A SEQUENCE LEARNING ARCHITECTURE BASED ON CORTICO-BASAL GANGLIONIC LOOPS AND REINFORCEMENT LEARNING

Raju S. Bapi and Kenji Doya

Email:{rajubapi, doya}@erato.atr.co.jp

Computational Neurobiology Group, Kawato Dynamic Brain Project, ERATO, JST
2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan

ABSTRACT

A sequence processing neural network architecture is proposed based on the known cortico-basal ganglionic structure[1]. Learning in the network is implemented using biologically plausible reinforcement learning paradigm. The network is designed in order to understand the functional roles of various brain areas involved in the performance of sequential arm movements. Findings from Tanji & Shima's [1] experiments on monkeys are used as specific test case. We propose that the matrix and patch compartments of the basal ganglia implement an Actor and a Critic, respectively and learn a sequence based on the motor efference copy of previous movement. Further, in order to flexibly shift among various sequences, supplementary motor area structures learn the sequence simultaneously and act via a biasing mechanism to enable the basal ganglia-actor recall the correct sequence. Simulation results are presented to illustrate how these proposals can be implemented.

KEYWORDS: Sequences, Reinforcement Learning, Cortico-Basal Ganglionic Loops

1. INTRODUCTION AND BACKGROUND

Serial order is an important aspect of human and animal behavior and continues to receive attention. In spite of research over many years, it is not yet completely understood how brain solves the serial order of motor behavior problem [2]. However, in recent years there has been rapid progress in neuroscience, especially with the availability of newer and better technology.

In the present paper, we will review the experimental findings of Tanji & Shima [1] on monkeys. The aim of the current work is to give a computational account of these experimental findings. In particular, the aim is to understand how a network of neuron-like elements can learn a sequence of actions using very coarse reinforcement signals that specify if the action sequence is correct or wrong only at the end of sequence performance. The structure of the network is designed based on known neuroanatomical information [3]. Learning is implemented using a biologically plausible scheme known as the reinforcement learning (RL).

2. TANJI'S EXPERIMENTS

Tanji and colleagues devised a sequence learning paradigm and made single cell recordings in behaving monkeys while they performed sequential arm movements. The recordings have been made in various areas of the brain, including the basal ganglia (BG) which comprises the striatum (STR) and the globus pallidus (GP), the supplementary motor area (SMA), the pre-supplementary motor area (preSMA), the premotor cortex (PM), and the motor cortex (MI). After describing the experimental setup, a summary of results from various experiments is given below.

2.1. Experimental Setup

Monkeys were initially trained to respond with a push, pull or turn movement on a handle when presented with a red, yellow, or green color light, respectively, on display panel. After establishing the color-to-action mapping, they were required to sequence these elementary actions by observing a succession of color lights and recalling the appropriate movement at every step. The trials when color lights preceded movements were called, 'Visual Trials (VT)' and the other trials when there were no colored lights were called 'Memory Trials (MT)'. In a block of trials, monkeys experienced 5 VT followed by 6 MT. They were taught four such sequences till they achieved 95% success rate on each sequence. After this stage, single cell recordings were made while monkeys performed several blocks of trials. Each block used a different sequence and the order of appearance of different sequences varied semi-randomly. The end of a block is signaled by random flashing of lights which also indicated a change in the sequence.

2.2. Experimental Results

In the present study, we focus on results from the recordings of SMA, PreSMA, PM and parts of the BG. SMA recordings [1] revealed a significant activation exhibiting two main properties: i) *sequence-specific activity*: activity of some cells preceded specifically before the performance of a particular sequence (say Push-Pull-Turn) and ii) *transition-specific activity*: some cells were tonically active after performing a specific movement and before initiating a particular next movement (say, Pull followed by Push). In contrast, the activity of majority of cells in the MI was selective to single movements but not sequence-specific. While SMA cells were mainly active in the MT phase, cells in PM were active only during the VT phase and cells in MI were active during all phases of the task. Some cells in preSMA were active only during the very first VT, possibly signifying the start of a new sequence and thus specifying the need for a change in motor plan [4].

Schultz *et al* in [3] reviewd, among other things, motor sequence-specific activity in putamen neurons. Putamen forms the input side of BG. GP forms the output side of BG and some neurons here showed more activity during the MT phase and specifically during particular transitions in some sequences or specifically to a particular sequence of arm movements [5]. Some of these experimental findings will be used to formulate hypotheses about functional roles played by various areas involved in sequential arm movements. Fur-

[1]Accepted for presentation at ICONIP'98-Kitakiyushu, Japan, Oct 21–26, 1998

ther these hypotheses will be tested in simulations to see if the whole scheme works as expected.

3. NEURAL NETWORK ARCHITECTURE

Monkeys had extensive training in order to establish the color-to-action mapping before starting the sequence task. Accordingly it is assumed in the model that this mapping is already established and only sequence processing is investigated. The color-to-action mapping is believed to be established in the PM area. When a color light and trigger signal are presented, the correct motor action is decoded from the PM and is fed both to the BG, MI and SMA structures. Because of the availability of perfect mapping in PM, network's performance is flawless during VT. The aim of the rest of the network is to learn the sequence so that sequence recall is accurate in the MT phase when visual guidance is absent.

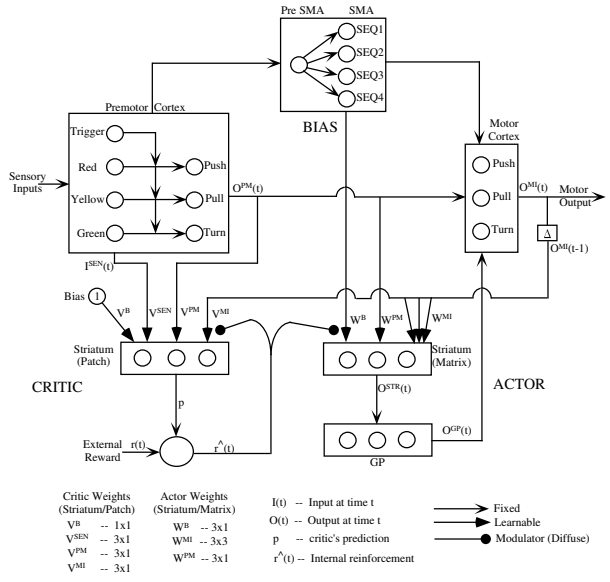

Figure 1: Sequence Processing NN Architecture

3.1. Design Hypotheses

It is hypothesized that BG learns by trial and error during the VT phase. Once learning is robust, network can recall the sequence during the MT phase also. In order to learn the sequential transitions, it is proposed that BG uses information about the previous movement and this information is fed from MI as a motor efferent signal. Thus, given the previous movement, learning in BG establishes the correct next action to perform. Since on the very first movement in a trial there is no motor efferent signal, it is proposed that a bias node that receives constant input is required. Thus bias node is active even on the very first movement in a trial and is geared toward recalling the first movement in MT. Once the first movement is performed, BG circuit kicks in and produces the correct subsequent transitions. It is proposed that SMA and preSMA structures implement the biasing mechanism.

It is proposed that learning in BG uses reinforcement learning scheme. Striatum forms the input side of BG and consists of matrix and patch compartments. The striatal matrix compartment functions as an actor for generating successive movements, whereas the patch compartment functions as a critic for generating internal reinforcement signals to enable learning in the absence of external reward signal (see Houk *et al* in [3]). Dopamine neurons in the substantia nigra pars compacta region have been shown to use the outputs from the BG and are specifically related to reward prediction. It is proposed that the internal reinforcement signal is related to dopamine activity and the relation of midbrain dopamine activity to reward prediction has been well documented [6].

3.2. Description of the Architecture

Network comprises an input module (PM), an output module (MI), an action prediction module (Actor), a reward prediction module (Critic), and a bias module (SMA modules) as illustrated in figure 1.

Input module This module located in the premotor cortex takes in sensory input of color and trigger information and generates the correct motor outputs. Color-to-action mapping is assumed fixed and preset. During VT, correct motor output is given by the PM output. However, during the MT only a small nonspecific trigger signal is fed as PM efferent signal and the correct movement has to be specified from elsewhere (using SMA and BG structures).

Actor The dorsal striatal matrix compartment acts as actor. It receives motor information - previous motor command from MI, current motor signal from PM and bias information from SMA. Given this information, it generates the correct movement to be output to the MI via GP. In this case GP is designed as a simple information carrier. Typical equations for the actor weights are shown below.

$$W^{PM}(t+1) = W^{PM}(t) + \alpha \cdot \hat{r} \cdot O^{PM}(t) \cdot O^{MI}(t) \cdot O^{STR}(t)$$

Critic Ventral straital patch compartment acts as the critic. It receives current sensory information, current PM output, previous motor movement from MI, and bias. Bias node has been found to be essential in preventing a blow-up of $\hat{r}$. Sensory and PM signals convey essentially the same information and hence are redundant. They enable establishing correct second-to-third movement transition in the MI actor weights. The critic weights from MI enable learning correct first-to-second movement transition in the actor weights. The internal reinforcement signal $\hat{r}$ enables learning of all the weights – both of the actor and the critic. Representative equations for the reward prediction, internal reinforcement signal, external reward, and the critic weights are shown below.

$$\begin{aligned} p(t) &= \sum V(t) \cdot I(t) \\ \hat{r} &= r(t) + \gamma \cdot p(t) - p(t-1) \\ r(t) &= \begin{cases} +1 & \text{SEQ is correct} \\ -1 & \text{Any Move wrong} \\ 0 & \text{Otherwise} \end{cases} \\ V(t+1) &= V(t) + \beta \cdot \hat{r} \cdot I(t-1) \end{aligned}$$

Bias module As will be shown in the results section, bias nodes situated in the SMA structures serve two fold purposes: i) they learn the order of sequence in the weight distribution and ii) they enable faster learning in the actor. The SMA

and preSMA circuit has not been simulated in detail is but provided for completeness of description.

Output module Output module located in the motor cortex receives the current PM output, GP output and generates a motor movement by using a winner-take-all strategy. During the VT, accurate information is provided by PM output and during MT, output from GP is the main source. A small noise signal is injected into the MI module in order to provide random activity during the initial blocks when the GP output has not reached a reliable accuracy level. An example equation for calculating striatal output is shown and other output equations are similar.

$$O^{STR}(t+1) = W^{PM}(t) \cdot O^{PM}(t) + \sum W^{MI}(t) \cdot O^{MI}(t-1)$$

4. SIMULATION RESULTS

4.1. Learning of Single Sequences

Simulations closely followed the event sequence of the actual experiment. Network was first trained with one sequence of movements continuously for 50 blocks. Each block consisted of 5 visual trials followed by 6 memory trials. During each visual trial, color and trigger inputs are given and the network is required to generate the appropriate action. A 3-bit binary vector is used as color input, with each bit representing one of the three colors (red, yellow, or green). Trigger signal is modeled as a small positive signal (0.1) injected nonspecifically to all the nodes. The motor output at both the PM and MI are also 3-bit binary vectors, with each bit representing one of the three actions (push, pull, or turn). Visual trials are followed by 6 memory trials. During each memory trial, input consisted of only a nonspecific trigger signal presented successively and the network is required to recall appropriate actions in the correct order. At the beginning of the experiment, all the critic weights are initialized to 0 and actor weights are set to a small positive value (0.01). All the node activities are set to 0. Learning rates of PM and Bias actor weights are set at a low value (0.025) compared to the MI actor weights (0.1) and the critic weight learning rate is set at 0.2. In the first part of simulatons, only one set of bias nodes for one sequence are considered in the bias module.

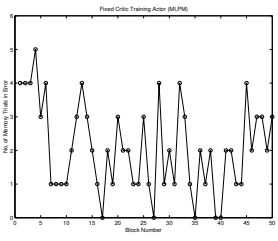

Figure 2: Performance with Fixed Critic & Basic Actor

In order to establish the utility of critic, a fixed policy critic is used to train a basic actor. A basic actor made decisions based on inputs from MI (previous movement information) and PM (current PM output). The polciy for the fixed critic was obtained by the experimenter based on heuristic trial and error search. The fixed policy consisted of small reward signal (0.1) when the sequence is correct, a low tonic reward signal for all the intermediate movements of the sequence, and a large punishment signal (-5.0) upon error. Fixed critic is also tested with a full actor with inputs from MI, PM, and bias nodes. A trial is treated erroneous and terminated if any one of the movements is wrong. The number of trials in error for each block is monitored.

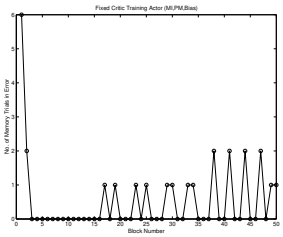

Figure 3: Performance with Fixed Critic & Full Actor

Fixed Critic with Basic Actor Network performance with fixed critic training a basic actor (prev MI and current PM) is shown in figure 2. Network performance is well above chance level, with a total of 103 trials in error out of 300 trials. In simulations it is observed that the actor weights from MI learn the transitions quite rapidly and if the first movement is correctly recalled then the subsequent movements are usually accurate. Thus there is a $\frac{2}{3}$ chance of making a mistake on the very first movement which then computed over all the 50 blocks (each with 6 MT) of simlulation gives a figure of 200 error trials for a chance-level-performance.

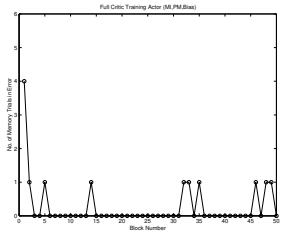

Figure 4: Performance with Full Critic & Full Actor

Fixed Critic with Full Actor Network performance with fixed critic trainig a full actor (prev MI, current PM, bias) is tested and the results are shown in figure 3. The performance improved dramatically with the bias node (26 error trials out of 300). It is also noticed that bias node also has the additional advantage of learning the order of the sequence in the weight distribution, that is, the first movement has the highest weight and the last one has the least weight. This phenomenon is later illustrated for the full critic.

Full Critic with Full Actor Finally a full critic is implemented with external reinforcement signal $r(t)$ as defined earlier. This signal consisted of reward (+1) if the sequence is correct, or punishment (-1) if the sequence is in error, or off (0) during other times. The performance of the network is further improved with the incorporation of full critic. As training proceeds, $\hat{r}$ value decreases gradually and enables stabilization of actor weights, whereas a fixed critic because of constant reinforcement signal does not facilitate weight

stabilization. Tests with full actor are shown in figure 4. The errors now reduced to 13 trials out of 300.

Relative performance profile is obtained by averaging results of four simulations in each of the above three conditions and summarized in figure 5.

Actor weights of bias nodes in the previous simulations with the full critic and full actor are averaged, by considering the weights corresponding to different movements (first, second, or third) separately. Results in figure figure 6 show clearly that because of training, the distribution of bias weights reflect the order of movements in the sequence. This fact enabled us to incorporate multiple sequences in the network.

4.2. Learning of Multiple Sequences

Simulations are performed with preset bias nodes to see the number of errors the network makes in coping with randomly changing sequences. Simulations are conducted by starting a trained network (trained on one of the four sequences) and on every fifth block a new sequence (randlomly chosen from the set of four sequences) is presented. In one case, bias nodes are set to appropriate values every time the sequence changed. In the other case, no such presetting of the bias weights is done. Comparative performance results are shown in figure figure 7. Results clearly show the advantage of bias switching.

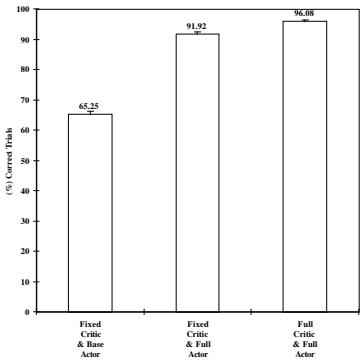

Figure 5: Relative Performance (averaged over 4 simulations)

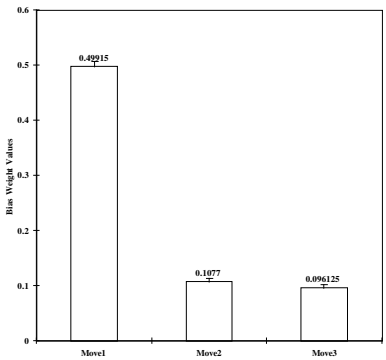

Figure 6: Bias Node Weights Reflect Sequence Order

5. DISCUSSION AND CONCLUSIONS

The proposal that BG learns and performs a sequence successfully by RL has been demonstrated in simulations. The key new idea here is that BG utilizes movement information from previous time step to learn the transition to the next movement. The chaining of successive movements is enabled further by biasing from SMA structures. It has been demonstrated that SMA structures enable switching to multiple sequences. However, in the simlations reported here switching is done externally and explicitly and is not adaptive.

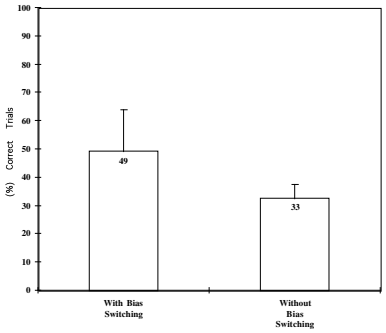

Figure 7: Performance on Multiple Sequence Testing

In a more detailed model, SMA structures can be seen to provide contextual information. However, the Learning of the bias nodes at the SMA level, that is, plasticity at the SMA level during multiple sequence performance, is not attempted in the current study. Also the selection of a bias node by actively monitoring the sequence performed during VT is also beyond the scope of present work.

References

[1] Tanji, J. & Shima, K. "Role for supplementary motor area cells in planning several movements ahead," *Nature*, **371**, 29, pp.413–416 (1994).

[2] Rosenbaum, D. A., *Human Motor Control*, San Diego, CA: Academic, 1991.

[3] Houk, J. C., Davis, J. L., & Beiser, D. G. (Eds.) *Models of information processing in the basal ganglia*, Cambridge, MA: MIT Press, 1995.

[4] Shima, K., Mushiake, H., Saito, N., & Tanji, J. "Role for cells in the presupplementary motor area in updating motor plans," *Proc. Natl. Acad. Sci. USA*, **93**, pp.8694-8698 (1996).

[5] Mushiake, H. & Strick, P. L. "Pallidal neuron activity during sequential arm movements," *J. Neurophys.*, **74**, 6, pp.2754-2758 (1996).

[6] Schultz, W., Dayan, P., & Montague, P.R. "A neural substrate of prediction and reward," *Science*, **275**, pp.1593–1599 (1997).