some commonalities at the conceptual level [8, 9, 11–14]. Furthermore, our understanding of how the biological system implements learning is limited to simple tasks. A few working theories successfully explain the biological principles of each part of the brain, but they do not yet account for the ways in which interactions between different brain regions lead to optimal learning of the environment. Another limitation is that biological models for investigating high-level brain functions, such as context or memory, are not tightly linked with biological processes at the single-neuron level or accompany many biological constraints that limit their applicability.

To reconcile the discrepancies between RL algorithms and biological RL, here we explore biological processes that potentially contribute to efficient learning. Based on recent neural findings explaining the ability of the hippocampus and prefrontal cortex to efficiently glean space and time information from the environment, we propose a novel RL model to reliably learn a highly uncertain environment. This paper is structured as follows.

- First, we propose a novel spacetime encoding method by combining successor representation with theta phase precession. This module acts as a rollout prediction.
- Second, we implement prefrontal population coding for reliable reward prediction, called Q switch ensemble, that exploits the spacetime information of the episodes from the first module.
- Third, to further improve learning efficiency, we propose a single biological learning rule, accommodating both hippocampal-prefrontal replay and synaptic homeostasis. This learning rule subserves confidence-based metacognitive learning.
- Finally, we designed a task array simulating different levels of environmental uncertainty and complexity to demonstrate the capacity of our model, called spacetime Q-Network (STQN). We also ran systematical ablation analyses to show the unique contributions of each component to resolving task uncertainty and complexity.

2 Related Works and Neural Basis

The successor representation (SR) is a concept that represents the temporal proximity between the current, past, and future state of the environment [15]. As it captures the proximity of two events being relational, it has become a valuable tool in various fields, such as human reinforcement learning [16], robotic control [17], and visual navigation [18]. Recently, SR has gained significant attention in deep reinforcement learning [19, 20]. However, it has so far been used as a subordinate tool; for example as the realization of the posterior sampling [21], the utility for feature representation with a set of expectations [22], or the intermediate for inference using a nonparametric Dirichlet process mixture model [23]. In this regard, they have limited biological relevance.

Next, the hippocampal replay, specifically when awake, is the high-frequency neural activity inside the envelope of a sharp wave ripple (SWR) that occurs during resting or receiving a reward. SWR has long been considered as the bridge between short- and long-term memory consolidation [24]. There are several studies on this relationship; however, most of them have focused on relating the generative and terminative mechanism of the hippocampal replay to various synaptic properties [25, 26]. One study attempted to interpret the replay as iterative executions of Bellman backup and showed that their model can imitate some basic properties of the hippocampal replay [27]. However, this study neither examines the correspondence between their model and the prefrontal-hippocampal network nor explains the biological mechanisms of initiating, proceeding, and terminating replay.

Third, the synaptic homeostasis is a consistent synaptic operation needed to maintain the stability of neural networks undergoing rapid spike timing dependent plasticity (STDP), to properly store or process information [28]. This demand is because of the exponential property of the STDP network; the amounts of weight updates depend on the weight itself. This characteristic sometimes leads to destructive forgetting, masking of useful information the agent had learned, under the pathological conditions [29]. While a few studies have investigated the effect of synaptic homeostasis on learning [30–32], none of them are sufficient to carry real-time regulations during complex learning.

3 Methods

3.1 Spacetime Encoding

To build an effective encoder, we combined two concepts; successor representation (SR) and theta phase precession. To model the spatial properties of the place cell, successor representation is a rational choice as it can explain the irregular modulations of its receptive field [33] and be simulated by the spiking neural network [34]. Moreover, recent observations show that place cells can represent locations of the other animals enable the SR matrix to encode the target object [35, 36].

Next, we assume that the theta phase precession sharpens the spatial coding by adding temporal context to the place cell firing sequences. This hypothesis is reinforced by various experiments. For instance, it is observed that the hippocampal theta sequence tends to proceed toward the current goals [37]. The theta phase can divide the trajectory that an animal had passed and will pass into temporal bins [38] support this idea. Moreover, our assumption is assisted by an experiment that shows that pharmacological inhibitions of the lateral septum (LS), the major source of hippocampal theta wave, extinguish the ability to navigate without disturbing the intact place cell activities [39] or an experiment that shows that LS extends the egocentric firing of place cells to the entire environment located by the animal [40], which is an invaluable tool for planning and tackling further situations inside the complex spatial learning [41].

3.1.1 Successor Representation for Predictive Coding

Assume that there are N_H place cells c_H , which have the center of their receptive fields $\mathbf{X}_H \subset \mathbb{R}^{N_H \times 2}$. To represent the correlational relationship between the place cells, we used SR to represent the correlational relationship between the place cells. The SR matrix $\boldsymbol{M}$ encodes the expected and discounted future occupancy of the cell c'_H along a trajectory starting from the cell c_H :

$$\boldsymbol{M}(c_H, c'_H) = \mathbb{E}\left[\sum_{t=0} \gamma^t \mathbb{I}(c^t_H = c'_H) | c^0_H = c_H\right],$$

where $\mathbb{I}(c^t_H = c'_H) = 1$ if $c^t_H = c'_H$, otherwise 0 [15]. To update the SR matrix, we group the place cells whose receptive fields contain an external stimuli (e.g., a ball in Pong game) inside. Specifically, if the Euclidean distance between the center $\mathbf{x}_H$ of the place cell c_H and the location of the input stimulus $\mathbf{x}^t_I \in \mathbb{R}^2$ is smaller than the update distance l_U, the cell c_H is in the candidate set,

$$S^t_U = \left\{c_H \mid d(\mathbf{x}_H, \mathbf{x}^t_I) \leq l_U\right\},$$

If S^t_U and S^{t+1}_U are not empty, the update begins. For every cell pairs (c^t_H, c^{t+1}_H) from $c^t_H \in S^t_U$ and $c^{t+1}_H \in S^{t+1}_U$, we do TD online learning for those pairs [42]:

$$\hat{\boldsymbol{M}}_{t+1}(c^t_H, c_H) = \hat{\boldsymbol{M}}_t(c^t_H, c_H) + \eta\left[\mathbb{I}(c^t_H = c_H) + \gamma \hat{\boldsymbol{M}}_t(c^{t+1}_H, c_H) - \hat{\boldsymbol{M}}_t(c^t_H, c_H)\right],$$

where η is a learning rate, γ is a discount factor, respectively.

3.1.2 Theta Phase Precession for Spacetime Encoding

It is known that the theta phases inside the CA1 and PFC align in a regular manners [43]. To imitate this systematic modulation, we divide the single period of theta wave, between two peaks, into four equal parts. Then, we allocate the first, second, and third parts, respectively, to the past, current, and future place cells individually. The firing of PFC cells locates at the peak behind; therefore, the firing sequences of future cells become closest among the place cells. The sign of the cosine value is the criteria used to divide the past and future cells among all fired place cells. Assume that the input stimulus is located at $\mathbf{x}_I$ with a velocity of $\vec{v}$. After defining the vector from the center $\mathbf{x}_H$ of place cell c_H to $\mathbf{x}_I$, we can discern whether or not the moving stimulus had already passed the center $\mathbf{x}_H$.

$$\vec{v} = \overrightarrow{\mathbf{x}_I^{t-1}\mathbf{x}_I^t}, \qquad \cos\theta = \frac{\vec{v}\cdot\overrightarrow{\mathbf{x}_H\mathbf{x}_I}}{\|\vec{v}\|\|\overrightarrow{\mathbf{x}_H\mathbf{x}_I}\|}.$$

If the cosine value is negative, we can state that the cell c_H lies on the past trajectory of the input stimulus and vice versa. However, the dot product above can only separate past and future cells, not current cells. To compensate for this limitation, we define the current cell as cells closer to the input stimulus than the others. Mathematically, a set of current cells is:

$$S_C^t = \left\{c_H \mid d(\mathbf{x}_H, \mathbf{x}_I^t) \leq l_C\right\},$$

where l_C is a new parameter, the current distance.

3.1.3 SR Firing Rate for Online Learning

The agent needs a channel to transfer processed spatiotemporal information. One of the biologically-plausible candidates is the firing sequence from the presynaptic to the postsynaptic neurons, which is supported by the LS [44]. However, based on our knowledge, there are no trials or attempts to use SR as the basis for modeling real-time activity of place cells during complex learning, specifically, not as an alternative representation of the abstract "state." Therefore, we defined the firing rate for the SR matrix in online learning. From a set of current place cells, we define the firing rates of individual place cells. One way is to average the column vectors of current place cells in the SR matrix $\boldsymbol{M}$:

$$\mathbf{r}_H = \frac{1}{\left|S_C^t\right|}\sum_{c_H\in S_C^t}\tilde{\boldsymbol{M}}(:,c_H) = \frac{1}{\left|S_C^t\right|}\sum_{c_H\in S_C^t}\frac{\hat{\boldsymbol{M}}(:,c_H)}{\left\|\hat{\boldsymbol{M}}(:,c_H)\right\|_\infty}, \qquad \tilde{\mathbf{r}}_H = \frac{\mathbf{r}_H}{\|\mathbf{r}_H\|_\infty}, \tag{1}$$

where $\left|S_C^t\right|$ is the number of current place cells, $\tilde{\boldsymbol{M}}$ is normalized SR matrix obtained by dividing each column of the learned SR matrix by its maximum element. After comparing the rate vector $\mathbf{r}_H$ to the uniform random vector, we build $\mathbf{f}_H$, binary firing vector of all place cells.

3.2 Q Switch Ensemble

Q switch ensemble consists of two components: PFC cells and their population coding for reward prediction. In this study, our PFC cell model is based on the experiment that shows showing that individual PFC cells represent the difference between Q values of two options (left or right) to its firing rate while solving the Markov decision problem [45]. Based on this observation, we generalize PFC cell activity to binary Q classifiers, continuously updating its Q value and firing when it predicts a specific event will happen, of which firing rate is the function of Q value. Based on the recent evidence on dopaminergic neurons having different criteria of reward feedback [46] and the capability of PFC cells to encode spatial information [43], we represent the reward prediction in the form of population coding. Similar to the experimental data [43], each PFC cells inherit only the least spatial information, acting as a decision center. The PFC cells fire when it forecasts that the reward event will occur above the center of decision, then the agent compiles the votes to predict the location.

3.2.1 Phase Distance for an Effective Weight Representations

Based on our assumption of the theta phase precession, there exists a relative attenuation of the neuronal signal intensity of the early (past) spikes, when compared to the late (future) spikes. Therefore, we calculated the effective coefficients following cable theory [47] and spike timing dependent plasticity (STDP) [48]. Specifically, using the equal phase distances between past-current, current-future, and future-PFC, we proposed the following weighting scheme:

$$\mathbf{w}_E = \exp\left[-\Delta\tilde{T}\right] = \exp\left[-\Delta T/\tau\right],$$

where $\tilde{T}$ is a time distance divided by time constant τ , to convert the exponents to the integer.

3.2.2 PFC Cell Population Coding for Binary Q Classification

To use the transferred temporal contexts by the phase distance, we multiplied the binary firing vector $\mathbf{f}_H$ with the effective weight $\mathbf{w}_E$, element-wisely, producing effective firing vector $\mathbf{f}^*_H$. The PFC cells receive and combine $\mathbf{f}^*_H$ linearly through the weight matrix $\boldsymbol{W}_P \subset \mathbb{R}^{N_P \times N_H}$, making the Q values,

$$\mathbf{Q} = \boldsymbol{W}_P \mathbf{f}^*_H = \boldsymbol{W}_P \left(\mathbf{f}_H \odot \mathbf{w}_E\right). \tag{2}$$

The firing rate of each PFC cell is a sigmoid function of the Q value. To separate the activity of PFC cells maximally, we used the mid-range of $\mathbf{Q}$ as the offset. To compute the firing rate of PFC cells $\mathbf{r}_P$:

$$\bar{\mathrm{Q}} = \ (\min \mathbf{Q} + \max \mathbf{Q})/2, \qquad\qquad \mathbf{r}_P = \sigma(T^{-1} \cdot \big(\mathbf{Q} - \bar{\mathrm{Q}}\big)), \tag{3}$$

where T^{-1} is the inverse temperature of the sigmoid function σ .

3.2.3 PFC Cell Population Coding for Reward Prediction

We interpreted the learning as a population process, starting from the binary Q classifier. We assume that there are N_P PFC cells c_P, and their center of decision is $\mathbf{x}_P \subset \mathbb{R}^{N_P}$. If the PFC cell predicts that the reward event will occur above its center of decision, it fires. Otherwise, when it forecasts that the reward will occur below its center, it does not fire. We then define the level of confidence, which quantifies the amount by which the single cell is confident with its decision,

$$\mathbf{C}_Q = |\mathbf{r}_P - 0.5|\,.$$

Here, the reward prediction is formalized in population coding. If the PFC centered at x_P fires, it indicates that the reward event was predicted to occur anywhere above the center of decision of the PFC cell. To compress the continuous domain of the prediction, we choose points arbitrarily for discretization. In our simulations, we select the midpoints between every decision centers and build a set $\mathbf{x}_A \subset \mathbb{R}^{N_P+1}$. We refer to this set as the action set. The certainty $\mathbf{C}_Q$ now becomes a poll value. For a single action point x_A , the firing of the single PFC cell centered at x_P provides the following:

$$\hat{P}\left(x_A | f_P = 1\right) = \begin{cases} 0.5 + C_Q, & x_A \geq x_P \\ 0.5 - C_Q, & x_A < x_P \end{cases}, \qquad \hat{P}\left(x_A | f_P = 0\right) = \begin{cases} 0.5 - C_Q, & x_A \geq x_P \\ 0.5 + C_Q, & x_A < x_P \end{cases},$$

where $f_P \in \mathbf{f}_P$ is a Boolean variable whether or not this PFC cell mentioned above fired and $C_Q \in \mathbf{C}_Q$ is a corresponding individual confidence of the cell. After summing all the votes for the midpoints and then normalizing them, a discrete probability distribution over $\mathbf{x}_A$ can be obtained. The coordinate x^*_A becomes the target; if the agent locates below x^*_A, the agent moves up and so on.

$$P\left(x_A\right) = \sum_{c_P \in \mathbf{C}_P} \hat{P}\left(x_A | f_P\right), \qquad\qquad x^*_A = \underset{x_A \in \mathbf{x}_A}{\arg\max}\, P\left(x_A\right).$$

3.3 Learning with HPC-PFC Replay

In this study, we view the HPC-PFC replay as a recursive TD learning process. First, we establish a hypothesis reasoned from optogenetic manipulation, in which the sharp wave ripples modulate the synaptic weights [49]. Furthermore, this view is supported by the evidence that the relative progress of task learning systemically modulates the trend of replay [50], and it actively selects the highly rewarded event [51]. When the trial ends, the agent receives the reward information (e.g., the actual location at which the ball arrived). From this spatial information, every PFC cell receives its distinct rewards via the dopaminergic pathway [52]. Then, the learning process is initiated and proceeds recursively, underpinned by recurrent, interlamellar connections inside the hippocampus [53].

When the action point x_B is the closest to the one associated with reward (e.g., the location at which the ball arrived), the initial feedback signal for learning is simply the logical result of comparison; whether or not the point ball arrived was above the decision center of each PFC cell:

$$\mathbf{F} = u\left(x_B - \mathbf{x}_D\right),$$

where $u\left(\cdot\right)$ is the unit step function. Then, the changes of synaptic weights are linear to the difference between the feedback and the firing rate of each PFC cell [42]. Moreover, to present the mechanism by which various phenomena seemingly contribute to the learning, we designed two simple functions; $\mathrm{H}\left(\cdot\right)$ is the entropy learning rate function, $\gamma\left(\cdot\right)$ is the synaptic stability function (The details are in the supplementary material). With these utilities, the equation becomes:

$$\frac{\mathrm{d}\boldsymbol{W}_P}{\boldsymbol{W}_P} = \left[\mathrm{H}\left(\mathbf{r}_P\right) \odot \gamma\left(\mathbf{Q}\right) \odot \left[\mathbf{F} - \mathbf{r}_P\right]\right] \otimes \mathbf{f}_H^*, \tag{4}$$

where $\mathbf{f}_H^*$ is the effective firing vector. After the learning is initiated, the agent must prepare the next interactions not just a right in front, but from a few steps behind. We assume that the HPC-PFC network achieves this by relaying the Q values and attempting to narrow the gap between adjacent decisions via weight modulation. Following the statistics and properties of the hippocampal replay, we implement the recursive algorithm to simulate such a behavior (supplementary material).

4 Simulation Result

4.1 Architecture of STQN

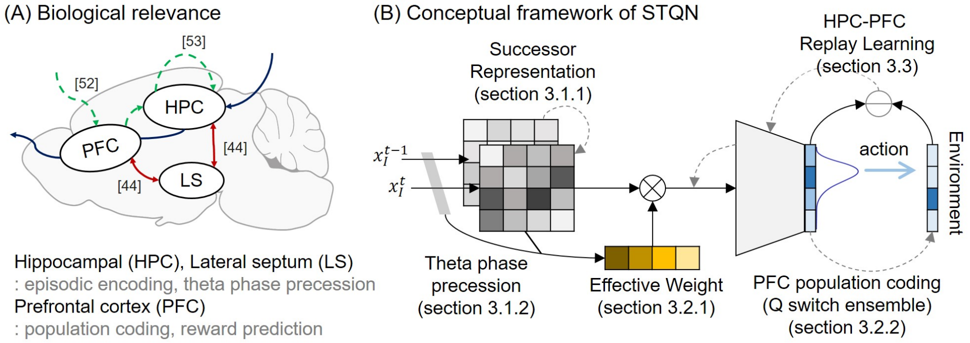

Figure 1: Concept Diagram of STQN

Our model consists of two components. First, the spacetime SR models the spatiotemporal encoding of the hippocampus by combining the successor representation as the spatial processor with the theta phase precession as the temporal channel. This module receives the current coordinate of the external input $\mathbf{x}_I$ and outputs the firing sequence of the relevant place cells. Next, it integrates the temporal order by multiplying the effective weight element-wise for the theta phase precession. Second, the Q switch ensemble receives the processed firing sequence and outputs the most probable reward location. After receiving the reward, the recursive TD learning updates the weights $\boldsymbol{W}_P$, with respect to the difference between the decision certainty of each PFC cell and the relative feedback signal.

4.2 Novel task design varying the degree of uncertainty and difficulty level

We designed our task based on the classic Atari game Pong. To incorporate both the structural complexity and state-transition uncertainty into the task, we added three walls to test the learning performance of various agents. First, we doubled the width and height of the Pong table, four times larger than the OpenAI gym Pong [54], without modifying the parameters of the agent behavior. Next, we varied the amount of uncertainty in ball bouncing, by making the trajectory of the ball difficult to predict. After the ball bounces off the wall, a uniform noise is added to the angle of the velocity vector. This noise is bounded by the "variation angle" θ_V. Using θ_V, we can incorporate the uncertainty in a controllable, natural manner. Additionally, the environment is modulated in a dynamic manner, from regular oscillation to the chaotic Brownian motion. Lastly, we implemented the opponent with an ideal agent. For this, we assumed that the agent has complete access to the environmental structure (e.g., the exact location where the ball will arrive, given the current position and velocity vector of the ball). At every single frame, the opponent (ideal agent) receives the future location of arrival, but it contains additive white Gaussian noise. The task difficulty can be varied by manipulating the probability with which the opponent makes mistakes (standard deviation of this noise σ_w). The difficulty level allows us to assess the capacity of the agent since the lesser mistakes the opponent makes, the longer the rally length becomes (Details are in supplementary material).

We set two hyperparameters, the "variation angle" θ_V and the standard deviation of opponent noise σ_w. Subsequently, we built six tasks by combining two hyperparameter sets; $\hat{\sigma}_w = \{0.1, 0.2\}$, and $\theta_V = \{0°, 10°, 30°\}$ where $\hat{\sigma}_w$ is the normalized standard deviation of the opponent, σ_w divided by the half-height of the Pong table. For simplicity, we hereinafter refer to the conditions $\hat{\sigma}_w = \{0.1, 0.2\}$ as EASY and HARD, and $\theta_V = \{0°, 10°, 30°\}$ as LOW, MID, and HIGH uncertainties, respectively.

4.3 Performance Comparison

To examine the way in which various deep RL agents deal with the aforementioned dynamic environments, we chose DQN [55], IQN [56], PPO [57], A3C-LSTM [12] as the representative comparison models. Notably, we found that our model is the only one that achieved guaranteed performances over the broad spectrum of uncertainty and complexity (Table 1) without the requirement of fine-tuning.

Table 1: Benchmark test (5M episodes)

	Score						Rally length (100 Frames)					
Uncertainty	LOW		MID		HIGH		LOW		MID		HIGH	
Difficulty	EASY	HARD	EASY	HARD	EASY	HARD	EASY	HARD	EASY	HARD	EASY	HARD
DQN	-11.2	-10.5	-10.7	-8.09	-14.1	-18.1	0.74	1.01	0.76	1.23	0.80	0.80
IQN	-14.9	-15.8	-15.7	-17.64	-14.0	-16.7	0.86	1.06	0.84	0.91	1.01	1.04
PPO	-18.5	-19.2	-18.6	-19.51	-18.0	-19.2	0.61	0.62	0.61	0.62	0.68	0.69
A3C-LSTM	-9.4	-14.2	-10.0	-14.66	-11.0	-15.1	1.15	1.33	1.18	1.35	1.17	1.30
STQN*	**15.2**	**10.2**	**14.1**	**9.1**	**14.1**	**9.2**	**3.99**	**6.36**	**4.51**	**6.59**	**4.44**	**6.84**

There are a few key points to note in RL learning. First, the score tended to decrease as the task difficulty increased, except for DQN. This is because DQN is unstable and it generally failed, but the minor (<15%) succeed to learn (supplementary material). Second, PPO failed to learn completely and this deficiency correlates significantly to task difficulty (supplementary material). The third is the reliability of the meta-RL agent (A3C-LSTM). To test its long-term learning, we increased the training episodes to 20M; however, we found that the effects are marginal (supplementary material). We used the rally length to quantify the capacity of the RL agents (see Table 1). When the capacity of the learning agent was comparable to that of the ideal agent (opponent), the rally length was observed increase abruptly, since there were no mistakes to terminate the rally. We were able to identify this effect in our model (>100%), regardless of the task uncertainty. It clearly demonstrated the high capacity of our model and also implied that the policy of our model is near-optimal with the scoreboard. However, this effect was not very pronounced in the other models (<20%) and tended to disappear under high uncertainty conditions, thereby indicating incomplete learning in these models.

4.4 Role of Predictive Spacetime Encoding in Learning (Ablation Study)

To examine the unique roles of individual components, we conducted the ablation study. More details on analyses and discussions are provided in Supplementary material. First, we quantified the degree of contribution of theta phase precession on task performance. Next, we examined how place cell encodes information and supports complex prediction by modulating the two hyperparameters; the update distance l_U and the current distance l_C based on successor representation. In here, we introduce two different measures, the "error" E and the "accuracy" R_C (see Figure 3). In-depth performance analysis and discussions about biological implications are provided in the supplementary material.

4.4.1 Theta Phase Precession

The theta phase precession implements the idea that the temporal distance controls the relative contribution of a single spike to the remaining process. This modulation allows the agent to take a predictive "vector decision": one should act differently if the situation were predicted to change in the foreseeable future. As learning proceeds, the learning agent becomes sensitive to the environmental changes by encoding the "context". However, the effectiveness of this strategy is constrained by the task uncertainty and the task difficulty.

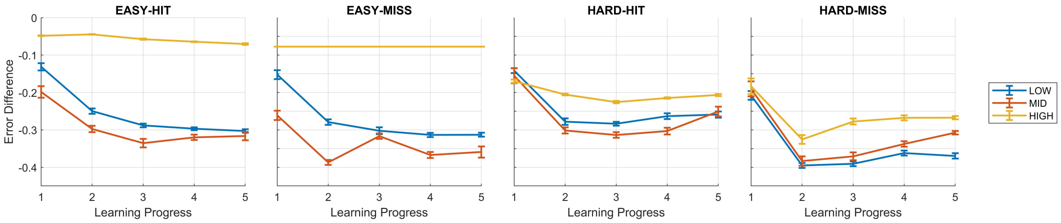

Figure 2: Theta Phase Precession
HIT corresponds to the case when the agent successfully passes the ball
MISS corresponds to the case when the agent makes mistakes and receives negative reward

By comparing the time courses of learning of the agent with and without theta phase precision, we found its significant contribution on performance (figure 2). We also found that the terminal error difference depends on the uncertainty and the difficulty level of the task. Especially, the effect of task uncertainty on the performance wears off in the high task difficulty conditions (Hard).

4.4.2 Successor Representation

Update distance l_U is the hyperparameter that represents the number of future states that a single current state should encode. The higher the uncertainty, the harder it is to predict future states from the given current state; consequently, the time required to achieve a certain level of accuracy also increases. This leads to the two distinct properties; (1) the slope of the error with respect to l_U becomes more negative proportionally (Figure 3A Amplitude), and (2) the error difference between the agent with the normal l_U and the small value increases (Figure 3A Offset) with the decrease in the task uncertainty.

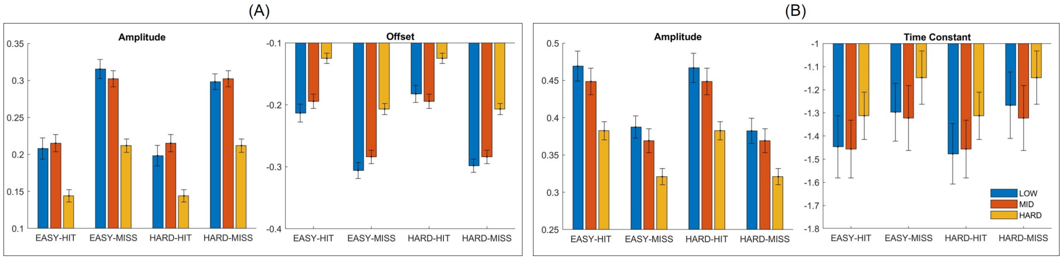

Figure 3: (A) Update Distance and (B) Current Distance
Fitting with (A) $\Delta E = A \cdot \sqrt{l_U^* - l_U} + C$ and (B) $\Delta R_C = A \cdot \exp\left[\tau_C \cdot l_C\right]$

If we view the SR matrix as an encoder and the Q switch ensemble as a decoder, then the current distance l_C is translated into the factor modulating the amount of information that the encoder can process. This provides two predictions; (P1) the agent needs more capacity in highly uncertain environment, and (P2) the measure R_C, the ratio of "correct" PFC cells, decreases when the amount of uncertainty increases with the same l_C. According to (P1), the difference of R_C between the agents with the changed and normal l_C, ΔR_C, should decrease more rapidly (Figure 3B Time Constant) since l_C^* that makes ΔR_C=0 decreases, where l_C^* is the minimal l_C that can support sufficient number of symbols (lower bound of l_C). According to (P2), the value at infinitesimal l_C increases when the amount of task uncertainty decreases (Figure 3B Amplitude).

5 Summary and discussion

Motivated by recent neural findings about space-time information processing in the hippocampus and prefrontal cortex, we propose a novel spacetime RL framework, called spacetime Q-Network (STQN), that reliably learns highly uncertain environment. The implications of our study are as follows. First, our spacetime encoding allows us to examine the hypothesis that hippocampal place cells with theta phase precession facilitate predictive encoding and learning. Second, we demonstrate that the group of binary Q classifiers as a proxy for PFC cells can predict the complex physical processes [46]. Third, the proposed learning rule underlines a direct relevance of the homeostatic synaptic plasticity to Q-learning [45]. In summary, our study lays the theoretical groundwork for integrating unique properties of separate brain regions in the context of RL. In addition, our formulation of biological processes in the form of a simple matrix offers valuable insights into biologically-plausible and highly-scalable neural architecture designs.

Broader Impact

This work is based on the computer simulation, so there is no direct impact on animals or humans. Therefore, there are no potential ethical issues. This work has the following potential impacts on society: The design of novel deep learning architecture. We proposed a biologically plausible RL model integrating unique characteristics of distinct brain regions. With taking into account a brain's capacity encoding the space and time information, the proposed model exploits the reliable learning the highly uncertain and complex environment. The model was implemented in a simple matrix form, so it is expected to accommodate expansion into highly-scalable and generalizable neural architecture. The training process is based on simulations in the virtual (game) environment. Hence, there is no psychological or physical harm to humans inflicted by our model.

References

[1] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[2] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

[3] Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Deep spatial autoencoders for visuomotor learning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 512–519. IEEE, 2016.

[4] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, pages 12498–12509, 2019.

[5] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pages 4754–4765, 2018.

[6] Suraj Srinivas and Francois Fleuret. Knowledge transfer with Jacobian matching. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4723–4731, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[7] Tingwu Wang and Jimmy Ba. Exploring model-based planning with policy networks. *arXiv preprint arXiv:1906.08649*, 2019.

[8] Sébastien Racanière, Théophane Weber, David Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adria Puigdomenech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al. Imagination-augmented agents for deep reinforcement learning. In *Advances in neural information processing systems*, pages 5690–5701, 2017.

[9] Vitchyr Pong, Shixiang Gu, Murtaza Dalal, and Sergey Levine. Temporal difference models: Model-free deep rl for model-based control. *arXiv preprint arXiv:1802.09081*, 2018.

[10] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*, 2017.

[11] Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7, 2019.

[12] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[13] Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868, 2018.

[14] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[15] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

[16] Ida Momennejad, Evan M Russek, Jin H Cheong, Matthew M Botvinick, Nathaniel Douglass Daw, and Samuel J Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680–692, 2017.

[17] Jingwei Zhang, Jost Tobias Springenberg, Joschka Boedecker, and Wolfram Burgard. Deep reinforcement learning with successor features for navigation across similar environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2371–2378. IEEE, 2017.

[18] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.

[19] Samuel J Gershman. The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, 38(33):7193–7200, 2018.

[20] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in neural information processing systems*, pages 4055–4065, 2017.

[21] David Janz, Jiri Hron, Przemysław Mazur, Katja Hofmann, José Miguel Hernández-Lobato, and Sebastian Tschiatschek. Successor uncertainties: exploration and uncertainty in temporal difference learning. In *Advances in Neural Information Processing Systems*, pages 4509–4518, 2019.

[22] Li Kevin Wenliang and Maneesh Sahani. A neurally plausible model for online recognition and postdiction in a dynamical environment. In *Advances in Neural Information Processing Systems*, pages 9641–9652, 2019.

[23] Tamas Madarasz and Tim Behrens. Better transfer learning with inferred successor maps. In *Advances in Neural Information Processing Systems*, pages 9026–9037, 2019.

[24] Alison R Preston and Howard Eichenbaum. Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, 23(17):R764–R773, 2013.

[25] Tatsuya Haga and Tomoki Fukai. Recurrent network model for learning goal-directed sequences through reverse replay. *Elife*, 7:e34171, 2018.

[26] Rich Pang and Adrienne L Fairhall. Fast and flexible sequence induction in spiking neural networks via rapid excitability changes. *eLife*, 8:e44324, 2019.

[27] Marcelo G Mattar and Nathaniel D Daw. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, 21(11):1609–1617, 2018.

[28] Kimberly M Huber. Synaptic homeostasis: quality vs. quantity. *Nature neuroscience*, 21(6):774, 2018.

[29] Boaz Styr and Inna Slutsky. Imbalance between firing homeostasis and synaptic plasticity drives early-phase alzheimer's disease. *Nature neuroscience*, 21(4):463–473, 2018.

[30] Jochen Triesch, Anh Duong Vo, and Anne-Sophie Hafner. Competition for synaptic building blocks shapes synaptic plasticity. *Elife*, 7:e37836, 2018.

[31] Milton Llera-Montero, João Sacramento, and Rui Ponte Costa. Computational roles of plastic probabilistic synapses. *Current opinion in neurobiology*, 54:90–97, 2019.

[32] Lee Susman, Naama Brenner, and Omri Barak. Stable memory with unstable synapses. *Nature communications*, 10(1):1–9, 2019.

[33] Kimberly L Stachenfeld, Matthew M Botvinick, and Samuel J Gershman. The hippocampus as a predictive map. *Nature neuroscience*, 20(11):1643, 2017.

[34] Johanni Brea, Alexisz Tamás Gaál, Robert Urbanczik, and Walter Senn. Prospective coding by spiking neurons. *PLoS computational biology*, 12(6), 2016.

[35] Jacob LS Bellmund, Peter Gärdenfors, Edvard I Moser, and Christian F Doeller. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415):eaat6766, 2018.

[36] David B Omer, Shir R Maimon, Liora Las, and Nachum Ulanovsky. Social place-cells in the bat hippocampus. *Science*, 359(6372):218–224, 2018.

[37] Andrew M Wikenheiser and A David Redish. Hippocampal theta sequences reflect current goals. *Nature neuroscience*, 18(2):289–294, 2015.

[38] Anoopum S Gupta, Matthijs AA Van Der Meer, David S Touretzky, and A David Redish. Segmentation of spatial experience by hippocampal theta sequences. *Nature neuroscience*, 15(7):1032, 2012.

[39] Kevin A Bolding, Janina Ferbinteanu, Steven E Fox, and Robert U Muller. Place cell firing cannot support navigation without intact septal circuits. *Hippocampus*, 30(3):175–191, 2020.

[40] David Tingley and György Buzsáki. Transformation of a spatial map across the hippocampal-lateral septal circuit. *Neuron*, 98(6):1229–1242, 2018.

[41] György Buzsáki and David Tingley. Space and time: The hippocampus as a sequence generator. *Trends in cognitive sciences*, 22(10):853–869, 2018.

[42] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[43] Mark C Zielinski, Justin D Shin, and Shantanu P Jadhav. Coherent coding of spatial position mediated by theta oscillations in the hippocampus and prefrontal cortex. *Journal of Neuroscience*, 39(23):4550–4565, 2019.

[44] Fatemeh Khakpai, Mohammad Nasehi, Ali Haeri-Rohani, Akram Eidi, and Mohammad Reza Zarrindast. Septo-hippocampo-septal loop and memory formation. *Basic and clinical neuroscience*, 4(1):5, 2013.

[45] Bilal A Bari, Cooper D Grossman, Emily E Lubin, Adithya E Rajagopalan, Jianna I Cressy, and Jeremiah Y Cohen. Stable representations of decision variables for flexible behavior. *Neuron*, 103(5):922–933, 2019.

[46] Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, pages 1–5, 2020.

[47] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544, 1952.

[48] Guo-qiang Bi and Mu-ming Poo. Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual review of neuroscience*, 24(1):139–166, 2001.

[49] Josef HLP Sadowski, Matthew W Jones, and Jack R Mellor. 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] Justin D Shin, Wenbo Tang, and Shantanu P Jadhav. Dynamics of awake hippocampal-prefrontal replay for spatial learning and memory-guided decision making. *Neuron*, 104(6):1110–1125, 2019.

[51] Frédéric Michon, Jyh-Jang Sun, Chae Young Kim, Davide Ciliberti, and Fabian Kloosterman. Post-learning hippocampal replay selectively reinforces spatial memory for highly rewarded locations. *Current Biology*, 29(9):1436–1444, 2019.

[52] Torben Ott and Andreas Nieder. Dopamine and cognitive control in prefrontal cortex. *Trends in Cognitive Sciences*, 2019.

[53] Sunggu Yang, Sungchil Yang, Thais Moreira, Gloria Hoffman, Greg C Carlson, Kevin J Bender, Bradley E Alger, and Cha-Min Tang. Interlamellar ca1 network in the hippocampus. *Proceedings of the National Academy of Sciences*, 111(35):12919–12924, 2014.

[54] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[55] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[56] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. *arXiv preprint arXiv:1806.06923*, 2018.

[57] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.