MODELING THE ROLE OF STRIATUM IN CONTEXT DEPENDENT REWARD BASED TASKS

A Project Report

submitted by

SABYASACHI SHIVKUMAR

in partial fulfilment of the requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY

&

MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

MAY 2017

THESIS CERTIFICATE

This is to certify that the thesis titled **Modeling the Role of Striatum in Context Dependent Reward Based Tasks**, submitted by **Sabyasachi Shivkumar**, to the Indian Institute of Technology Madras, Chennai for the award of the degree of **Bachelor of Technology & Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. V. Srinivasa Chakravarthy

Research Guide

Professor
Dept. of Biotechnology
IIT-Madras, 600 036

Prof. Harishankar Ramachandran

Research Co-Guide

Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai
Date: 2nd May 2017

ACKNOWLEDGEMENTS

The path towards completing this thesis has been challenging and filled with many memories. This journey would not have been possible without the support of so many people. I would like to thank Indian Institute of Technology for giving me the opportunity and the requisite skill set for completing this project. I would like express my gratitude towards Prof. V. Srinivasa Chakravarthy for his continuous guidance and motivation. Working under him gave me the chance to experience the ups and downs of research and taught me the drive the see the problem through. I would also like to thank Dr. Nicolas P. Rougier, my mentor during the summer of 2016 in Bordeaux, for constantly giving me encouragement and opportunity to solve various problems. I would also like to thank Indo-French project 'Basal Ganglia at Large' for giving me a chance to work in a cross-cultural lab and learn many skills as part of the experience. I thank Prof. Harishankar Ramachandran for inspiring and supporting me to complete this project.

This project wouldn't have possible without the support of my parents, family, teachers, labmates and all my friends. My teachers have been a constant source of inspiration and have given me the skills required to tackle various challenges. I would specially like to thank Vignesh Muralidharan, PhD student for helping me with this project. My labmates and friends have been of constant help and have provided me with numerous interesting conversations. Lastly, my parents and family have given me continuous support and I thank them for helping me at every step of the way.

ABSTRACT

KEYWORDS: Striatum, Basal Ganglia, Context Dependant Learning, Striosomes and Matrisomes, Self Organizing Maps, Modular Reinforcement Learning, Stochastic Multi Context Tasks, Bayesian Model

Basal Ganglia circuit is an important subcortical system of the brain thought to be responsible for reward based learning. Striatum, the largest nucleus of the Basal Ganglia, serves as an input port that maps cortical information. Microanatomical studies show that the striatum is a mosaic of specialized input-output structures called Striosomes and regions of the surrounding matrix called the Matrisomes. We have developed a computational model of the striatum using layered self-organizing maps to capture the centre-surround structure seen experimentally and explain its functional significance. We believe that these structural components could build representations of state and action spaces in different environments. The striatum model is then integrated with other components of Basal Ganglia, making it capable of solving reward based tasks. We have proposed a biologically plausible mechanism of action based learning where the striosome biases the matrisome activity towards a preferred action. Several studies indicate that the striatum is critical in solving context dependant problems. We build on this hypothesis and the proposed model exploits the modularity of the striatum to efficiently solve such tasks. We have also looked at stochastic multi context tasks and developed a Bayesian theoretical model to solve these problems. The striatum model is also catered to solve these tasks. We have shown that the striatal model matches the theoretical model for low stochasticity in the environment and could be thought of as a neural implementation of the theoretical model.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **1**

ABSTRACT **2**

LIST OF TABLES **5**

LIST OF FIGURES **6**

ABBREVIATIONS **10**

NOTATIONS **11**

OUTLINE OF THE THESIS **12**

SECTION I **13**

1. INTRODUCTION **14**

2. METHODS **17**

Modeling the microanatomy of the Striatum

Reinforcement learning in Basal Ganglia

Reinforcement Learning in Environments with Multiple Contexts

Using the striatal modularity to solve modular reinforcement learning tasks

3. RESULTS **29**

Modeling the microanatomy of the Striatum

Reinforcement Learning in a Single Context Gridworld Task

Reinforcement Learning in a Multi-Context Grid-world Problem

4. DISCUSSION **39**

Striosome-Matrisome Dynamics with their Dopaminergic Projections

Mapping Representations to Action Primitives

Contextual Learning and Striatal Modularity

Behavioral Observations

SECTION II **43**

5. STOCHASTIC MULTI CONTEXT TASK **44**

6. METHODS **45**

Bayesian Model Formulation

Theoretical Model

Stochastic Reward Based Task Learning in Striatum

Exploiting the Striatal Modularity for solving context dependent tasks

7. RESULTS **59**

Performance of theoretical model on T-Maze tasks

Solving Stochastic Reward Based Tasks using the Striatum Model

Comparing the Theoretical and Neural model

8. DISCUSSION **69**

CONCLUSION **71**

REFERENCES **72**

LIST OF PUBLICATIONS BASED ON THESIS **76**

LIST OF TABLES

Table 1: Parameter values for single context and multi context tasks

Table 2: Parameter values for cue based decision tasks

LIST OF FIGURES

Fig. 1 **A)** A schematic of the striosome-matrisome centre surround mapping in the striatum. The red structures represent the striosomes and the surrounding green structures represent the matrisomes. **B) A** Schematic of the layered SOM structure modeling the striosomes and matrisomes. The Strio-SOM (Red) represents the striosomes and the Matri-SOM (Green) represents the matrisomes; each Strio-SOM neuron has projections to the surrounding Matri-SOM neurons.

Fig. 2 Schematic Diagram for the Basal Ganglia model. The arrows indicate connections and their type. The component sizes are proportional to their dimensions. The feedback connections from the thalamus project the information about the action chosen back to the striatum

Fig. 3 Schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

Fig. 4 **A)** Activity of the Strio-SOM and the corresponding Matri-SOM neurons for different actions in a state. The centre map shows only the activity of the Strio-SOM in the absence of any action and the other four maps in the corners show the activity of the Strio-SOM and the four possible Matri-SOM neurons that best respond to the particular action. **B)** Same as (A) for another state. **C)** Combined activity for all the action pairs in (A). Shows one configuration of the centre-surround mapping. **D)** Combined activity for all the action pairs in (B). Shows another configuration of the centre-surround mapping.

Fig. 5 **A)** Schematic of the grid-world used in the task. A goal is located at the top right corner of the grid **B)** State value map estimated by the agent at different spatial locations. We can see that the state value peaks at the goal location. **C)** Plot of the Number of Steps

taken by the agent in each episode averaged across 50 independent sessions. We see that the number of steps reduces as the agent learns across episodes.

Fig. 6 **A)** Schematic of the gridworld used in the task. A goal is switched between the top left and bottom right corner every 150 episodes. **B)** State value map estimated by the agent at different spatial locations across different contexts. We can see that the state value peaks at the goal location corresponding to the context. **C)** Environment Feature Signal maps estimated by the agent at different spatial locations across different contexts. We can see that the state value peaks at the goal location corresponding to the context. **D)** Modules chosen by the agent at different episodes. We can see that the module chosen switched with change in context indicating that the agent is able to identify the context it is currently present in.

Fig. 7 **A)** Plot of Number of Steps taken by the single module agent in each episode averaged across 50 independent sessions. We see that the agent needs to relearn after each context switch **B)** Plot of Number of Steps taken by the multi module agent in each episode averaged across 50 independent sessions. We see that the agent efficiently switches modules after each context switch **C)** Peak number of steps needed to reach the goal after a context switch averaged across 50 sessions. **D)** Number of episodes for the number of steps required to reach the goal to go below a certain threshold averaged across 50 sessions **E)** Peak value for the average number of steps needed to reach the goal after a context switch. The experimental values have been adapted from (Brunswik 1939) **D)** Number of episodes for the average number of steps required to reach the goal to go below a certain threshold. The experimental values have been adapted from (Brunswik 1939)

Fig. 8 Flowchart depicting steps to solve a stochastic multi context task.

Fig. 9 **A)** Schematic of the centre surround mapping of seen in the striatum. The red centre represents the striosomes and the surround green neurons represent the matrisomes. **B)** Schematic of the layered SOM architecture where each neuron in the Strio-SOM (Red) projects to the neurons in the Matri-SOM (Green) **C)** Schematic diagram of the Striatum

model where the arrows indicate the connections and their types.

Fig. 10 Schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

Fig. 11 **A)** Demonstration of change in performance with varying reward magnitudes (Figure adapted from (Lloyd and Leslie 2013)). **B)** Performance of our model on the varying reward magnitude task **C)** Demonstration of change in performance with varying reward probabilities (Figure adapted from (Lloyd and Leslie 2013)). **D)** Performance of our model on the varying reward probability task

Fig. 12 **A)** Percentage of trials where the animal chooses the arm which is non-profitable for the first 24 trials and becomes profitable following that. (Figure adapted from (Lloyd and Leslie 2013)). **B)** Performance of the model on the task described in **A**. We see that the model shows similar trends where the definite reward tasks show faster reversal learning. **C)** Percentage of trials where the animal chooses the arm which was rewarding before 24 trials following which both arms are not rewarded (Figure adapted from (Lloyd and Leslie 2013)). **D)** Performance of the model on the task described in **C** where the model shows similar trends as the definite reward task show faster unlearning.

Fig. 13 **A)** Schematic of the cue based decision making task where the agent has to choose between the two shapes shown in the screen and each shape has a different probability of reward associated with it. **B)** Percentage of correct responses averaged over 25 sessions for 200 trials. **C)** Mapping of the action inputs forms a centre-surround structure when we view the combined activity of the Matri-SOM for all action inputs **D)** Ratio of choosing response 1 with associated probability P1 wrt to the sum P1+P2. The model follows a similar trend to the experimental plot adapted from (Pasquereau, Nadjar et al. 2007)

Fig. 14 **A)** Probability of context 1 estimated by the theoretical model. **B)** Probability of context

1 estimated by the neural model.

Fig. 15 **A)** Percentage of correct responses by the theoretical model. **B)** Percentage of correct responses by the neural model.

Fig. 16 Schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

ABBREVIATIONS

BG	Basal Ganglia
SOM	Self-Organizing Maps
TD	Temporal Difference
STN	Subthalamic nucleus
GPe	Globus Pallidus external
GPi	Globus Palidus internal
RL	Reinforcement Learning

NOTATIONS

x	Variable x (italicized)
$\boldsymbol{x}$	Vector x (bold, italicized)
dim($\boldsymbol{x}$)	Dimension of vector x
[n]	Spatial Location of neuron n
x→y	Projection from x to y

OUTLINE OF THE THESIS

The thesis is organized into several chapters most of which are assigned to two main sections which describe the two main models of striatum. Section I of the thesis deals with the non-stochastic version of reward based learning. The first chapter of this section introduces the earlier studies for understanding the structure and function of the striatum and basal ganglia. The next chapter lays down the architecture of the proposed model for dealing with these tasks. This is followed by the results of section I which sets up a testbench problem and demonstrates the model performance in building representations, solving both the stationary and non-stationary versions of the reinforcement learning problems and finally shows some experimental results which validate the model. The last chapter of section I is the discussion chapter which presents an analysis of the various predictions made by the model and their support with existing literature.

Section II of the thesis mainly deals with the stochastic reward based tasks. Since the inherent complexity of these tasks is much higher than their non-stochastic counterparts, the approach to solve these problems is presented in a different route. The first chapter of this section introduces the problem. This is followed by the Methods chapter for this section. This chapter first lays out a detailed formulation of the problem. Next, a Bayesian model is formulated which provides a bound on the optimal performance in such tasks. This Bayesian model is modified to relax some of its strong assumptions to give rise to a theoretical model catering to stochastic multi-context tasks. This is followed by a neural model of the striatum which borrows some base features from the model in section I but presents an alternate variation to the striatal model. The next chapter consists of the results which first demonstrate the theoretical model followed by the demonstration of the neural model. These two models are then compared with each other on various problem formulations. The final chapter of this section presents a short discussion for the various ideas in this section.

The last part of the thesis concludes all the results in the previous sections and then describes the possible future route for the model.

SECTION I

CHAPTER 1

INTRODUCTION

In order to understand the role of the striatum within the basal ganglia (BG) circuit, it is essential to understand the rich and complex microcircuitry of this structure. It is well known that the striatum has a modular architecture, containing specialised input-output structures called the 'striosomes' and regions of the surrounding matrix called the 'matrisomes' (Graybiel, Flaherty et al. 1991). The striosomes are known to receive limbic inputs and send their projections to the substantia nigra pars compacta, a midbrain dopaminergic nucleus, whereas the matrisomes mostly receive sensorimotor and associative inputs and project to downstream BG nuclei (Graybiel, Aosaki et al. 1994). The cortico-striatal connectivity seems to show a divergence property, where there is spread of connections coming from the cortex to the striatum followed by a convergence at the level of the globus pallidus (GP) (Graybiel, Aosaki et al. 1994). There have also been suggestions that the striatum constructs low dimensional representations of the cortical states via the cortico-striatal projections (Bar-Gad, Havazelet-Heimer et al. 2000; Bar-Gad, Morris et al. 2003). Indirect evidence for this comes from experiments which indicate hebbian like learning in cortico-striatal projections (Charpier and Deniau 1997). Therefore the striatum has the cellular and molecular machinery to possibly construct such reduced representations of cortical states. These facts about striatal microanatomy lead us to believe that the striatum could build representations for several state and action spaces.

Anatomically the striosome-matrisome complex has a center-surround structure (Graybiel, Flaherty et al. 1991), and the proposed computational architecture for the striatum is inspired by this fact. Studies investigating the projection of prefrontal areas to the striosomes show specificity to certain cortical areas (Eblen and Graybiel 1995). These

cortical projections to anterior striosomes are mostly from frontal regions like the orbitofrontal cortex, anterior insula and the anterior cingulate cortex (Eblen and Graybiel 1995) which could very well represent the task or state space (Wilson, Takahashi et al. 2014). The matrisome which receives more sensorimotor information would well represent the action space (Flaherty and Graybiel 1994). In classical reinforcement learning (RL) literature, the expected reward signal in a given state is called the value function (Sutton and Barto 1998). The striosomes are known to have reciprocal projections to both the Ventral Tegmental Area (VTA) and the Substantia Nigra pars compacta (SNc) and thus would receive the prediction error signal from these midbrain nuclei, which can serve as a reinforcement signal that aids in the computation of the state value function (Granger 2006). On the other hand the action representations perhaps evolve at the level of matrisomes, and get mapped on to action primitives at the level of GPi (Pasquereau, Nadjar et al. 2007). Thus using the reward information from the environment and the representations built in the striatum, the BG can learn to perform reward based decision making tasks.

This functional organization and the modularity of the striatum has been hypothesized to perform context dependent tasks (Amemori, Gibb et al. 2011). Mulitple spatio-temporal contexts could then be mapped to different striatal modules, leading to decomposition of goals (context information) a facet of modular reinforcement learning (Kalmár, Szepesvári et al. 1999). We then consider the selection of the module appropriate to a given context to be driven by a responsibility signal, which is a function of the uncertainty in the environment. Uncertainty in the environment from previous approaches has been represented by reward variance (Balasubramani, Chakravarthy et al. 2015). Since change in context leads to increased uncertainty, reward variance could help identify this change.

In the current study, we propose a hierarchical self-organizing structure to model the striosome-matrisome compartments. Self-organizing maps (SOMs) have been used to represent high-dimensional information in 2-D sheets of neurons (Kohonen 1998). The striosome and the matrisome layers are both modelled as a double SOM layer, consisting

of Strio-SOM and Matri-SOM respectively, where a single Strio-SOM neuron has projections to the surrounding Matri-SOM neurons. The activity of the Matri-SOM is mapped to action primitives via the direct and indirect pathways of the BG to perform action selection. The reward information from the environment is utilized by the Strio-SOM to bias the surrounding Matri-SOM activity towards a preferred action. This provides a biologically plausible way of carrying out action based Q-learning (Sutton and Barto 1998) and is a novel feature of our model. This model has been tested on standard grid-world problems.

The model has been extended to cater to problems with varying contexts (changing reward locations). Different striatal modules map different contexts and Tonically Active Neurons (TANs) (Apicella 2007) aid in module selection This selection is driven by the risk (reward variance) in the environment which is used to calculate the responsibility signal (Amemori, Gibb et al. 2011) for a particular module. We have tested this model on grid-world problems with varying reward distributions and the model is able to solve these problems efficiently.

CHAPTER 2

METHODS

Modeling the microanatomy of the Striatum

We have proposed an architecture consisting of two layers of SOMs as a method for mapping centre-surround structures seen in the striatum (Fig. 1A). This architecture is used to model striosomes and matrisomes which map the state space and action space respectively.

The first layer called Strio-SOM models the striosomes and maps the state space. The second layer activated by the Strio-SOM is called the Matri-SOM which models the matrisomes and maps the action space (Fig. 1B).

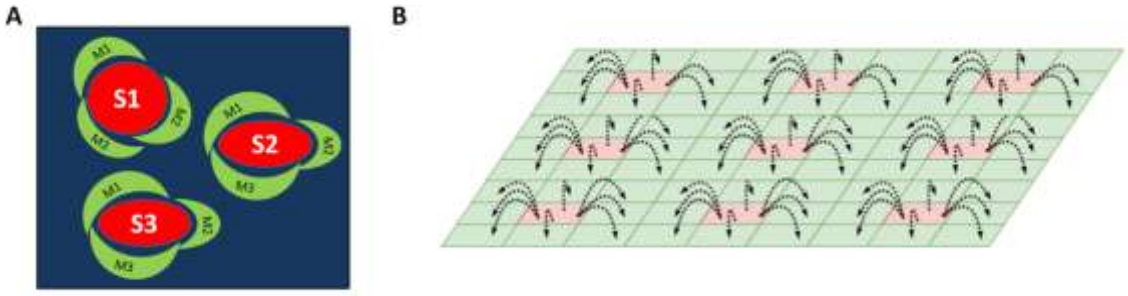

Fig. 1 **A)** A schematic of the striosome-matrisome centre surround mapping in the striatum. The red structures represent the striosomes and the surrounding green structures represent the matrisomes. **B) A** Schematic of the layered SOM structure modeling the striosomes and matrisomes. The Strio-SOM

(Red) represents the striosomes and the Matri-SOM (Green) represents the matrisomes; each Strio-SOM neuron has projections to the surrounding Matri-SOM neurons.

In order to map the state space, we have a Strio-SOM of size m_1 x n_1. If **s** is a state vector, the weights of the Strio-SOM (W^S) are of dimension m_1 x n_1 x dim(***s***), where dim(***s***) stands for the dimension of the state vector ***s***. Similarly, to map the action space, we have a Matri-SOM of size m_2 x n_2. If ***a*** is an action vector, the weights of all the Matri-SOMs (W^M) are of dimension m_1 x n_1 x m_2 x n_2 x dim(**a**) as each neuron in the Strio-SOM is connected to a Matri-SOM.

The activity for a neuron n in the Strio-SOM for a state input ***s*** is given in Eq. 1.

$$X^S_{[n]} = \exp\left(\frac{-\| W^S_{[n]} - s \|_2^2}{\sigma_S^2}\right) \quad \text{Eq. 1}$$

where [n] represents the spatial location of the neuron n and σ_S controls the sharpness of the neuron activity. The complete activity of the Strio-SOM (X^S) is the combination of individual activity of all the neurons. The neuron with the highest activity ("winner") for a state **s** is denoted by n_s^*.

Similarly, the activity for a neuron n in the Matri-SOM for an action input *a* in a state s is given in Eq. 1.

$$X^M_{[n_s^*][n]} = \exp\left(\frac{-\| W^M_{[n_s^*][n]} - a \|_2^2}{\sigma_M^2}\right) \quad \text{Eq. 2}$$

where σ_M controls the sharpness of the neuron activity. The complete activity of the Matri-SOM corresponding to neuron n_s^* ($X^M_{[n_s^*]}$) is the combination of individual activity of all the neurons in the Matri-SOM corresponding to n_s^*. The neuron with the highest activity ("winner") for an action **a** in a state **s** is denoted as $n_{s,a}^*$.

The weight of a neuron **n** in the Strio-SOM for a state input **s** is updated according to the following rule

$$W^{S}_{[n]} \leftarrow W^{S}_{[n]} + \eta_S . \exp(\frac{-\|[n]-[n_s^*]\|_2^2}{\sigma_S^2}).(s - W^{S}_{[n]}) \quad \text{Eq. 3}$$

The weight of neuron **n** in the Matri-SOM for an action input **a** in a state **s** is updated according to the following rule:

$$W^{M}_{[n_s^*][n]} \leftarrow W^{M}_{[n_s^*][n]} + \eta_M . \exp(\frac{-\|[n]-[n_{s,a}^*]\|_2^2}{\sigma_M^2}).(a - W^{M}_{[n_s^*][n]}) \quad \text{Eq. 4}$$

Reinforcement learning in Basal Ganglia

The striatum model developed in the previous section was useful in developing representations for states and actions. In this section, we incorporate the striatum model in a BG model and apply the model to standard reinforcement learning tasks. A schematic diagram of the model is given in Fig. 2.

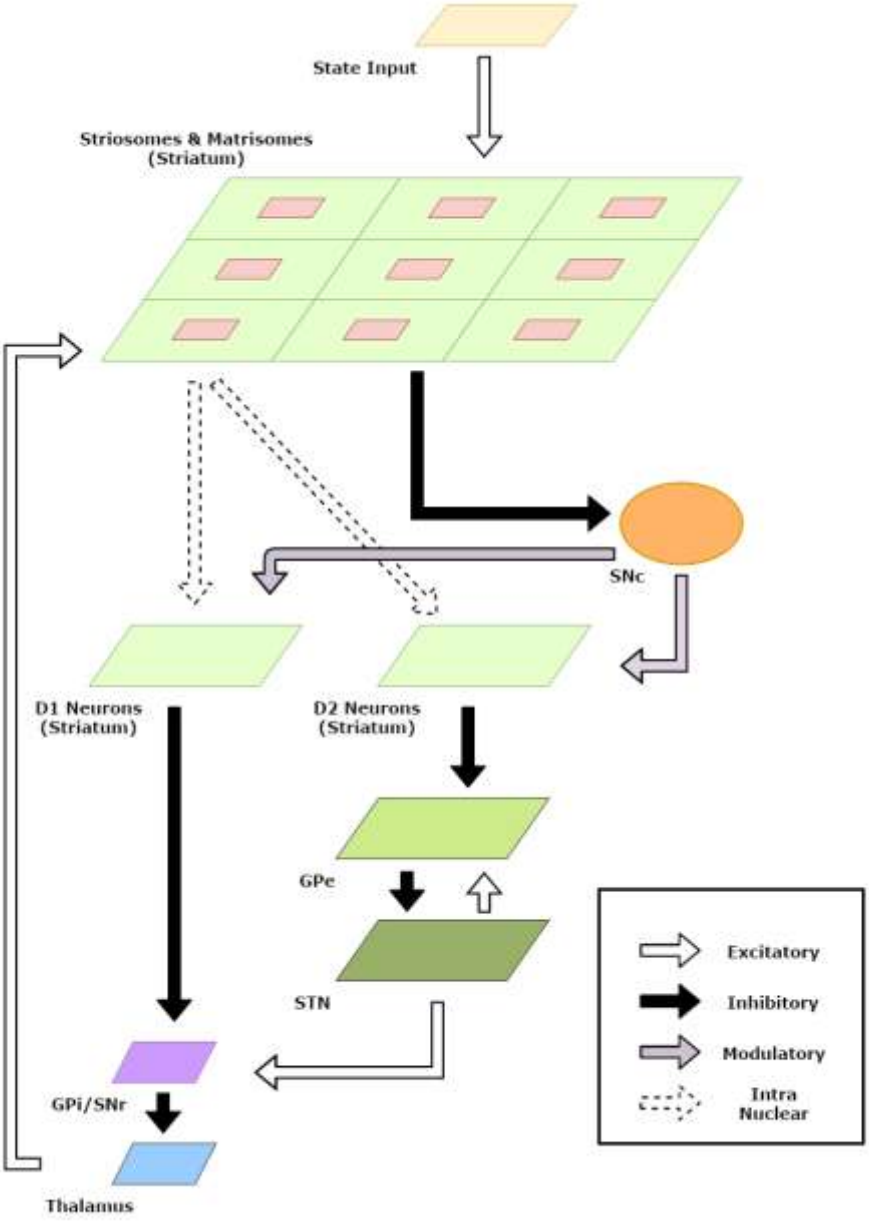

Fig. 2 Schematic Diagram for the Basal Ganglia model. The arrows indicate connections and their type. The component sizes are proportional to their dimensions. The feedback connections from the thalamus project the information about the action chosen back to the striatum

Let us assume that the animal is in a state **s**. The activity of the striosomes gives us the representation of the state in the striatum. In our model, the activity of the striosomes is given as the activity of the neurons in the Strio-SOM where the activity of a single neuron is given by Eq. 1. Thus the activity is of dimension m_1 x n_1.

This activity of the Strio-SOM projects to the SNc and represents the value for the state **s** in our model (Eq. 5). These weights from the striatum to SNc ($W^{Str \rightarrow SNc}$) are trained using the signal from SNc which is representative of Temporal Difference (TD) error (δ) (Eq. 6). The TD error is calculated as $\delta = r + \gamma V(s') - V(s)$ where s' is the new state after taking action **a** (Eq. 19), r is the reward obtained and γ is the discount factor.

$$V(s) = \sum_{\forall n} W^{Str \rightarrow SNc}{}_{[n]} X^{S}{}_{[n]} \qquad \textbf{Eq. 5}$$

$$\Delta W^{Str \rightarrow SNc}{}_{[n]} = \eta^{Str \rightarrow SNc} \delta X^{S}{}_{[n]} \qquad \textbf{Eq. 6}$$

where V(s) represents the value for state **s**, $\eta^{Str \rightarrow SNc}$ is the learning rate for $W^{Str \rightarrow SNc}$.

The representation for the various actions the agent in state **s** can perform is given by the activity of the matrisomes surrounding the corresponding striosome neuron for the state. In our model, this is given by the activity of the neurons of the Matri-SOM corresponding to the neuron with the highest activity in the Strio-SOM (n_s^*) where the activity of a single neuron in the Matri-SOM is given in Eq. 2. Thus the activity is of dimension m_2 x n_2. The action input **a** is given as feedback input from the thalamus to the striatum (Fig. 2).

The activity of Matri-SOM neurons is further tuned by the connections between the neurons in the Strio-SOM and the Matri-SOM ($W^{S \rightarrow M}$). These connections are also trained using TD error as above using the Matri-SOM activity for the action (**a**) chosen, as follows:

$$Y^{M}{}_{[n_s^*][n]} = \alpha X^{M}{}_{[n_s^*][n]} + (1-\alpha) W^{S \rightarrow M}{}_{[n_s^*][n]} X^{S}{}_{[n_s^*]} \qquad \textbf{Eq. 7}$$

$$\Delta W^{S\to M}{}_{[n_S^*][n]} = \eta^{S\to M}\delta X^{M}{}_{[n_S^*][n]} \qquad \text{Eq. 8}$$

where α controls the contribution of the action and the lateral connections to the activity of the Matri-SOM and $\eta^{Str\to SNc}$ is the learning rate for $W^{S\to M}$. Choosing a low value of α and low initial weights for $W^{S\to M}$ ensures that the activity is driven by the action representation initially and then driven by the lateral weights once the $W^{S\to M}$ have been trained sufficiently. The Strio-SOM/Matri-SOM weights ($W^{S\to M}$) are thresholded and normalized by their sum to ensure stability.

The matrisomes activity is projected to the direct and indirect pathways by the D1 and D2 neurons of the striatum. In our model, the Matri-SOM activity is modulated by a value difference signal (δ_V). If the agent goes from state $\mathbf{s^{(1)}}$ to $\mathbf{s^{(2)}}$, δ_V is the difference between the value of the two states, i.e. $\delta_V = V(\mathbf{s^{(1)}})-V(\mathbf{s^{(2)}})$.

This value difference signal modulates the switching between the direct and indirect pathways and is thought to be represented by the dopamine signaled by SNc (Chakravarthy and Balasubramani 2015). The activity of the D1 and D2 neurons are given in Eq. 9 and Eq. 10.

$$Y^{D1}{}_{[n]} = f(\lambda_{D1}\delta_V)Y^{M}{}_{[n_s^*][n]} \qquad \text{Eq. 9}$$

$$Y^{D2}{}_{[n]} = f(\lambda_{D2}\delta_V)Y^{M}{}_{[n_s^*][n]} \qquad \text{Eq. 10}$$

where f is a tanh nonlinearity and λ_{D1} and λ_{D2} are the gains of the D1 and D2 neurons respectively. The indirect pathway consisting of the GPe and STN is modeled as network of coupled non-linear oscillators. The dynamics of these oscillators is highly dependent on the input, which constitutes the projections from the D2-expressing neurons of the striatum. The dynamics of GPe is given below:

$$\tau^{GPe}\frac{dX^{GPe}{}_{[n]}}{dt} = -X^{GPe}{}_{[n]} - \varepsilon^{GPe}W^{GPe\to GPe}{}_{[n]}Y^{GPe}{}_{[n]} + W^{STN\to GPe}{}_{[n]}Y^{STN}{}_{[n]} + Y^{D2}{}_{[n]} \qquad \text{Eq. 11}$$

$$Y^{GPe}_{[n]} = \tanh(\lambda^{GPe} X^{GPe}_{[n]})$$ **Eq. 12**

where $W^{GPe \to GPe}$ are the lateral weights within the GPe, ϵ^{GPe} is the connection strength, $W^{STN \to GPe}$ are the connections between STN and Gpe, and λ^{GPe} is a non-linear scaling parameter.

The STN layer in the model exhibits correlated activity suppressed for high striatal input, and uncorrelated oscillatory activity for low striatal inputs(Chakravarthy and Balasubramani 2015). The uncorrelated oscillations of the STN are a key source of exploration for the agent. The dynamics of STN is given below:

$$\tau^{STN} \frac{dX^{STN}_{[n]}}{dt} = -X^{STN}_{[n]} + \varepsilon^{STN} W^{STN \to STN}_{[n]} Y^{STN}_{[n]} - W^{GPe \to STN}_{[n]} Y^{GPe}_{[n]}$$ **Eq. 13**

$$Y^{STN}_{[n]} = \tanh(\lambda^{STN} X^{STN}_{[n]})$$ **Eq. 14**

where $W^{STN \to STN}$ are the lateral weights within the STN, ϵ^{STN} is the connection strength, $W^{GPe \to STN}$ are the connections between Gpe and STN and λ^{STN} is a non-linear scaling parameter.

The D1 neurons of the striatum and the STN neurons project to the GPi leading to the convergence of the direct and indirect pathways in GPi. In the model, the number of GPi neurons equals number of actions (= dim(**a**)). The weights $W^{D1 \to GPi}$ and $W^{STN \to GPi}$ map the corresponding activities of D1 striatum and STN onto the GPi. The Matri-SOM

activity (Y^{D1}) corresponding to the chosen action (**a**) (which comes via feedback) is used to train the two sets of weights, $W^{D1\rightarrow GPi}$ and $W^{STN\rightarrow GPi}$ using Hebb's rule. The output of GPi neurons are computed according to (eqn. 15), and the update for the weights $W^{D1\rightarrow GPi}$ and $W^{STN\rightarrow GPi}$ are done according to (Eq. 16 and Eq. 17).

$$Y^{GPi}_{[n']} = W^{D1\rightarrow GPi}_{[n'][n]} Y^{D1}_{[n]} - W^{STN\rightarrow GPi}_{[n'][n]} Y^{STN}_{[n]} \quad \textbf{Eq. 15}$$

$$\Delta W^{D1\rightarrow GPi}_{[n][n']} = \eta^{D1\rightarrow GPi} Y^{D1}_{[n]} X^{GPi}_{[n']} \quad \textbf{Eq. 16}$$

$$\Delta W^{STN\rightarrow GPi}_{[n][n']} = \eta^{STN\rightarrow GPi} Y^{STN}_{[n]} X^{GPi}_{[n']} \quad \textbf{Eq. 17}$$

The neurons in the GPi project to the thalamus. In our model, action selection takes place in the thalamus, following the integrator-race model (Bogacz 2007) with thalamic neurons having self-exciting and mutually inhibiting interactions. The thalamic neuron that first crosses a threshold value (Y_{thresh}) determines the action. The thalamic neurons have low initial random activity which converge to a high activity for the chosen action and low values for the others. The dynamics of thalamic neurons is given as:

$$\dot{Y}^{Thal}_{[n]} = \sum_{n' \in Thal} W^{Thal}_{[n][n']} Y^{Thal}_{[n']} + Y^{GPi}_{[n]} \quad \textbf{Eq. 18}$$

$$a = \{[n] : Y^{Thal}_{[n]} > Y_{thresh}\} \quad \textbf{Eq. 19}$$

This action (**a**) chosen is carried out and the reward (r) is obtained. The action chosen is also projected back to the striatum to obtain the activity. Both the action and the reward are used for updates in Eq. 6, Eq. 8 and Eq. 16.

Reinforcement Learning in Environments with Multiple Contexts

Standard Reinforcement Learning techniques are suited for problems where the environment is stationary. However, in some tasks the environment suddenly changes and the agent has to adopt a policy suitable for the new environment. In such a case, the agent identifies the context either using a cue which is representative of the context or using its experience in the preceding trials. One of the techniques to solve problems of the second category is the modular RL framework. In this method, the agent allocates separate modules to separate contexts. Each of the modules has its own copy of the environment in a particular context, represented by an environment feature signal (ρ). This copy is used to generate a responsibility signal, denoted by λ, which indicates how close the current context is to the one represented by the module. Thus by identifying the module with the highest responsibility signal we can follow the policy developed in that module to solve the problem in an efficient manner.

Using the striatal modularity to solve modular reinforcement learning tasks

The striatum model developed above forms the basic module capable of solving simple RL tasks. Multiple such modules in the striatum could then be exploited to tackle multi-context tasks using modular RL framework. A schematic of this extended model is given in Fig. 3.

We believe that context selection happens at the level of the striatum and the context modulated activity is projected to the downstream nuclei of the BG for further processing. Thus, for clarity, we have expanded the intra-nuclear activity of the striatum in the model schematic (Fig. 3). Supposing there are K modules denoted by M_1, M_2 … ,M_K. We now define the weights and activities in the previous sections for each module and denote $\{M_i\}$ with each term associated with module M_i. Thus, for a module *m*, the

following variables undergo a change in notation: $X^{S} \rightarrow X^{S,\{m\}}$ (Eq. 1), $X^{M} \rightarrow X^{M,\{m\}}$ (Eq. 2), $W^{S} \rightarrow W^{S,\{m\}}$ (Eq. 3), $W^{M} \rightarrow W^{M,\{m\}}$ (Eq. 4), $V(s) \rightarrow V^{\{m\}}(s)$ (Eq. 5), $W^{Str \rightarrow SNc} \rightarrow W^{Str \rightarrow SNc,\{m\}}$ (Eq. 6), $X^{M} \rightarrow X^{M,\{m\}}$ (Eq. 7), $W^{S \rightarrow M} \rightarrow W^{S \rightarrow M,\{m\}}$ (Eq. 8).

We propose that in addition to the value of the state **s**, the activity of the Strio-SOM also projects to the SNc to represent the environment feature signal ($\rho^{\{m\}}$). The weights of these projections are denoted as $W_{\rho}^{Str \rightarrow SNc,\{m\}}$ and are trained using the signal from SNc which is representative of context prediction error (δ^{*}). The corresponding equations are given in Eq. 20 and Eq. 21. The context prediction error is calculated as $\delta^{*} = r - \rho^{\{m\}}(s)$

$$\rho^{\{m\}}(s) = \sum_{\forall n} W_{\rho}^{Str \rightarrow SNc,\{m\}}{}_{[n]} X^{S,\{m\}}{}_{[n]} \quad \textbf{Eq. 20}$$

$$\Delta W_{\rho}^{Str \rightarrow SNc,\{m\}}{}_{[n]} = \eta_{\rho}^{Str \rightarrow SNc} \delta^{*} X^{S,\{m\}}{}_{[n]} \quad \textbf{Eq. 21}$$

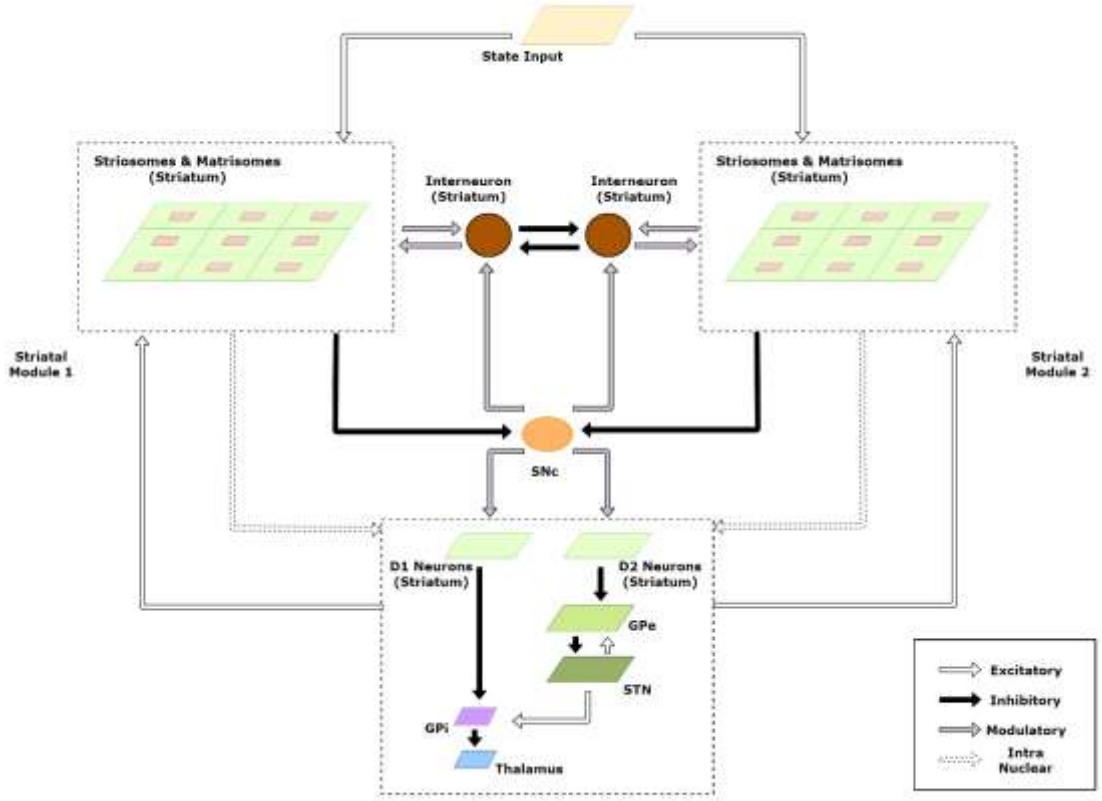

Fig. 3 Schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

We believe that the selection of the appropriate module for the context is guided by the striatal interneurons. In our model, the activity of the interneurons represents the responsibility signal for each module, denoted by $\lambda^{\{m\}}$ for module m. In a given state **s**,

the inter-neurons compete among themselves and the one with the highest λ chooses the module responsible for deciding the action in that state. Let the winning module in the state **s** be denoted by m^*. This module guides the projection to the direct and indirect pathway (Eq. 9 and Eq. 10) as given in Eq. 22 and Eq. 23.

$$Y^{D1}_{[n]} = f(\lambda_{D1}\delta_V)Y^{M,\{m^*\}}_{[n_s^*][n]} \quad \text{Eq. 22}$$

$$Y^{D2}_{[n]} = f(\lambda_{D2}\delta_V)Y^{M,\{m^*\}}_{[n_s^*][n]} \quad \text{Eq. 23}$$

Following this stage, the equations governing the signal flow are same as that in the previous section. The weight updates in the striatum are however done only to the module m^*.

The dynamics of the responsibility signal is given in Eq. 24

$$\dot{\lambda} = -\lambda - \alpha_\lambda(\delta^*)^2 \quad \text{Eq. 24}$$

where α_λ controls the influence of context prediction error on the responsibility signal and δ^* is the context prediction error.

CHAPTER 3

RESULTS

Modeling the microanatomy of the Striatum

We use a grid-world problem as a preliminary benchmark to test our model. The grid is of size 10 x 10 and the agent can take one of the four actions- Up, Down, Right and Left in a state. A reward is placed at one of the corners of the maze. The goal of the task is to make the model (agent) learn to reach this goal. We use the terms model and agent interchangeably in these sections since we use the model as a reinforcement learning agent in the various tasks. We used a 10 x 10 Strio-SOM to represent the state space and a 3 x 3 Matri-SOM, associated with each of the Strio-SOM neurons, for representing the action space.

In order to develop these representations, we make the agents explore various states and choose random actions in those states. Following this, we look at the neuron with the highest activity in the Strio-SOM for a particular state and the neurons with the highest activity for each action in the corresponding Matri-SOM for that state (Fig. 4A, Fig. 4B). Upon looking at the combined Matri-SOM activity for all the actions, we observed predominantly two different configurations of the centre-surround mapping (Fig. 4C, Fig. 4D).

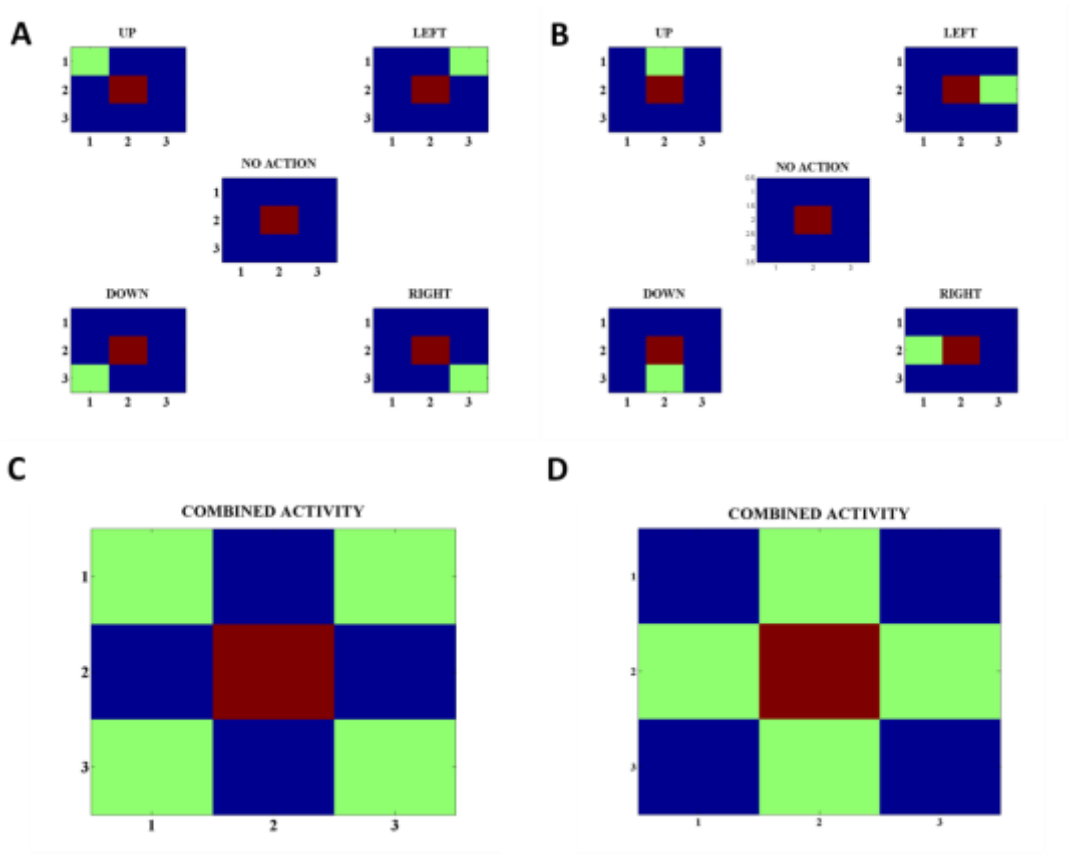

Fig. 4 **A)** Activity of the Strio-SOM and the corresponding Matri-SOM neurons for different actions in a state. The centre map shows only the activity of the Strio-SOM in the absence of any action and the other four maps in the corners show the activity of the Strio-SOM and the four possible Matri-SOM neurons that best respond to the particular action. **B)** Same as (A) for another state. **C)** Combined activity for all the action pairs in (A). Shows one configuration of the centre-surround mapping. **D)** Combined activity for all the action pairs in (B). Shows another configuration of the centre-surround mapping.

Reinforcement Learning in a Single Context Gridworld Task

The goal was placed at the top right of the grid as seen in Fig. 5A. The agent received a

reward of +20 when it reached the goal and 0 for all the other steps. At the beginning of an episode, the agent started at random and the episode ended when the agent reached the goal or when it reached the upper limit on number of steps allowed in the episode. The agent carried on the task for 150 episodes. This procedure was carried out for 50 independent sessions and the mean number of steps to reach the goal in a particular episode was plotted in Fig. 5C. The heat map of the state value function (Eq. 5) estimated by the agent at different spatial locations is given in Fig. 5B and peaks at the goal location. This combined with the fact that number of steps reduces as the episodes progress indicate that the agent is able to learn the single context task. The various parameter values for this task are given in Table1.

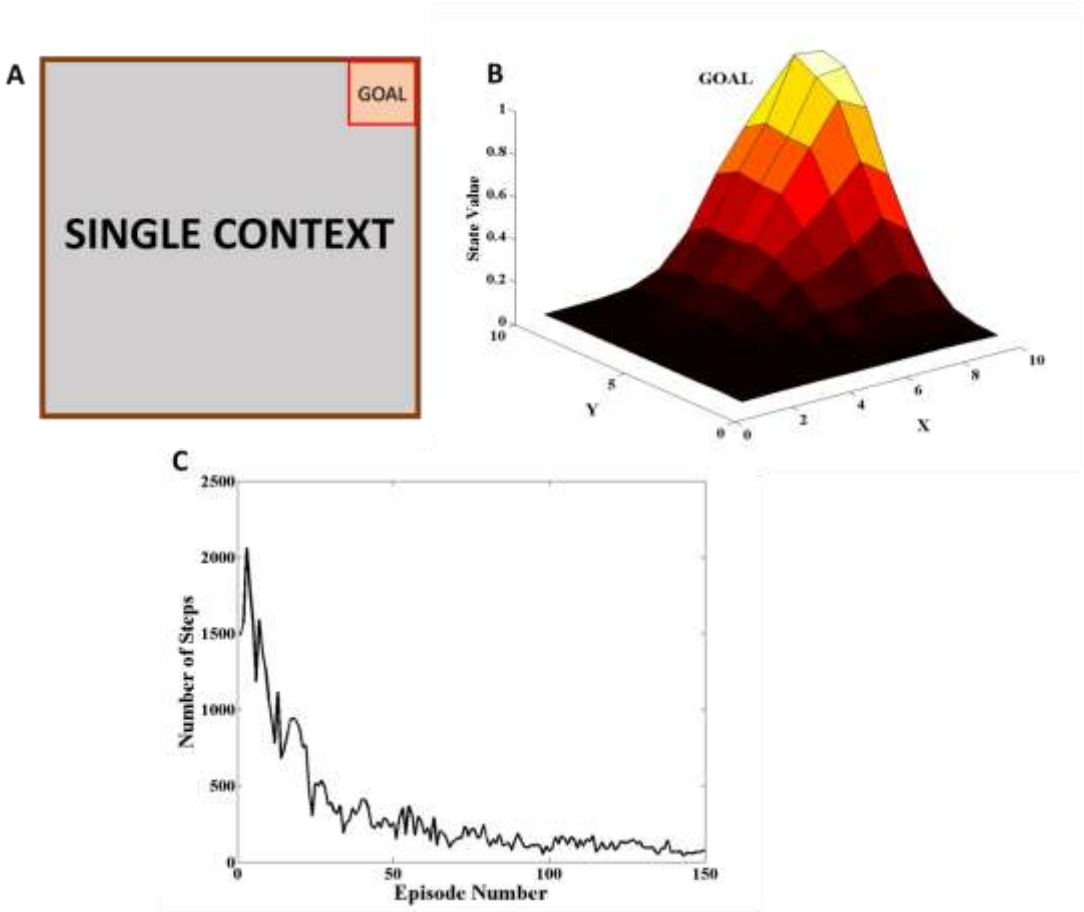

Fig. 5 **A)** Schematic of the grid-world used in the task. A goal is located at the top right corner of the grid **B)** State value map estimated by the agent at different spatial locations. We can see that the state value peaks at the goal location. **C)** Plot of the Number of Steps taken by the agent in each episode averaged across 50 independent sessions. We see that the number of steps reduces as the agent learns across episodes.

Reinforcement Learning in a Multi-Context Grid-world Problem

In the multi context grid-world tasks, the agent had to reach the goal like the previous section but the goal location changed after a certain number of episodes. The goal was present either at the top right corner or at the bottom left corner as shown in Fig. 6A. The goal was switched to the other location after 150 episodes. The task was carried out in 50 independent sessions with each session containing 900 episodes. The parameters used have the same values as given in Table 1. Fig. 6B shows the value function (Eq. 5) heat map and Fig. 6C shows the environment feature signal (Eq. 20) heat map estimated by the agent for the two contexts. We can observe that the agent is able to learn these values for both the contexts. Fig. 6D shows the context chosen by the agent in different episodes and we can observe that the agent is able to switch context in sync with the switch in reward distribution. These results illustrate that the agent is able to successfully identify the context it is presently in, and complete the corresponding grid-world task.

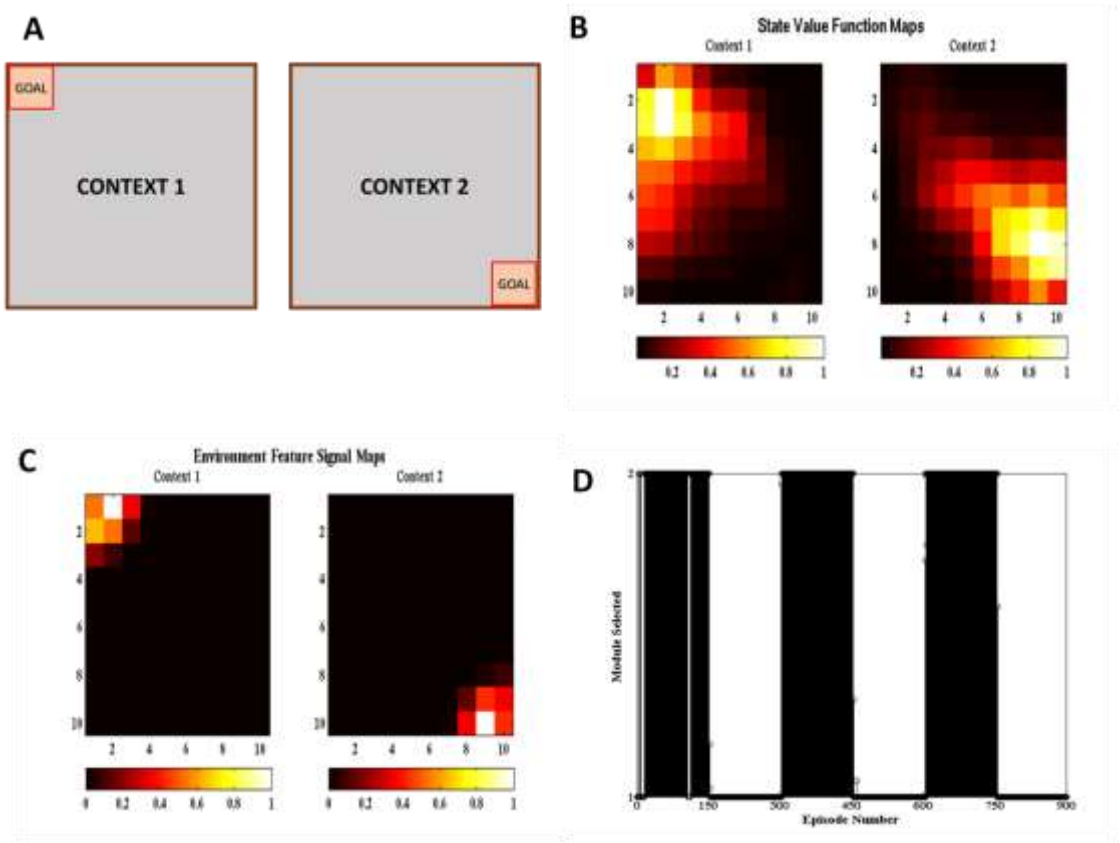

Fig. 6 **A)** Schematic of the gridworld used in the task. A goal is switched between the top left and bottom right corner every 150 episodes. **B)** State value map estimated by the agent at different spatial locations across different contexts. We can see that the state value peaks at the goal location corresponding to the context. **C)** Environment Feature Signal maps estimated by the agent at different spatial locations across different contexts. We can see that the state value peaks at the goal location corresponding to the context. **D)** Modules chosen by the agent at different episodes. We can see that the module chosen switched with change in context indicating that the agent is able to identify the context it is currently present in.

Table 1: Parameter values for single context and multi context tasks

Parameter	Value	Parameter	Value
Strio-SOM Dimension (m_1xn_1)	**10x10**	**Matri-SOM Dimension (m_2xn_2)**	**3x3**
σ_S	**0.01**	$\boldsymbol{\sigma_M}$	**0.1**
η_S	**0.4**	$\boldsymbol{\eta_M}$	**0.4**
γ	**0.97**	$\boldsymbol{\eta^{Str\rightarrow SNc}}$	**0.1**
α	**0.1**	$\boldsymbol{\eta^{S\rightarrow M}}$	**0.1**
λ_{D1}	**1**	$\boldsymbol{\lambda_{D2}}$	**-1**
τ^{GPe}	**3**	$\boldsymbol{T^{STN}}$	**1**
ϵ^{GPe}	**-0.01**	$\boldsymbol{\epsilon^{STN}}$	**0.01**
λ^{GPe}	**3**	$\boldsymbol{\lambda^{STN}}$	**3**
$\eta^{D1\rightarrow GPi}$	**0.01**	$\boldsymbol{\eta^{STN\rightarrow GPi}}$	**0.01**
Y_{thresh}	**1**	$\boldsymbol{\eta_\rho^{Str\rightarrow SNc}}$	**0.1**
α_λ	**0.8**		

The average number of steps required by the agent to reach the goal for each episode across 50 sessions is given in Fig. 7B. The same plot for an agent with only a single module is given in Fig. 7A. We can clearly see that the learning is more efficient for multi module agent as compared to the single module case. In order to quantify this improvement, we use two values to measure the agent's performance after a context switch. These are the peak number of steps to reach the goal after a context switch and the number of episodes for the number of steps needed to go below a certain threshold.

We calculate these two values for each context switch in a session. These values are averaged across sessions and presented in Fig. 7C and Fig. 7D resp. In both cases we see that the multi module agent is better than the single module agent for solving the task. We use these measures to compare the model against experimental data in (Brunswik 1939). Since we only have the average performance across sessions available in the reference, we calculate the corresponding values from our model and present these for the single module, multi module and the experimental case in Fig. 7E and Fig. 7F respectively. We can observe that multi module results have a similar trend to the experimental results as compared to the single module model, thus demonstrating that the BG could be using the modular architecture of the striatum to solve context switching tasks.

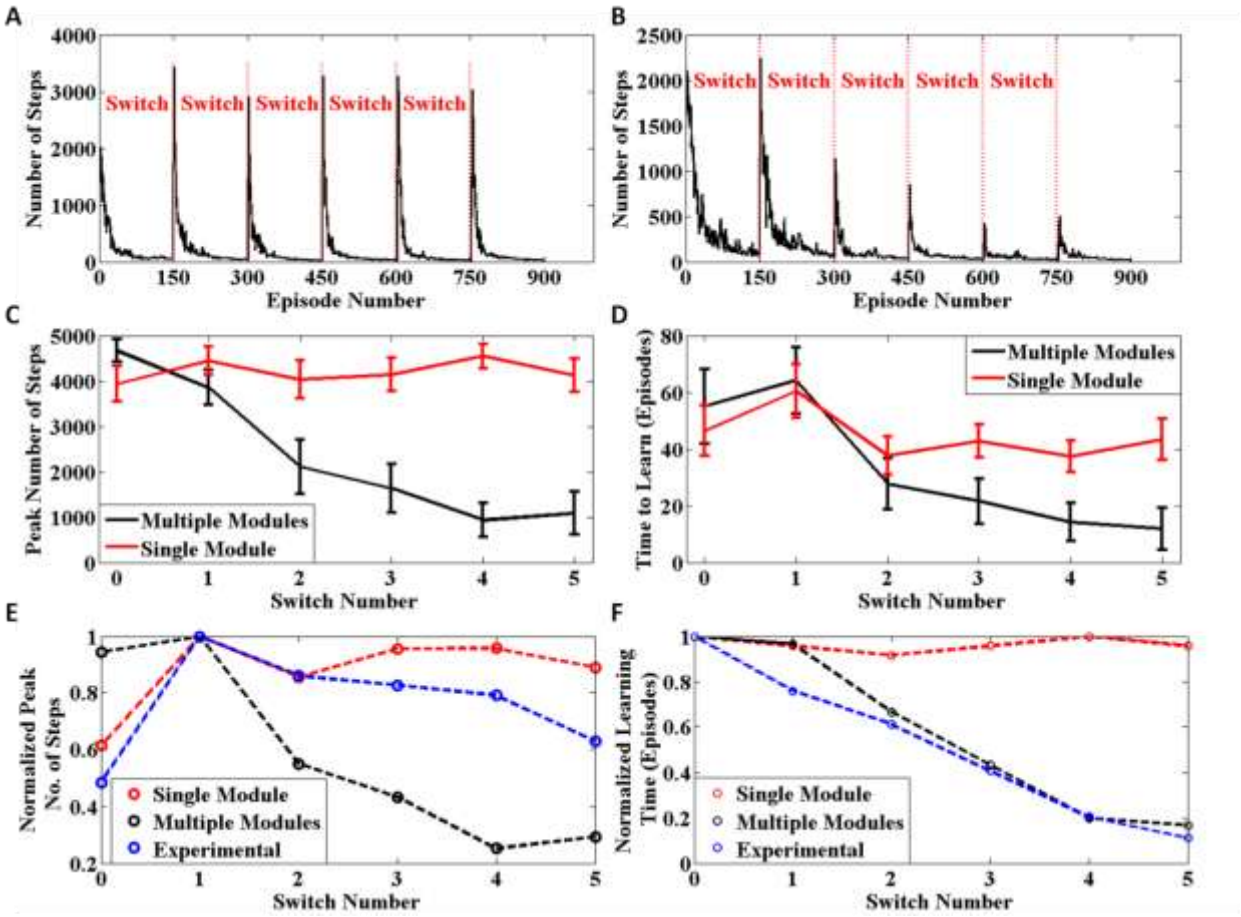

Fig. 7 **A)** Plot of Number of Steps taken by the single module agent in each episode averaged across 50 independent sessions. We see that the agent needs to relearn after each context switch **B)** Plot of Number of Steps taken by the multi module agent in each episode averaged across 50 independent sessions. We see that the agent efficiently switches modules after each context switch **C)** Peak number of steps needed to reach the goal after a context switch averaged across 50 sessions. **D)** Number of episodes for the number of steps required to reach the goal to go below a certain threshold averaged across 50 sessions **E)** Peak value for the average number of steps needed to reach the goal after a context switch. The experimental values have been adapted from

(Brunswik 1939) **D)** Number of episodes for the average number of steps required to reach the goal to go below a certain threshold. The experimental values have been adapted from (Brunswik 1939)

CHAPTER 4

DISCUSSION

We have proposed a network model of BG incorporating a computational framework to capture the microanatomy of the striatum. Our model shares features with existing models of BG designed to solve reinforcement learning (RL) tasks. In addition to solving RL tasks, our model exploits the modularity of the striatum to solve tasks with varying reward distributions in multiple contexts.

Striosome-Matrisome Dynamics with their Dopaminergic Projections

Our model is based on the assumption that striosomes map state information and matrisomes map action information. Earlier results suggest that the striosomes receive input from the orbitofrontal cortex (Eblen and Graybiel 1995) known for coding reward related states (Wilson, Takahashi et al. 2014). Matrisomes receive connections from primary motor and somatosensory cortices and could have action representations (Flaherty and Graybiel 1994), thereby supporting the assumptions of our model. Anatomical studies show that striosome medium spiny neurons (MSNs) project directly to SNc (Lanciego, Luquin et al. 2012). We believe that these projections could code for the state value of the agent as seen from the Strio-SOM to SNc connections in our model.

We propose that the striosome neurons influence the behaviour of the surrounding matrisome neurons. Earlier results show that Fast Spiking Interneurons (FSIs) and Persistent and Low-Threshold Spike (PLTS) interneurons are anatomically suitable candidates for this role since they branch across the patch and matrix (Gittis and Kreitzer 2012). We believe that the dopaminergic projections to these interneurons (Bracci, Centonze et al. 2002) could allow the striosome to bias the surrounding matrisome activity towards a preferred action. To our knowledge, this modulation (Eq. 8) is a unique

feature to our model and gives a biologically plausible mechanism to perform Q-learning. This is also supported by experiments which indicate that the striatum contributes to action selection by biasing its output towards the most desirable action (Samejima, Ueda et al. 2005; Hikosaka, Nakamura et al. 2006).

Mapping Representations to Action Primitives

Striatal MSN recordings show that they encode action representations and are modulated by the expected reward for the actions (Isomura, Takekawa et al. 2013). Our model agrees with this as both the Matri-SOM D1 and D2 neurons represent the action space and are correspondingly modulated by the TD error which is representative of the expected reward. Experiments also show activity in the MSNs corresponding to the outcome of the chosen action (Kim, Sul et al. 2009). We believe again that this could be the signal required to bias the activity of the striatal MSN as seen in the model (Eq. 8).

GPi forms the output nucleus of the BG and receives projections from Striatal MSNs through the direct and indirect pathways. Lesion studies show that GPi control movement by inhibitory projections to the thalamus and lesioning GPi impairs motor responses (Baunez and Gubellini 2010). Experiments also show that in the executive part of the task, the GPi activity is strongly related to the action performed (Pasquereau, Nadjar et al. 2007).

We propose that the connections from striatal D1 MSNs and STN to the GPi map the projections from action representations to action primitives. We believe that this mapping provides a flexible method to switch different action primitives for the same representations and vice versa, providing a plausible mechanism of adaptation in learning. Experiments show evidence of transformation of action information seen as higher degree of correlation in GPi activity as it passes from striatum to the GPi (Garenne, Pasquereau et al. 2011).

Contextual Learning and Striatal Modularity

Contextual Learning refers to the ability of the agent to adapt and learn in different

contexts. Some earlier operant conditioning experiments in such tasks have an explicit indication of contexts (different room or colour for each context) using which the agent can choose its actions (Bouton & King, 1983; Bouton & Peck, 1989). In such tasks, the agent shows renewal upon context switching indicating a mechanism for context identification. Experimental results indicate that the BG encodes the context as well as the choices in those contexts (Garenne, Pasquereau et al. 2011).

A recent study (Amemori, Gibb et al. 2011) hypothesized that the modular architecture of the striatum makes it a suitable candidate for solving multi context RL problems. We build on this by providing a computational neural model for the same. We describe the plausible correlates for computing the necessary variables needed to solve multi-context problems using a modular setting. The context prediction signal is very similar to a state value and we propose that neurons in the SNc code for this signal as well (Tobler, Fiorillo et al. 2005). In our model this is represented by the projections from Strio-SOM to the SNc. There is also a need for a reward prediction variance signal or a risk signal. Dopamine in the midbrain is proposed to also represent the risk component in the environment (Schultz 2010). In addition, it has been proposed that serotonin activity in the striatum correlates to risk or reward variance, just as dopamine codes for reward prediction error (Balasubramani, Chakravarthy et al. 2015).

We propose that the module selection and switching in different contexts could be carried out by Tonically Active Neurons (TANs). TANs exert a strong influence on striatal information processing and lesioning inputs to TANs impair learning after a change in reward distribution (Ragozzino, Jih et al. 2002). In our model, the TANs compete with each other and select the module appropriate for the task. Experiments support this hypothesis by showing that TANs can compete with each other using inhibitory connections similar to the model (Sullivan, Chen et al. 2008) and can cause widespread inhibition of MSNs by activating a GABAergic subpopulation (English, Ibanez-Sandoval et al. 2012). Another plausible method for context switching by TANs is by producing Acetylcholine (ACh) which can inhibit targeted MSNs. Dynamic changes in Acetylcholine output in the medial striatum (Ragozzino and Choi 2004) during reversal

learning supports this claim.

Behavioral Observations

Several behavioral processes were also observed from the results of the experiments on the model. We saw in Fig. 6B that the agent increases its Down and Right actions when the goal is placed at the bottom right corner. The agent thus exhibits acquisition (Graham and Gagné 1940) since it strengthens certain actions over the others based on the reinforcement given. We saw in Fig. 6D, that once the context has changed, the agent stops choosing the initial preferred response. This demonstrated extinction (Graham and Gagné 1940) since the behavior associated with a certain task gets elimininated when the reinforcement associated is removed. The experiments also indicate that the agent is able to show stimulus generalization and stimulus discrimination as the agent is able to distinguish between two different contexts which are two distinct stimuli (Till and Priluck 2000). Also the value function peaks where the goal is given, therefore goals which are near each other will have similar value profiles. From Fig. 7B, we saw that after two changes when the initial context reappears, the agent is able to bring back the policy learnt almost immediately exhibiting spontaneous recovery (Graham and Gagné 1940) referring to the reappearance and faster relearning of a previously extinguished response.

SECTION II

CHAPTER 5

STOCHASTIC MULTI CONTEXT TASK

A stochastic multi context task is an extension of the standard task used in a reinforcement learning setting. In this section, we introduce the various task settings and parameters and introduce the notation used in the rest of the chapter. In a standard task, the agent is present in a state ***s*** and can take action ***a***. Upon taking an action ***a***, the agent goes to a state $\boldsymbol{s}^*$ and is given a reward *r*. The reward *r* is obtained from the reward distribution function $R : S \times A \mapsto \mathcal{R}$ as $r = R(s, a)$ where S and A are state and action spaces of dimensions dim(***s***) and dim(***a***) respectively and $\mathcal{R}$ is the reward space which is a subset of real numbers ($\mathbb{R}$). (dim(**x**) denotes the dimension of the vector **x**)

This problem becomes harder when the environment is not stationary and the reward distribution changes based on which context the environment is present in. Mathematically, this means that the reward distribution function is redefined as $R : S \times A \times C \mapsto \mathcal{R}$ and $r = R(s, a, c)$ where C is the context space of dimension dim(***c***) and ***c*** is the context in which the agent is present. The problem is harder in this case since the agent has to identify the context in which it is present and then choose the action accordingly. This set of tasks are known as multi context tasks.

We can make the task much harder by introducing stochasticity in the problem. This is done by defining R as a probability distribution over $\mathcal{R}$ and *r* is a sample drawn from this distribution. While individually having multiple contexts or stochasticity is reasonably solvable, together they make the problem highly non-trivial. This set of problems are the stochastic multi context problems. Such problems can be viewed as an extension of contextual bandits (Langford and Zhang 2008)) where the context information is not presented to the agent.

CHAPTER 6

METHODS

Bayesian Model Formulation

We defined the stochastic multi context problem in the previous section. In this section, we look at an algorithm to solve the problem. We consider a simpler version of the problem but the discussions can be extended to harder tasks. We consider a single state so that the reward only depends on the context and the action chosen. We look at a setting where there are two possible actions, a_1 and a_2 and two contexts c_1 and c_2. Let a_1 be the optimal action in c_1 and a_2 in c_2. Also we restrict $\mathcal{R}$ to have 2 values- $R_{success}$ and $R_{failure}$. Since there are two possible actions and contexts, we define a reward distribution matrix as follows

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$$

where r_{ij} is the probability of getting a reward $R_{success}$ while taking action a_j in context c_i. We get $R_{failure}$ with a probability (1- r_{ij}) while taking action a_j in context c_i. With the help of this, we define the reward distribution function as

$$R(c_i, a_j) = \begin{cases} R_{success} & \text{with probability } r_{ij} \\ R_{failure} & \text{with probability } 1 - r_{ij} \end{cases}$$

Having formulated the problem, we notice that solving the problem essentially reduces to estimating the current context since we know the optimal action in each context. Assuming we choose action ***a*** and get a reward r, using Bayes Theorem

$$P(c=c_1 \mid a,r) = \frac{P(a,r \mid c=c_1)P(c=c_1)}{P(a,r \mid c=c_1)P(c=c_1)+P(a,r \mid c=c_2)P(c=c_2)} \quad \text{Eq. 25}$$

$$P(c=c_2 \mid a,r) = \frac{P(a,r \mid c=c_2)P(c=c_2)}{P(a,r \mid c=c_1)P(c=c_1)+P(a,r \mid c=c_2)P(c=c_2)} \quad \text{Eq. 26}$$

Assuming we do not have any knowledge of the current context, $P(c=c_1)=P(c=c_2)=0.5$. Also $P(c=c_2 \mid a,r)=1-P(c=c_1 \mid a,r)$. Hence we need to only track Eq. 25 which reduces to

$$P(c=c_1 \mid a,r) = \frac{P(a,r \mid c=c_1)}{P(a,r \mid c=c_1)+P(a,r \mid c=c_2)} \quad \text{Eq. 27}$$

We can now extend this to multiple trials by keeping track of the history of action selection and rewards obtained. At the i^{th} trial, let the action chosen be a^i and the reward obtained be r^i. We get at the n^{th} trial

$$P((c^n=c_1),...,(c^1=c_1) \mid (a^n,r^n),...(a^1,r^1)) =$$

$$\frac{P((a^n,r^n),...(a^1,r^1) \mid (c^n=c_1),...,(c^1=c_1))}{P((a^n,r^n),...(a^1,r^1) \mid (c^n=c_1),...,(c^1=c_1))+P((a^n,r^n),...(a^1,r^1) \mid (c^n=c_2),...,(c^1=c_2))} \quad \text{Eq. 28}$$

and correspondingly for context 2 as well. Due to independence of trials, the Eq. 28 can be simplified as

$$P((c^n = c_1),...,(c^1 = c_1) \mid (a^n, r^n),...(a^1, r^1)) =$$

$$\frac{\prod_{i=1}^{n} P(a^i, r^i \mid c^i = c_1)}{\prod_{i=1}^{n} P(a^i, r^i \mid c^i = c_1) + \prod_{i=1}^{n} P(a^i, r^i \mid c^i = c_2)} \quad \text{Eq. 29}$$

Instead of keeping the full history since beginning, we can consider the history for a particular window length m, making Eq. 29

$$P((c^n = c_1),...,(c^1 = c_1) \mid (a^n, r^n),...(a^1, r^1)) =$$

$$\frac{\prod_{i=n-m+1}^{n} P(a^i, r^i \mid c^i = c_1)}{\prod_{i=n-m+1}^{n} P(a^i, r^i \mid c^i = c_1) + \prod_{i=n-m+1}^{n} P(a^i, r^i \mid c^i = c_2)} \quad \text{Eq. 30}$$

These terms can be read from the reward distribution function. However, the reward distribution function is not accessible to the agent. Thus this model is not realistic and we need to estimate these terms which gives rise to the proposed theoretical model.

Theoretical Model

The Bayesian model developed in the previous section seems to solve the problem of estimating the context in which the agent is present. However it uses $P(a^i, r^i \mid c^i = c_1)$ which is not available to the agent. Thus, the next best option is to estimate the context the agent is in and then choose the actions accordingly. We denote the context estimated

by the agent using $\hat{c}$. Following the same steps as above we get the expression for the estimated context as

$$P((\hat{c}^n = \hat{c}_1),...,(\hat{c}^1 = \hat{c}_1) \mid (a^n, r^n),...(a^1, r^1)) =$$

$$\frac{\prod_{i=n-m+1}^{n} P(a^i, r^i \mid \hat{c}^i = \hat{c}_1)}{\prod_{i=n-m+1}^{n} P(a^i, r^i \mid \hat{c}^i = \hat{c}_1) + \prod_{i=n-m+1}^{n} P(a^i, r^i \mid \hat{c}^i = \hat{c}_2)} \quad \text{Eq. 31}$$

Now we can get values for the terms in Eq. 7 since the agent knows which context it estimated it was in when taking the action. Using the information from the preceding trials, we can estimate the probability as

$$P(a^i, r^i \mid \hat{c}^i = \hat{c}_1) = \frac{N((a^i, r^i) \mid \hat{c} = \hat{c}_1)}{N(\hat{c} = \hat{c}_1)} \quad \text{Eq. 32}$$

where $N((a^i, r^i) \mid \hat{c} = \hat{c}_1)$ is the number of times the agent chose a^i when it estimated the context as $\hat{c}_1$ and got the reward r^i and $N(\hat{c} = \hat{c}_1)$ is the number of times the agent estimated it's context as $\hat{c}_1$. This expression was derived so that agent can estimate the context it is in by looking at the term $P((\hat{c}^n = \hat{c}_1),...,(\hat{c}^1 = \hat{c}_1) \mid (a^n, r^n),...(a^1, r^1))$.

But to calculate this, we require terms that expect the agent to estimate the context and choose actions according to that context. Thus, there is an inherent circularity in the problem. To break this, we solve the problem in an iterative manner. We try to estimate the reward distribution function at trial number *t* and denote this as $\hat{\mathbf{R}}^t$. In addition, we

keep track of another matrix $\mathbf{N}^t$ which has the number of times, the agent chose a particular action in a particular estimated context. The two matrices are as follows

$$\mathbf{R}^t = \begin{bmatrix} \hat{r}_{11}^t & \hat{r}_{12}^t \\ \hat{r}_{21}^t & \hat{r}_{22}^t \end{bmatrix}$$

where $\hat{r}_{ij}^t$ represents the estimated probability of getting a reward $R_{success}$ when choosing action a_j in estimated context $\hat{c}_i$ at trial *t*.

$$\hat{\mathbf{N}}^t = \begin{bmatrix} \hat{n}_{11}^t & \hat{n}_{12}^t \\ \hat{n}_{21}^t & \hat{n}_{22}^t \end{bmatrix}$$

where $\hat{n}_{ij}^t$ represents the number of times the agent chose action a_j in estimated context $\hat{c}_i$ at trial *t*. For ease of notation, we also define $L_k^i = P(a^i, r^i \mid \hat{c}^i = \hat{c}_k)$ and k varies from 1 to 2. With this Eq. 31 becomes

$$P((\hat{c}^n = \hat{c}_1), ..., (\hat{c}^1 = \hat{c}_1) \mid (a^n, r^n), ...(a^1, r^1)) =$$

$$\frac{\prod_{i=n-m+1}^{n} L_1^i}{\prod_{i=n-m+1}^{n} L_1^i + \prod_{i=n-m+1}^{n} L_2^i} \qquad \textbf{Eq. 33}$$

Since the reward probabilities are equally likely at the beginning of the trial, we have

$$\mathbf{R}^0 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\mathbf{N}^0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$P(\hat{c}^0 = \hat{c}_1) = P(\hat{c}^0 = \hat{c}_2) = 0.5$$

In trial *t*, the agent estimates its current context ($\hat{c}_i$) based on its estimate in the previous trial and chooses the action (a_j) as given in Eq. 34 and Eq. 35 respectively.

$$i = \arg\max_{k\in\{1,2\}} P(\hat{c}^{t-1} = \hat{c}_k) \quad \text{Eq. 34}$$

$$j = \begin{cases} \arg\max_{k\in\{1,2\}} \hat{r}_{ik} & \text{with probability 1-}\epsilon \\ 1+b & \text{with probability } \epsilon \end{cases} \quad \text{Eq. 35}$$

where ϵ denoted the probability of exploration and $b \sim Ber(0.5)$. The exploration ensures that all the actions are sampled in the initial trials.

Based on the choice of $\hat{c}_i$ and a_j, the agent can update the values of $\hat{\mathbf{R}}^t$ and $\mathbf{N}^t$ as given in Eq. 36 and Eq. 37 respectively.

$$\hat{n}_{ij}^t = \hat{n}_{ij}^{t-1} + 1 \quad \text{Eq. 36}$$

$$\hat{r}_{ij}^t = \frac{(\hat{n}_{ij}^{t-1} * \hat{r}_{ij}^{t-1} + r^t)}{\hat{n}_{ij}^t} \quad \text{Eq. 37}$$

where r^t denotes the reward obtained at trial t.

Since $\hat{r}_{ij}^t$ represents the estimated probability of getting a reward $R_{success}$ when choosing action a_j in estimated context $\hat{c}_i$ at trial *t*, 1-$\hat{r}_{ij}^t$ represents the estimated probability of getting a reward $R_{failure}$. Thus L_i^t is given in

$$L_i^t = \begin{cases} \hat{r}_{ij}^t & r^t = R_{success} \\ 1-\hat{r}_{ij}^t & r^t = R_{failure} \end{cases} \quad \text{Eq. 38}$$

Substituting values of Eq. 38 in Eq. 33, we can get the estimates of the context in trial *t* as given in

Eq. 39

$$P(\hat{c}^t = \hat{c}_1) = \frac{\prod_{f=t-m+1}^{t} L_1^f}{\prod_{f=t-m+1}^{t} L_1^f + \prod_{f=t-m+1}^{t} L_2^f}$$

Eq. 34 to Eq. 39 can be used to formulate an algorithm for the agent to solve a stochastic multi context task as shown in Fig. 8

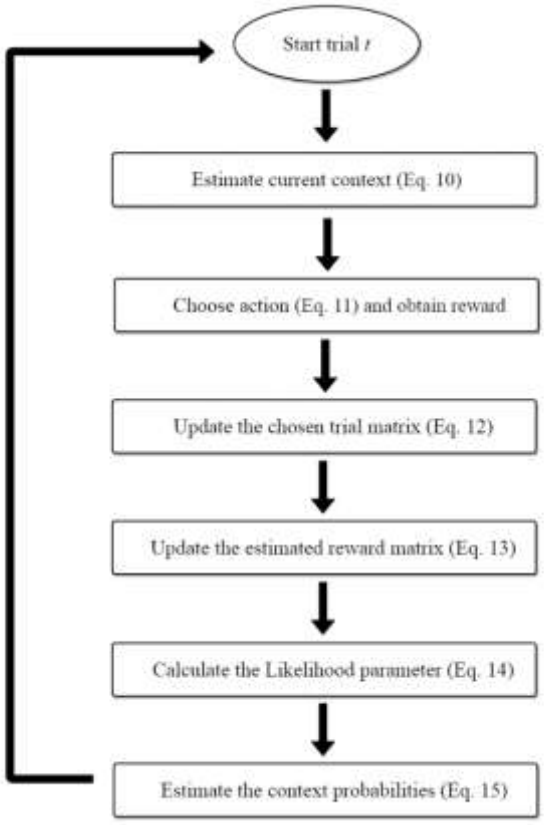

Fig. 8 Flowchart depicting steps to solve a stochastic multi context task.

Stochastic Reward Based Task Learning in Striatum

We proposed a theoretical model in the last section to solve stochastic multi context tasks. In this section we develop a biologically plausible model of the striatum for these tasks. We model the centre-surround structures seen in the striatum (Fig. 9A) using a

layered SOM model. In a layered SOM model, each neuron in a top SOM layer projects to a secondary SOM layer.

The top layer in our striatal model is the Strio-SOM, which maps the state space and is believed to model the striosomes. The neurons in the Strio-SOM project to the Matri-SOM which maps the action space and is believed to model the matrisomes Fig. 9B.

If we have m_1 x n_1 neurons in the Strio-SOM and m_2 x n_2 neurons in the Matri-SOM, the weights of the Strio-SOM(W^S) have dimension m_1 x n_1 x dim(**s**) where ***s*** is the state vector. Similarly, for an action vector ***a*** the weights of all the Matri-SOMs (W^M) are of dimension m_1 x n_1 x m_2 x n_2 x dim(***a***) as each neuron in the Strio-SOM projects to a Matri-SOM.

For a state input ***s***, the activity for a neuron *n* in the Strio-SOM is given in Eq. 40.

$$X^S_{[n]} = \exp\left(\frac{-\| W^S_{[n]} - s \|_2^2}{{\sigma_S}^2}\right) \qquad \text{Eq. 40}$$

where [n] represents the spatial location of the neuron n and σ_S controls the spread of the neuron activity. The complete activity of the Strio-SOM (X^S) is the combination of individual activity of all the neurons. The neuron with the highest activity ("winner") for a state **s** is denoted by n_s^*.

Similarly, for an action input ***a*** corresponding to a state ***s***, the activity for a neuron *n* in the Matri-SOM is given in Eq. 2.

$$X^M_{[n_s^*][n]} = \exp\left(\frac{-\| W^M_{[n_s^*][n]} - a \|_2^2}{{\sigma_M}^2}\right) \qquad \text{Eq. 41}$$

where σ_M controls the spread of the neuron activity. The complete activity of the Matri-SOM corresponding to neuron n_s^* ($X^M_{[n_s^*]}$) is the combination of individual activity of

all the neurons in the Matri-SOM corresponding to n_s^*. The neuron with the highest activity ("winner") for an action **a** in a state **s** is denoted as $n_{s,a}^*$.

The weight of a neuron *n* in the Strio-SOM for a state input **s** is updated according to the following rule (Eq. 3)

$$W_{[n]}^S \leftarrow W_{[n]}^S + \eta_S . \exp\left(\frac{-\|[n]-[n_s^*]\|_2^2}{\sigma_S^2}\right).(s - W_{[n]}^S) \quad \text{Eq. 42}$$

The weight of neuron *n* in the Matri-SOM for an action input **a** in a state **s** is updated according to Eq. 4.

$$W_{[n_s^*][n]}^M \leftarrow W_{[n_s^*][n]}^M + \eta_M . \exp\left(\frac{-\|[n]-[n_{s,a}^*]\|_2^2}{\sigma_M^2}\right).(a - W_{[n_s^*][n]}^M) \quad \text{Eq. 43}$$

These representations can be used to evaluate the states and actions and guide the decision making process. The schematic of our striatal model to solve stochastic RL tasks is given in Fig. 9C.

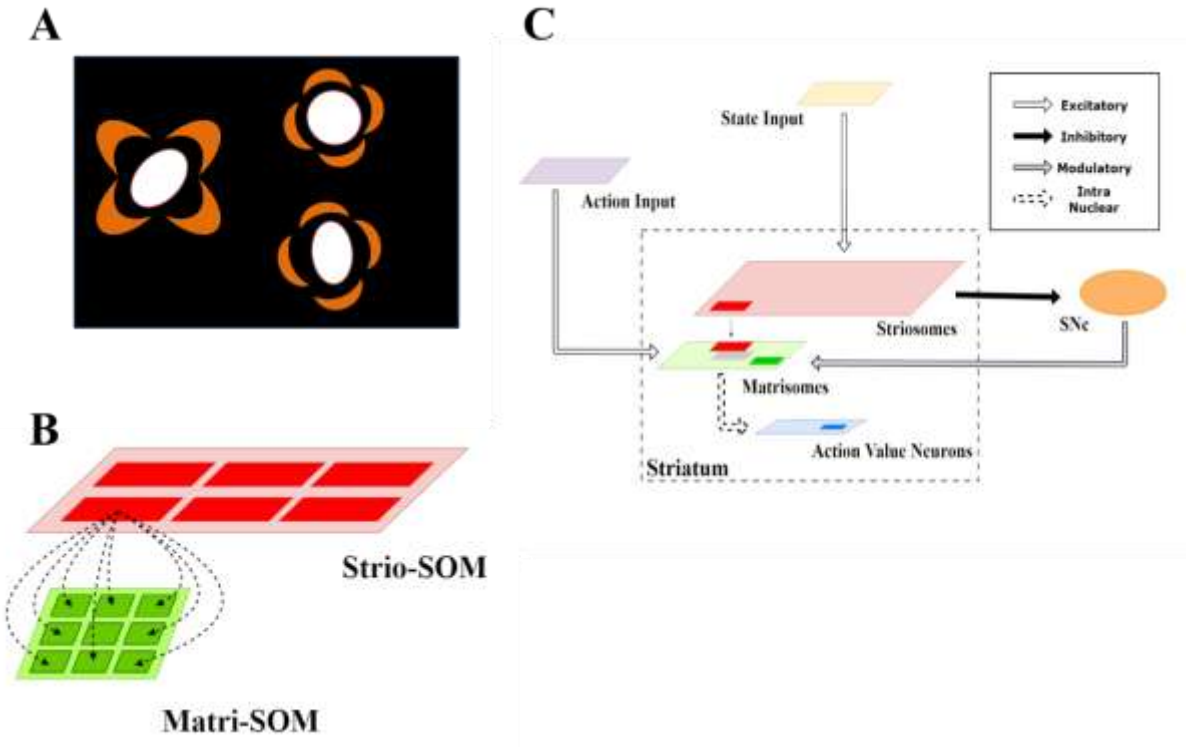

Fig. 9 **A)** Schematic of the centre surround mapping of seen in the striatum. The red centre represents the striosomes and the surround green neurons represent the matrisomes. **B)** Schematic of the layered SOM architecture where each neuron in the Strio-SOM (Red) projects to the neurons in the Matri-SOM (Green) **C)** Schematic diagram of the Striatum model where the arrows indicate the connections and their types.

Let the agent performing the task be in state ***s***. The striosome activity gives us the representation of the state in the striatum. This activity is modeled by the Strio-SOM as given in Eq. 40. Thus the activity is of dimension m_1 x n_1.

This activity of the Strio-SOM projects to the SNc and represents the value for the state **s** in our model (Eq. 5). The Striatal-SNc ($W^{Str \rightarrow SNc}$) are trained using the signal from SNc which is representative of Temporal Difference (TD) error (δ) (Eq. 6). The TD error is calculated as $\delta = r + \gamma V(s') - V(s)$ where ***s'*** is the new state after taking action **a** (), r is the reward obtained and γ is the discount factor.

$$V(s) = \sum_{\forall n} W^{Str \to SNc}{}_{[n]} X^{S}{}_{[n]}$$ **Eq. 44**

$$\Delta W^{Str \to SNc}{}_{[n]} = \eta^{Str \to SNc} \delta X^{S}{}_{[n]}$$ **Eq. 45**

where V(s) represents the value for state **s**, $\eta^{Str \to SNc}$ is the learning rate for $W^{Str \to SNc}$.

The actions that can be performed in a state ***s*** are represented by the matrisome activity surrounding the striosome neuron for that state. This is given by the activity of the Matri-SOM corresponding to the neuron with the highest activity in the Strio-SOM (n_s^*) in our model. The activity of a Matri-SOM neuron for an action ***a*** is given in Eq. 2 and is of dimension m_2 x n_2.

The Matri-SOM activity x for action ***a*** is projected to the action value neurons as given in Eq. 46. If n_a is the action value neuron for the action ***a***, $X^{Q}{}_{[n_a]}$ corresponds to the action value for the action in the state ***s*** in our model. These connections are also trained using TD error as above and the update equation is given in Eq. 47

$$X^{Q}{}_{[n_a]} = \sum_{\forall n} W^{Str(X_m) \to Str(Q)}{}_{[n_s^*][n]} X^{M}{}_{[n_s^*][n]}$$ **Eq. 46**

$$\Delta W^{Str(X_m) \to Str(Q)}{}_{[n_s^*][n]} = \eta^{Str(X_m) \to Str(Q)} \delta X^{M}{}_{[n_s^*][n]}$$ **Eq. 47**

where X^Q represents the activity of the action value neurons, $\eta^{Str(X_m) \to Str(Q)}$ is the learning rate for $W^{Str(X_m) \to Str(Q)}$.

The activity of the action value neurons are used for action selection by using a softmax policy (Sutton and Barto) in our model (Eq. 48). We believe that this is carried out by the dynamics of the STN-GPe oscillations with the striatal action value neurons projecting to the GPe.

$$P(a \mid s) = \frac{\exp(\beta X^{Q}_{[n_a]})}{\sum_{a' \in \mathcal{A}} \exp(\beta . X^{Q}_{[n_{a'}]})}$$ **Eq. 48**

where β is the inverse temperature and $\mathcal{A}$ denotes the action set for the agent.

Exploiting the Striatal Modularity for solving context dependent tasks

We propose that the modular nature of the striatal anatomy could be responsible for solving context dependent tasks using a modular RL framework. In this method, the agent allocates separate modules to separate contexts. Each of the modules has its own copy of the environment in a particular context, represented by an environment feature signal (ρ). This copy is used to generate a responsibility signal, denoted by λ, which indicates how close the current context is to the one represented by the module. Thus by identifying the module with the highest responsibility signal we can follow the policy developed in that module to solve the problem in an efficient manner. We can extend the model above to incorporate the modular RL framework. The schematic for the extended model is given in Fig. 10.

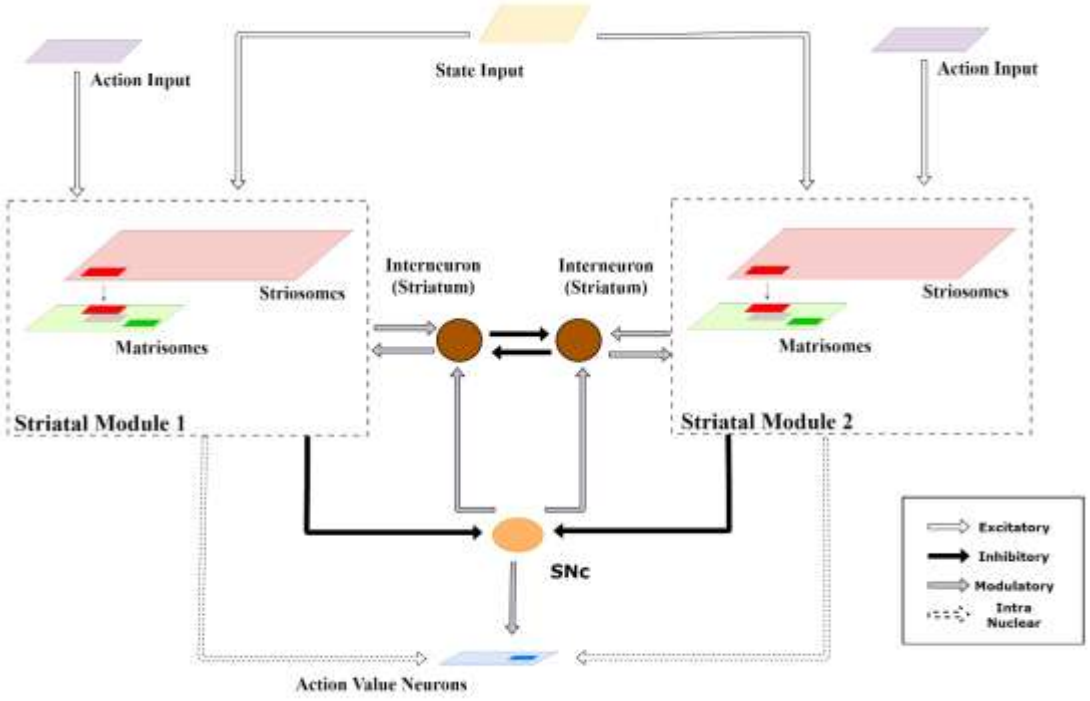

Fig. 10 Schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

We believe that context selection happens at the level of the striatum and the context modulated activity is projected to the action value neurons. Thus, for clarity, we have expanded the intra-nuclear activity of the striatum in the model schematic (**Fig. 3**). Supposing there are K modules denoted by M_1, M_2 … ,M_K. We now define the weights and activities in the previous sections for each module and denote $\{M_i\}$ with each term associated with module M_i. Thus, for a module *m*, the following variables undergo a change in notation: $X^S \rightarrow X^{S,\{m\}}$ (Eq. 40), $X^M \rightarrow X^{M,\{m\}}$ (Eq. 2), $W^S \rightarrow W^{S,\{m\}}$ (Eq. 3), $W^M \rightarrow W^{M,\{m\}}$ (Eq. 4), $V(s) \rightarrow V^{\{m\}}(s)$ (Eq. 5), $W^{Str\rightarrow SNc} \rightarrow W^{Str\rightarrow SNc,\{m\}}$ (Eq. 6), $W^{Str(X_m)\rightarrow Str(Q)} \rightarrow W^{Str(X_m)\rightarrow Str(Q),\{m\}}$ (Eq. 47).

We propose that in addition to the value of the state **s**, the activity of the Strio-SOM also projects to the SNc to represent the environment feature signal ($\rho^{\{m\}}$). The weights of these projections are denoted as $W_\rho^{\text{Str}\rightarrow\text{SNc},\{m\}}$ and are trained using the signal from SNc which is representative of context prediction error (δ^*). The corresponding equations are given in Eq. 20 and Eq. 21. The context prediction error is calculated as $\delta^* = r - \rho^{\{m\}}(s)$

$$\rho^{\{m\}}(s) = \sum_{\forall n} W_\rho^{Str\rightarrow SNc,\{m\}}{}_{[n]} X^{S,\{m\}}{}_{[n]} \quad \textbf{Eq. 49}$$

$$\Delta W_\rho^{Str\rightarrow SNc,\{m\}}{}_{[n]} = \eta_\rho^{Str\rightarrow SNc} \delta^* X^{S,\{m\}}{}_{[n]} \quad \textbf{Eq. 50}$$

We believe that the selection of the appropriate module for the context is guided by the striatal interneurons. In our model, the activity of the interneurons represents the responsibility signal for each module, denoted by $\lambda^{\{m\}}$ for module m. In a given state **s**, the inter-neurons compete among themselves and the one with the highest λ chooses the module responsible for deciding the action in that state. Let the winning module in the state **s** be denoted by m^*. The winning module projects to the action value neurons (Eq. 51) following which the processing is the same as in the previous section.

$$X^{Q}{}_{[n_a]} = \sum_{\forall n} W^{Str(X_m)\rightarrow Str(Q),\{m^*\}}{}_{[n_s^*][n]} X^{M,\{m^*\}}{}_{[n_s^*][n]} \quad \textbf{Eq. 51}$$

The dynamics of the responsibility signal is given in Eq. 24

$$\dot{\lambda} = -\lambda - \alpha_\lambda (\delta^*)^2 \quad \textbf{Eq. 52}$$

where α_λ controls the influence of context prediction error on the responsibility signal and δ^* is the context prediction error.

CHAPTER 7

RESULTS

Performance of theoretical model on T-Maze tasks

The study of context dependent stochastic tasks is a reasonably underexplored area owing to the complexity in these tasks. However, some of the earlier results (Lloyd and Leslie 2013) make some predictions which we aim to replicate with our model.

The task performed by the agent is a T-maze task (Olton 1979) where the agent has to choose one of the arms in a maze. Upon choosing the arm, the agent gets a reward R_{max} with a given probability ($P_{success}$) and a reward R_{min} with a given probability ($P_{failure}$). The task can be extended to a context dependent problem by reversing the reward distributions with trials.

We study the performance with changing R_{max}/R_{min} and $P_{success}/P_{failure}$. Animals tend to choose rewards which have a higher magnitude and greater rewards lead to a faster convergence (Fig. 11A). Similarly, with the same magnitude, animals tend to prefer distributions which reward with a higher probability (Fig. 11C). These effects are captured by our model as shown in Fig. 11B and Fig. 11D respectively. The figures show the ratio of the correct choices by the agent in 50 trials averaged over 50 sessions. The value of exploration factor, ϵ (Eq. 35) was set as 0.1 and the window length, m (Eq. 31) was chosen as 5.

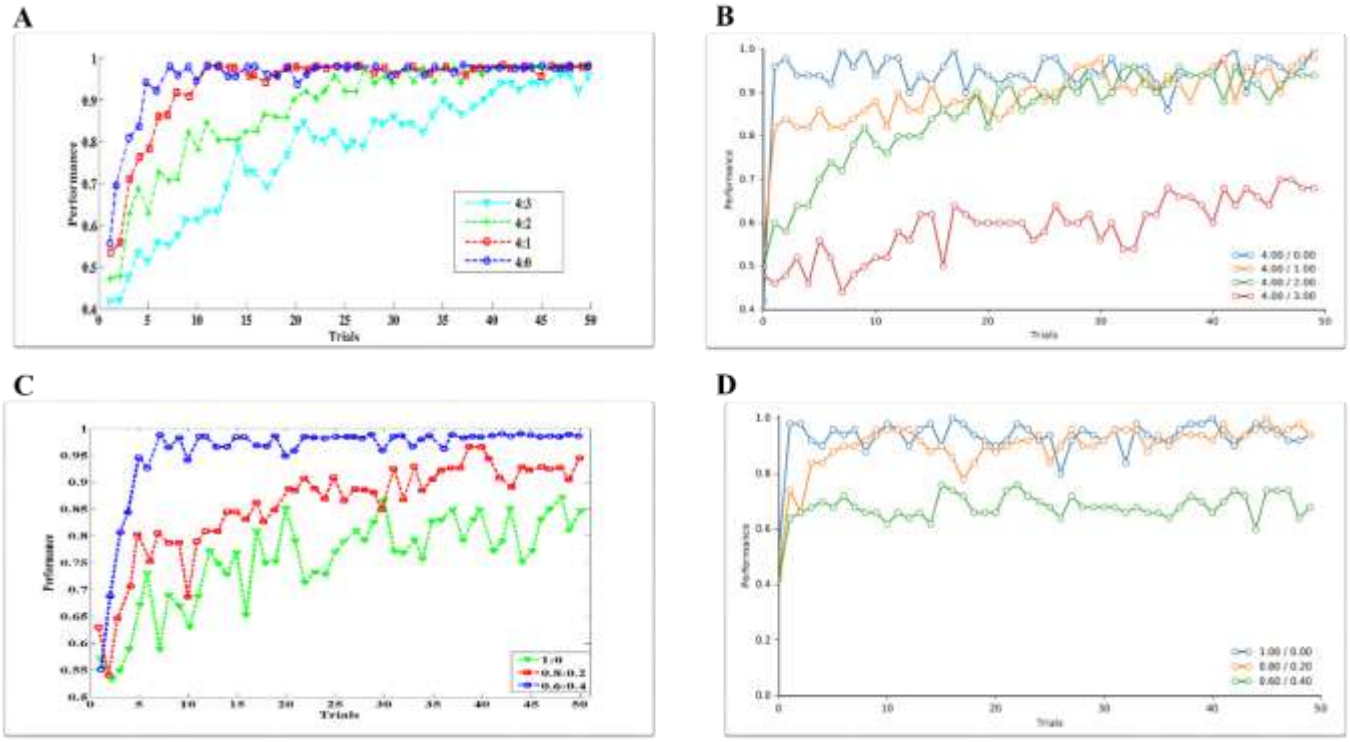

Fig. 11 **A)** Demonstration of change in performance with varying reward magnitudes (Figure adapted from (Lloyd and Leslie 2013)). **B)** Performance of our model on the varying reward magnitude task **C)** Demonstration of change in performance with varying reward probabilities (Figure adapted from (Lloyd and Leslie 2013)). **D)** Performance of our model on the varying reward probability task

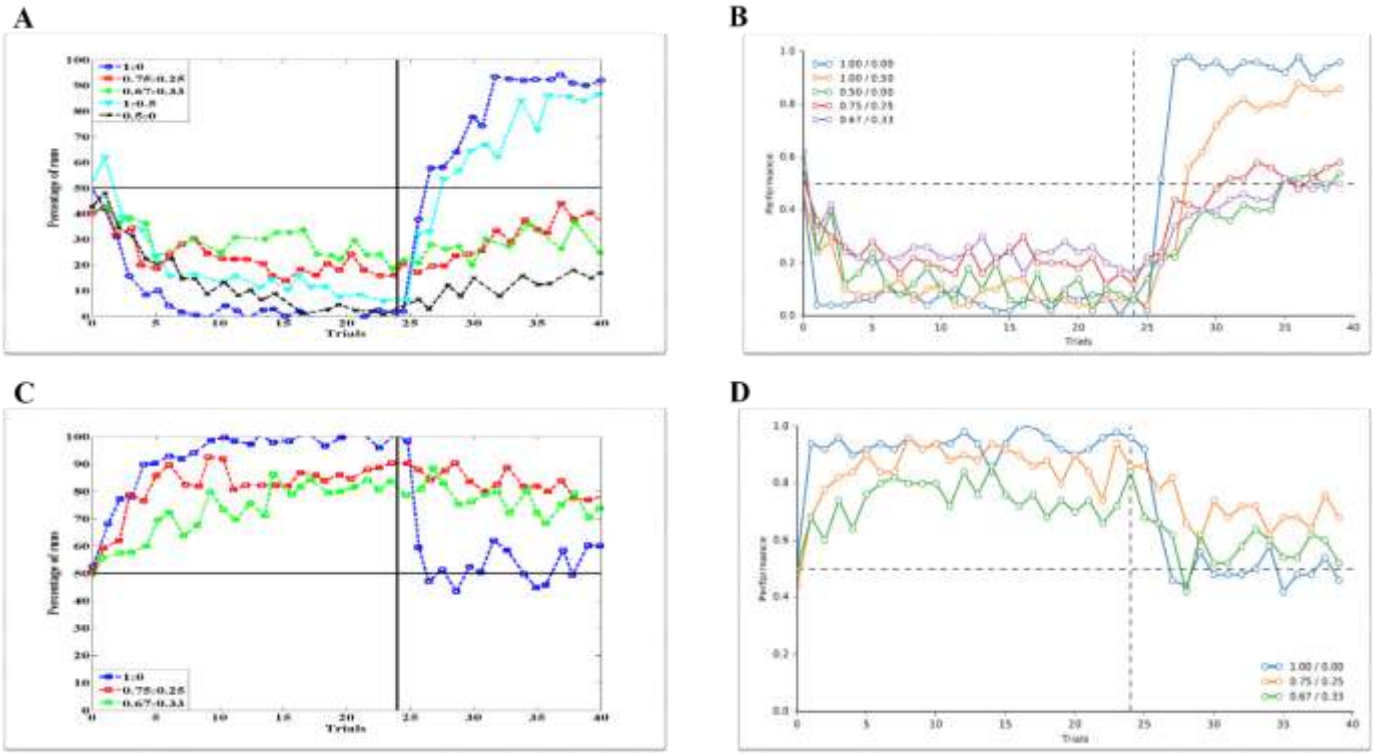

Fig. 12 **A)** Percentage of trials where the animal chooses the arm which is non-profitable for the first 24 trials and becomes profitable following that. (Figure adapted from (Lloyd and Leslie 2013)). **B)** Performance of the model on the task described in **A**. We see that the model shows similar trends where the definite reward tasks show faster reversal learning. **C)** Percentage of trials where the animal chooses the arm which was rewarding before 24 trials following which both arms are not rewarded (Figure adapted from (Lloyd and Leslie 2013)). **D)** Performance of the model on the task described in **C** where the model shows similar trends as the definite reward task show faster unlearning.

Experimental evidence (Brunswik 1939) shows that partial reinforcement and stochastic rewards have a significant effect on reversal learning. We consider a task where the animal is trained on a T-maze with different reward probabilities for 24 trials and then the rewarding probabilities are reversed. We look at the percentage of the trials where the animal chooses the arm which is unprofitable at first and becomes profitable after the reversal. We can observe that the model results (Fig. 12B) show similar trends to earlier

results (Fig. 12A). The tasks where one arm had a definite reward showed quick reversal. However probabilistic rewards show slower policy modulation by the agent.

Stochastic reward distributions also have an effect on extinction (Miltenberger 2011) of a learned policy. To test this, we consider a task where the animal on a T-maze for 24 trials as above. However, the rewards for both arms are set as 0 following the 24 trials and the rate of unlearning is studied. We observe that definite rewarding tasks show faster extinction as compared to the tasks with stochastic rewards (Fig. 12C) which is captured by the model (Fig. 12D).

Solving Stochastic Reward Based Tasks using the Striatum Model

In this section, we demonstrate that the striatum model developed is capable of solving stochastic tasks. We consider a cue based decision making task where the animal has to choose one of the cues displayed on the screen. This task was first described in (Pasquereau, Nadjar et al. 2007) and a schematic of the task is given in Fig. 13A. The animal is presented with two cues in each trial at two locations as seen in the figure. Each shape is associated with a different probability of reward. The agent has to choose one of the shapes and gets awarded a reward accordingly.

We show that our striatal model is able to solve this task. We consider a 4 dimensional state vector, where each dimension is 1 if the shape is shown and 0 otherwise. The action vector is also 4 dimensional with each dimension denoting the action that is chosen by the agent. The various parameters of the model are given in Table 2.

Table 2: Parameter values for cue based decision making task

Parameter	Value	Parameter	Value
Strio-SOM Dimension (m_1xn_1)	**3x2**	**Matri-SOM Dimension (m_2xn_2)**	**3x3**
σ_S	**0.01**	$\boldsymbol{\sigma_M}$	**0.1**
η_S	**0.4**	$\boldsymbol{\eta_M}$	**0.4**
γ	**0.95**	$\boldsymbol{\eta^{Str\rightarrow SNc}}$	**0.05**
$\boldsymbol{\eta^{Str(Xm)\rightarrow Str(Q)}}$	$\mathbf{5x10^{-4}}$	**B**	**50**
α_λ	**0.8**	$\boldsymbol{\eta_\rho^{Str\rightarrow SNc}}$	**0.1**

The agent (model) is pre-trained where it is given various state and action inputs. We show that the representational maps developed have a centre surround structure (Fig. 13C) when we view the activity corresponding to all the actions for a particular state. The ratio of correct choices chosen in 200 trials averaged over 25 sessions in given in Fig. 13B. Thus, we can see that the agent is able to solve stochastic reward based tasks. Experimental evidence shows that the percentage of times the agent chooses the arm with reward probability P1, when the ratio of the reward probabilities is P1+P2 follows a sigmoid activity with centre at 0.5 which is well captured by the model (Fig. 13D).

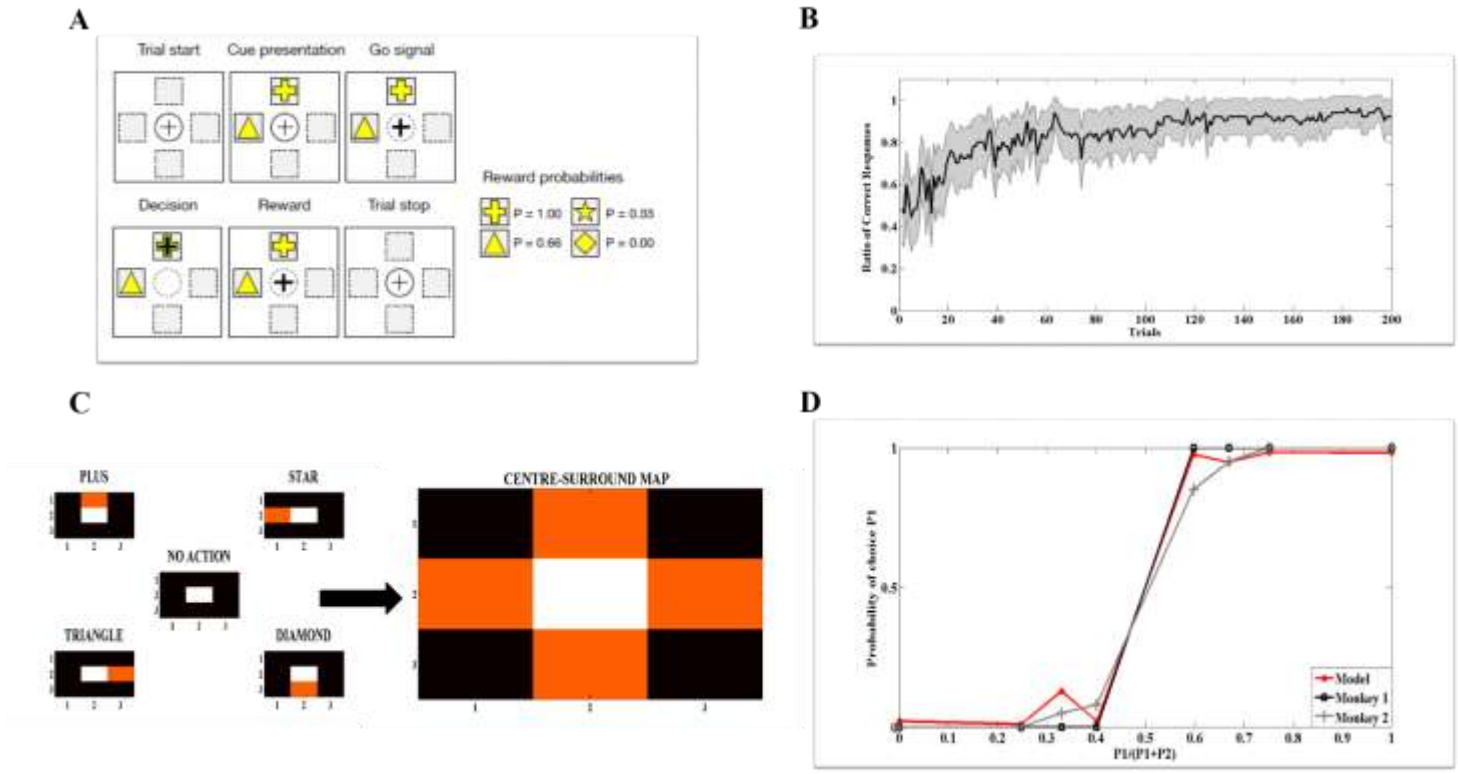

Fig. 13 **A)** Schematic of the cue based decision making task where the agent has to choose between the two shapes shown in the screen and each shape has a different probability of reward associated with it. **B)** Percentage of correct responses averaged over 25 sessions for 200 trials. **C)** Mapping of the action inputs forms a centre-surround structure when we view the combined activity of the Matri-SOM for all action inputs **D)** Ratio of choosing response 1 with associated probability P1 wrt to the sum P1+P2. The model follows a similar trend to the experimental plot adapted from (Pasquereau, Nadjar et al. 2007)

Comparing the Theoretical and Neural model

We have developed a theoretical model capable of solving stochastic multi-context tasks and also developed a neural model which provides a biologically plausible mechanism for the same. Since there are no concrete experiments dealing with these tasks, we use the theoretical model to understand the performances of the neural model. We use a

stochastic two arm bandit task which was the underlying problem in both the tasks described above. The reward distributions reverse after 500 trials and the performance of the agent is characterized by looking at 25 sessions. We look at the performances for different values of ε which represents the probability of reward for the non-profitable arm.

Fig. 14A demonstrates the probability of context 1 estimated by the theoretical model whereas Fig. 14B gives the estimate by the neural model. We observe that the theoretical model is able to identify the context even for larger values of the ε. The neural model is able to identify the context well in most cases but fails for larger values of ε. A similar trend can be seen in Fig. 15A and Fig. 15B where we measure the percentage of correct choices by the agent. We observe that the theoretical model is able to learn faster upon context reversal for all values of ε but the neural model needs to relearn for higher values of ε.

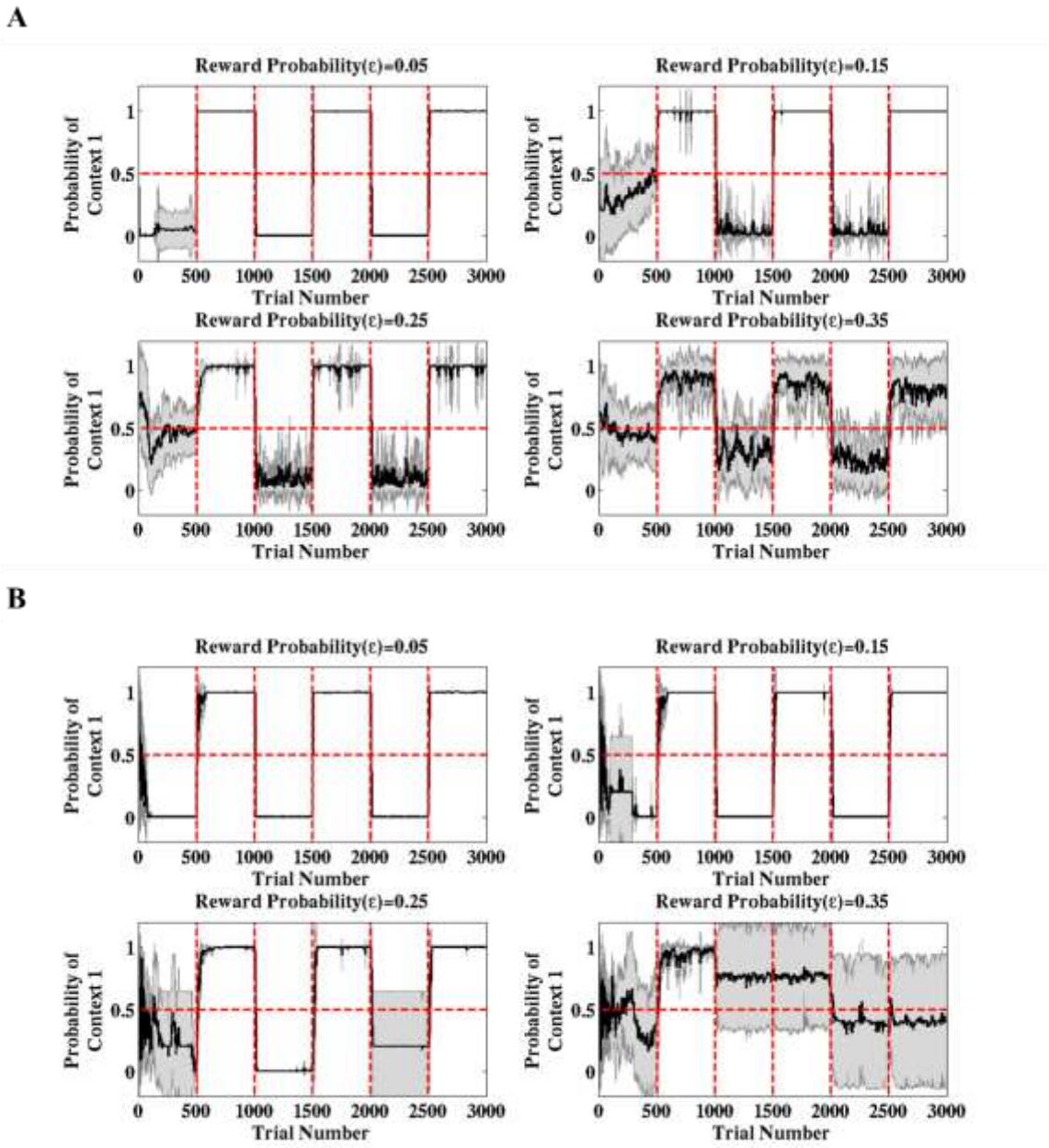

Fig. 14 **A)** Probability of context 1 estimated by the theoretical model. **B)** Probability of context 1 estimated by the neural model.

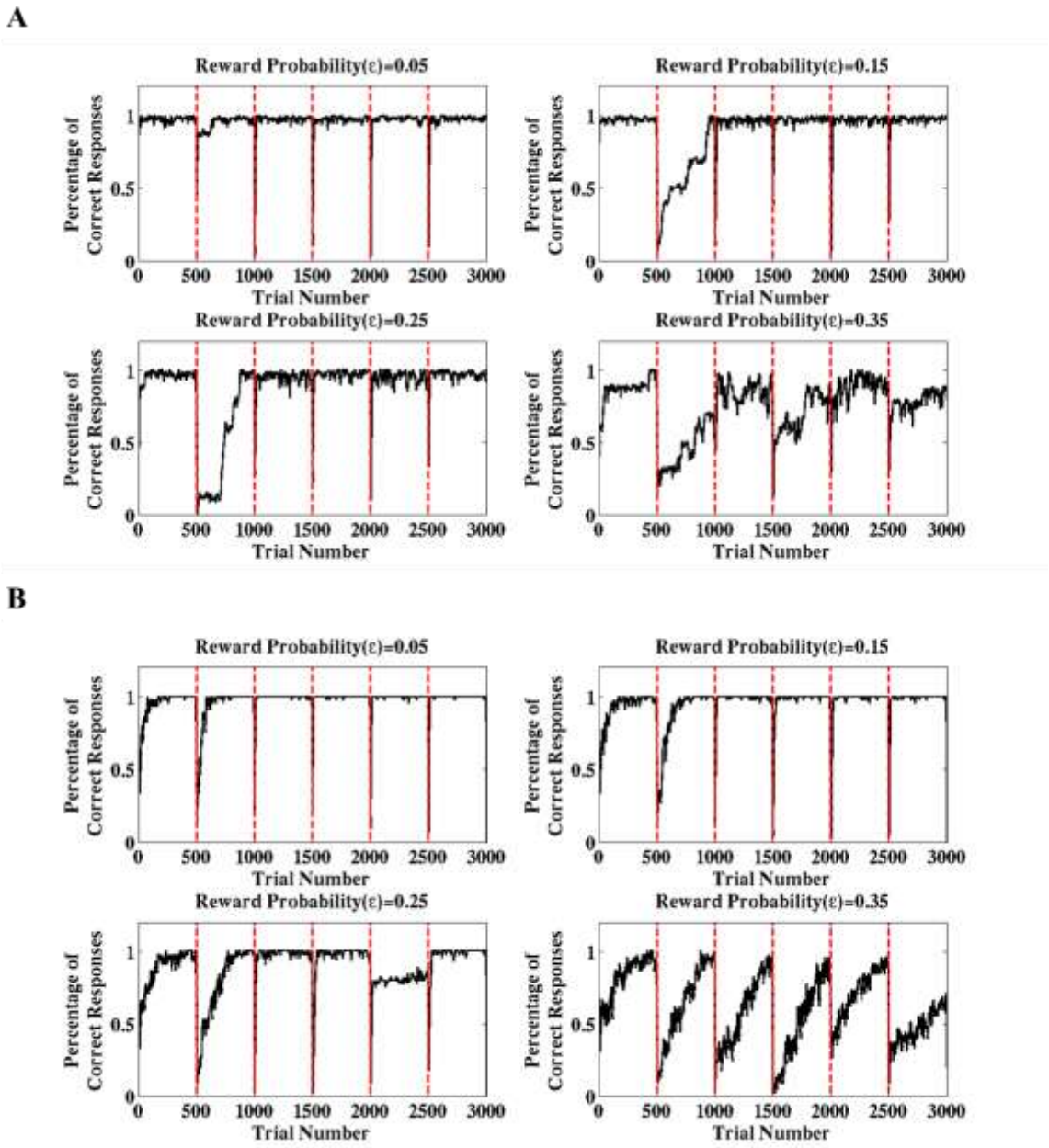

Fig. 15 **A)** Percentage of correct responses by the theoretical model. **B)** Percentage of correct responses by the neural model.

Thus we can conclude that the neural model is able to follow the theoretical model for low values of ε but behaves like a single context agent for larger values. This is shown in Fig. 16 which shows that the neural model lies between the theoretical optimal and a

single context model and could be the biological mechanism used for solving stochastic multi context tasks.

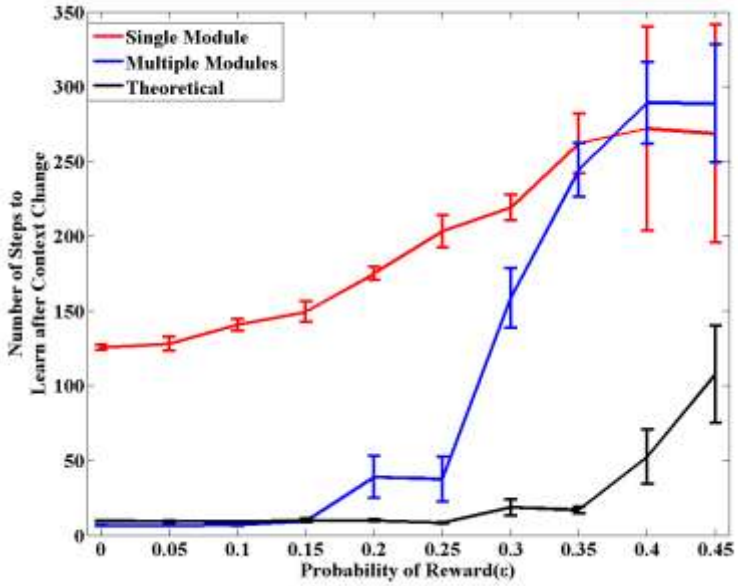

Fig. 16 Schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

CHAPTER 8

DISCUSSION

We have presented a theoretical model to solve stochastic multi-context tasks. This is also accompanied by a biologically plausible computational model of the striatum which also attempts to tackle these problems. The striatal model borrows some ideas from the earlier model described in the previous section and presents a alternate model of the striatum capable of handling stochastic RL problems.

Our model derives from the model in the previous section that the striosomes map the state space and the matrisomes map the action space. This is supported from earlier results that the striosomes receive input from the orbitofrontal cortex (Eblen and Graybiel 1995) known for coding reward related states (Wilson, Takahashi et al. 2014). Anatomical studies also show that striosome medium spiny neurons (MSNs) project directly to SNc (Lanciego, Luquin et al. 2012) which could compute values as in our model.

Evidence suggests that similar to how projections from the striosomes code for state value, projections from the matrisomes code for action value (Doya 2002). Experimental results show the existence of such neurons in the striatum which code specifically for action value (Samejima, Ueda et al. 2005). This is well captured in our model as the Matri-SOM projects to action value neurons in out striatal model.

Action selection is done using the softmax policy (Eq. 48) following the action value computation in the striatum. This policy uses a parameter β which controls the exploration of the agent. We believe that this could be the role of STN, GPe and GPi

before action selection is done at the level of the thalamus. This is supported by earlier results which suggest that The underlying stochasticity in the soft-max rule could be achieved indirectly by the chaotic dynamics of the STN-GPe loop (Kalva, Rengaswamy et al. 2012).

We have developed a theoretical model and a neural model for tackling stochastic multi-context tasks. Due to the inherent complex nature of the task, there is a lack of experimental data for tasks which are both stochastic and non-stationary. However, what we can observe from the results (Eq. 32) is that the neural model falls between the performance of the theoretical model and the neural model with a single module. Thus the theoretical model acts as a lower bound to the performance of the neural model for the given stochasticity in the problem. Also the neural model is able to achieve performance on par with the theoretical model for low values of ϵ but fails to do so for larger ϵ where it becomes similar to a single module system. Thus, we predict that our neural model can explain behavior in stochastic multi context tasks for $\epsilon<0.3$.

Another feature of our theoretical model is that it is a very simple model with no assumptions on the reward or the context distributions. However, despite its simplistic formulation, the model is quite powerful and is able to capture all the previous results quite well. The modular arrangement of identifying context and using it for task selection is very similar to the proposed striatal model. Thus, the striatal model could be a biologically plausible neural implementation of the theoretical model.

CONCLUSION AND FUTURE WORK

In this thesis, we have proposed two models of the striatum to handle stochastic multi context tasks. We first developed a layered-SOM architecture to model the centre-surround mapping seen in micro anatomical studies of the striatum. This was used to map the state and action spaces in the reward based task. In section I, we focused on handling non-stochastic context dependent tasks. We extended the model to a full network model of the Basal Ganglia. We proposed a biologically plausible mechanism of action based learning where the striosome biases the matrisome activity towards a preferred action. Using this model we were able to solve simple reinforcement learning tasks. We also exploited the modularity of the striatum to handle multi-context tasks. We tested the model on a grid world problem. We also demonstrated that our model captures experimental data better than existing single module models of the basal ganglia.

In section II, we looked at harder problems where the rewards were both stochastic and non stationary. We proposed a theoretical model which was capable of handling such tasks and showed that it captures trends seen in several earlier experiments. Following this, we proposed a striatal model which borrowed base features from the model in Section I but used action values for decision making. We tested this model on different stochastic problems and validated it with experimental data. We also compared the two models and showed that the neural model could match the theoretical model for low levels of ϵ. We also proposed that the neural model could be a neural implementation of the proposed theoretical model owing to their similar structure and performances.

The next step would be to integrate the two models are present a unified framework for solving all context dependent problem. Another improvement would be to obtain the state and action inputs by multi-sensory integration to present a complete cortico-basal ganglia model.

REFERENCES

Amemori, K.-i., L. G. Gibb, et al. (2011). "Shifting responsibly: the importance of striatal modularity to reinforcement learning in uncertain environments." Frontiers in human neuroscience **5**: 47.

Apicella, P. (2007). "Leading tonically active neurons of the striatum from reward detection to context recognition." Trends in neurosciences **30**(6): 299-306.

Balasubramani, P. P., V. S. Chakravarthy, et al. (2015). "A network model of basal ganglia for understanding the roles of dopamine and serotonin in reward-punishment-risk based decision making." Frontiers in computational neuroscience **9**: 76.

Bar-Gad, I., G. Havazelet-Heimer, et al. (2000). "Reinforcement-driven dimensionality reduction--a model for information processing in the basal ganglia." J Basic Clin Physiol Pharmacol **11**.

Bar-Gad, I., G. Morris, et al. (2003). "Information processing, dimensionality reduction and reinforcement learning in the basal ganglia." Progress in neurobiology **71**(6): 439-473.

Baunez, C. and P. Gubellini (2010). "Effects of GPi and STN inactivation on physiological, motor, cognitive and motivational processes in animal models of Parkinson's disease." Progress in brain research **183**: 235-258.

Bogacz, R. (2007). "Optimal decision-making theories: linking neurobiology with behaviour." Trends in cognitive sciences **11**(3): 118-125.

Bracci, E., D. Centonze, et al. (2002). "Dopamine excites fast-spiking interneurons in the striatum." Journal of neurophysiology **87**(4): 2190-2194.

Brunswik, E. (1939). "Probability as a determiner of rat behavior." Journal of Experimental Psychology **25**(2): 175.

Chakravarthy, V. S. and P. P. Balasubramani (2015). "Basal ganglia system as an engine for exploration." Encyclopedia of Computational Neuroscience: 315-327.

Charpier, S. and J. Deniau (1997). "In vivo activity-dependent plasticity at cortico-striatal connections: evidence for physiological long-term potentiation." Proceedings of the National Academy of Sciences **94**(13): 7036-7040.

Doya, K. (2002). "Metalearning and neuromodulation." Neural Networks **15**(4): 495-506.

Eblen, F. and A. M. Graybiel (1995). "Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey." Journal of neuroscience **15**(9): 5999-6013.

English, D. F., O. Ibanez-Sandoval, et al. (2012). "GABAergic circuits mediate the reinforcement-related signals of striatal cholinergic interneurons." Nature neuroscience **15**(1): 123-130.

Flaherty, A. and A. M. Graybiel (1994). "Input-output organization of the sensorimotor striatum in the squirrel monkey." The Journal of neuroscience **14**(2): 599-610.

Flaherty, A. and A. M. Graybiel (1994). "Input-output organization of the sensorimotor striatum in the squirrel monkey." Journal of Neuroscience **14**(2): 599-610.

Garenne, A., B. Pasquereau, et al. (2011). "Basal ganglia preferentially encode context dependent choice in a two-armed bandit task." Frontiers in systems neuroscience **5**.

Gittis, A. H. and A. C. Kreitzer (2012). "Striatal microcircuitry and movement disorders." Trends in neurosciences **35**(9): 557-564.

Graham, C. and R. M. Gagné (1940). "The acquisition, extinction, and spontaneous recovery of a conditioned operant response." Journal of Experimental Psychology **26**(3): 251.

Granger, R. (2006). "Engines of the brain: The computational instruction set of human cognition." AI Magazine **27**(2): 15.

Graybiel, A., A. Flaherty, et al. (1991). Striosomes and matrisomes. The basal ganglia III, Springer**:** 3-12.

Graybiel, A. M., T. Aosaki, et al. (1994). "The basal ganglia and adaptive motor control." SCIENCE-NEW YORK THEN WASHINGTON-: 1826-1826.

Hikosaka, O., K. Nakamura, et al. (2006). "Basal ganglia orient eyes to reward." Journal of neurophysiology **95**(2): 567-584.

Isomura, Y., T. Takekawa, et al. (2013). "Reward-modulated motor information in identified striatum neurons." Journal of neuroscience **33**(25): 10209-10220.

Kalmár, Z., C. Szepesvári, et al. (1999). "Modular reinforcement learning." Acta Cybernetica **14**(3): 507-522.

Kalva, S. K., M. Rengaswamy, et al. (2012). "On the neural substrates for exploratory dynamics in basal ganglia: a model." Neural Networks **32**: 65-73.

Kim, H., J. H. Sul, et al. (2009). "Role of striatum in updating values of chosen actions." Journal of neuroscience **29**(47): 14701-14712.

Kohonen, T. (1998). "The self-organizing map." Neurocomputing **21**(1): 1-6.

Lanciego, J. L., N. Luquin, et al. (2012). "Functional neuroanatomy of the basal ganglia." Cold Spring Harbor perspectives in medicine **2**(12): a009621.

Langford, J. and T. Zhang (2008). The epoch-greedy algorithm for multi-armed bandits with side information. Advances in neural information processing systems.

Lloyd, K. and D. S. Leslie (2013). "Context-dependent decision-making: a simple Bayesian model." Journal of The Royal Society Interface **10**(82): 20130069.

Miltenberger, R. G. (2011). Behavior modification: Principles and procedures, Cengage Learning.

Olton, D. S. (1979). "Mazes, maps, and memory." American psychologist **34**(7): 583.

Pasquereau, B., A. Nadjar, et al. (2007). "Shaping of motor responses by incentive values through the basal ganglia." Journal of Neuroscience **27**(5): 1176-1183.

Ragozzino, M. E. and D. Choi (2004). "Dynamic changes in acetylcholine output in the medial striatum during place reversal learning." Learning & Memory **11**(1): 70-77.

Ragozzino, M. E., J. Jih, et al. (2002). "Involvement of the dorsomedial striatum in behavioral flexibility: role of muscarinic cholinergic receptors." Brain research **953**(1): 205-214.

Samejima, K., Y. Ueda, et al. (2005). "Representation of action-specific reward values in the striatum." Science **310**(5752): 1337-1340.

Schultz, W. (2010). "Dopamine signals for reward value and risk: basic and recent data."

Behavioral and brain functions **6**(1): 24.

Sullivan, M. A., H. Chen, et al. (2008). "Recurrent inhibitory network among striatal cholinergic interneurons." Journal of neuroscience **28**(35): 8682-8690.

Sutton, R. S. and A. G. Barto Reinforcement learning: An introduction.

Sutton, R. S. and A. G. Barto (1998). Reinforcement learning: An introduction, MIT press Cambridge.

Till, B. D. and R. L. Priluck (2000). "Stimulus generalization in classical conditioning: An initial investigation and extension." Psychology & Marketing **17**(1): 55-72.

Tobler, P. N., C. D. Fiorillo, et al. (2005). "Adaptive coding of reward value by dopamine neurons." Science **307**(5715): 1642-1645.

Wilson, R. C., Y. K. Takahashi, et al. (2014). "Orbitofrontal cortex as a cognitive map of task space." Neuron **81**(2): 267-279.

LIST OF PUBLICATIONS BASED ON THESIS

Sabyasachi Shivkumar, Vignesh Muralidharan, V. Srinivasa Chakravarthy, 'A computational architecture to model the microanatomy of the striatum and its functional properties', in BMC Neuroscience 2016, 17(Suppl 1):P189

Sabyasachi Shivkumar, Vignesh Muralidharan, V. Srinivasa Chakravarthy , 'A biologically plausible architecture of the striatum to solve context-dependant reinforcement learning tasks', Frontiers in Neural Circuits **(Under Review)**