Modeling Reward Dependent Activity Pattern of Caudate Neurons

Thomas Trappenberg and Hiroyuki Nakahara
RIKEN Brain Science Institute, Lab. for Information Synthesis
2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan
{thomas,hiro}@brain.riken.go.jp

Okihide Hikosaka
Department of Physiology, School of Medicine, Juntendo University
2-1-1 Hongo, Bunkyo, Tokyo 113-0033, Japan
hikosaka@med.juntendo.ac.jp

Abstract

A hypothesis on the function of the basal ganglia was recently proposed based on reinforcement learning, however, only at conceptual level [1]. Our ongoing project is to quantify this hypothesis in cooperation with the experiment of reward-modulated activities of caudate neurons by Kawagoe et al. [2]. This paper, as our initial effort, aims to summarize the followings: (1) predictions of experimental results drawn from a minimal model, (2) comparison between these predictions and currently-obtained experimental results, (3) some extensions of a minimal model, (4) requirement for further experimental and computational studies.

1 Introduction

Inspired by the experiment on dopaminergic (DA) neurons in the substantia nigra pars compacta (SNc) [3], a recently-proposed hypothesis on the basal ganglia[1, 4, 5] is that the spiny neurons (Sps) in the striatum perform reinforcement learning based on the actor-critic scheme (see [5]) by use of a prediction error carried by DA neurons in the SNc. This hypothesis, however, is not yet quantitatively examined, particularly with respect to neural activities in the striatum. Our ongoing project is, based on this hypothesis, to have a quantitative analysis of neural activities in the striatum in cooperation with the ongoing experiment of reward-modulated neural activities in caudate (CD) nuclei (a part of the striatum) of monkeys by Kawagoe et al. [2]. In the following we first briefly introduce the experiment in [2] and outline some of their currently obtained results. Second, to start a quantitative analysis of those we introduce a minimal model based on the above hypothesis and state predictions given by this model. Third, we show that several modifications of the minimal model can explain some additional details of the experimental results. Finally, we discuss requirements for experimental and computational studies for further quantitative investigations.

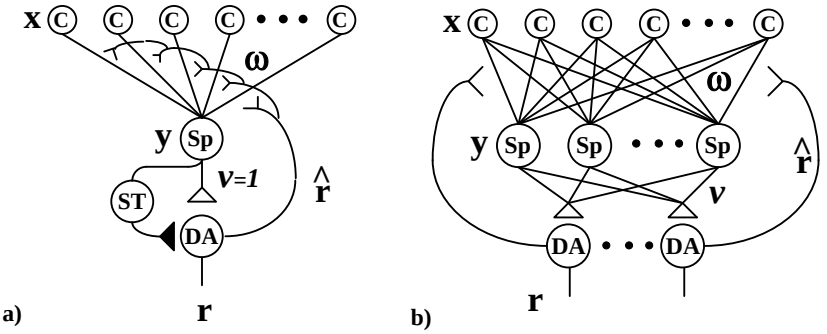

Figure 1: Model architecture of the caudate-SNc loop. Cortical neurons are denoted by **c**, spiny neurons by Sp, dopaminergic neurons by DA, and the subthalamus by ST.

2 Reward dependent caudate activities

Kawagoe et al. [2] have studied the activity pattern of Sps in CD in a memory-guided saccade task with several reward conditions. In each trial, one of four possible target locations was randomly illuminated for a short duration, and a monkey was required to saccade to the memorized target when the fixation point was removed. In each block of trials, a successful saccade to all of four possible targets was rewarded with liquid in ADR condition. In exclusive-1DR condition, only one of four targets comes with a reward, and in relative-1DR condition, one of four comes with a bigger reward than the other three targets. Kawagoe et al. [2] found a variety of neural activity patterns in these different reward conditions. Most salient results are (1) that many Sps in CD showed the biggest response for rewarded directions in each block of all four possible targets in exclusive-1DR and (2) that the reversed behavior of (1) was also observed, that is, some Sps responded increasingly only for the non-rewarded direction. Let us call such Sps of (1) and (2) reward-dependent (R-dependent) Sps and, particularly, such Sps of (2) reversed Sps for convenience. The activities of R-dependent Sps were also consistent in ADR in that the enhancement or suppression of responses occurred with all the rewarded directions. We will mention further details of results in [2] during the course of this paper.

3 Towards a quantitative description

In the actor-critic scheme (see [5]) the 'critic' works to evaluate a current state while the 'actor' works to choose an optimal action given the state. Figure 1a shows a possible neural implementation of one critic module based on some known anatomy of the basal ganglia [1]. A simplification we made in our minimal model is to neglect the subthalamic (ST) loop, which is suggested to convey information of the predicted value of the previous states to the DA as

in the temporal difference (TD) learning framework [4]. Even though that is very helpful for delayed reward in general, we neglect it in our minimal model because there is not much delayed reward aspect nor varying timing of reward in the experiments of Kawagoe et al [2]. Also, our minimal model is just a linear perceptron for sake of simplicity. Then, provided a cortical input vector $\mathbf{x}$, the output, $\mathbf{y}$, of the Sp in CD can be given by

$$y = \sum_j \omega_j x_j, \tag{1}$$

where ω denotes the cortico-striatal synaptic efficacy (weights).

The Sp project in an inhibitory way to a DA in the SNc, which will also receive an excitatory input r related to the physical liquid reward. These two inputs, y and r, determine the output $\hat{r}$ of the DA at the dopamine receptors of the Sp,

$$\hat{r} = (\hat{r}_0 + r - \sum_i v_i\ y_i)\Theta(\hat{r}_0 + r - \sum_i v_i\ y_i). \tag{2}$$

This rule differs from the one in [1] only slightly in that we restricted $\hat{r}$ to positive values with the step function Θ and included $\hat{r}_0$ to represent a constant background activity of DA. The cortico-striatal weights are only changed if the dopamine level is different from this constant background amount according to

$$\omega_{ij}^{new} = \omega_{ij}^{old} + \alpha(\hat{r} - \hat{r}_0)x_i, \tag{3}$$

where α is a learning rate. Note that we do assume here for simplicity that the timing of the reward is synchronized with the arrival of cortical activity.

3.1 The basic predictions of the minimal model

A major aim of this paper is to outline the basic predictions of this minimal model. We start here with the additional simplifying choice of using orthogonal input vectors representing the targets. The simulation of two blocks in exclusive-1DR is shown in Figure 2 (a & b), where we used input vectors which have an entry 1 for the illuminated location and 0 otherwise (e.g. $\mathbf{x} = (0, 1, 0, 0)$ for the target at location $j = 2$) and used the reward value $r = 1$ for a rewarded direction. The curves for the Sp activity of the model (Figure 2a) are similar, as a first approximation, to typical R-dependent Sp activities (not reversed ones) in [2]. The DA activity (Figure 2b) for an unexpected reward is large and converges to the background activity once the conditioning is established, which is in general agreement with cell recording data from Schultz et al. [3].

When the amount of reward is varied in exclusive-1DR, the saturated Sp response of the minimal model is linearly proportional to that amount, simply because the increase of Sp response stops when it correctly predicts this amount, i.e. $y = r$ (Figure 2c). We can not quantify this relation directly with the data in [2] since we do not have a direct measure for the reward input to DAs, r, but only indirectly as the amount of liquid in the experiment. However, the magnitude of r is expected to monotonically increase as the amount

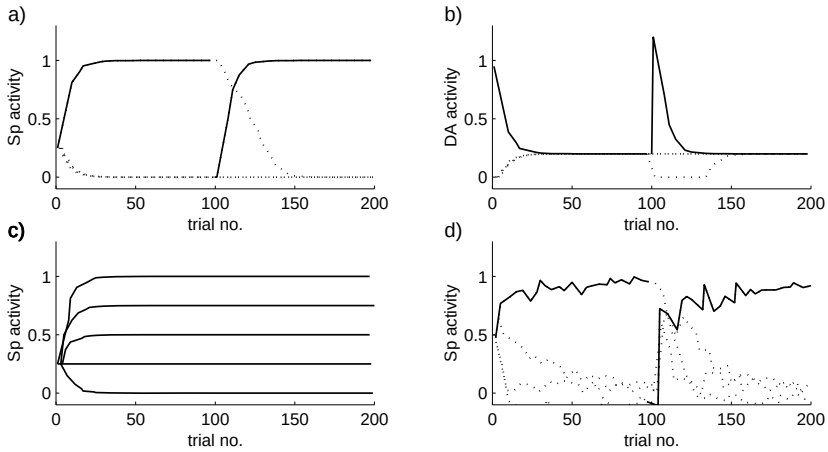

Figure 2: Simulated Sp and DA responses (rewarded as solid and non rewarded as dotted lines) in exclusive-1DR by the minimal model. (a,b) 2 blocks with 100 trials with orthogonal input vectors. (c) 1 block of 200 trials only for rewarded direction, superimposed for different reward values r between 0 and 1 with $\Delta r = 0.25$. (d) Same as (a) with partial overlapping input vectors.

of liquid increases in a certain range, which was also generally consistent in exclusive-1DR [2]. Then, the prediction of the minimal model should be a monotone relation between R-dependent Sp activity and the amount of liquid for exclusive-1DR and also for ADR and relative-1DR. However, this does not seem to hold for some data of the relative-1DR experiments and should be investigated further.

3.2 Partial overlapping input vectors

The choice of orthogonal input vectors seems rather special. Hinted by the topographical organization of the cortical areas that also have a topographical projection to CD, one way to relax orthogonal condition is to introduce partially overlapping input vectors, while the input vectors are still separable by our perceptron model (eq. 1). The example of such input vectors used in this study is

$$x_i^j = \begin{cases} 1 & \text{if} \quad i = j \\ 0.5 & \text{if} \quad i = j \pm 1 \\ 0 & elsewhere \end{cases} \qquad \text{with} \qquad i = 0, ..., 5 \tag{4}$$

where $j = 1, ..., 4$ denotes four possible targets.

Figure 2d Shows the Sp responses with these input vectors in the same setting as of Figure 2a. Note that fluctuation of Sp responses in Figure 2d is induced by the overlapping components of input vectors for different directions. The data in [2] do show fluctuations, and it should be explored if some parts of those fluctuations can be related to overlapping input vectors.

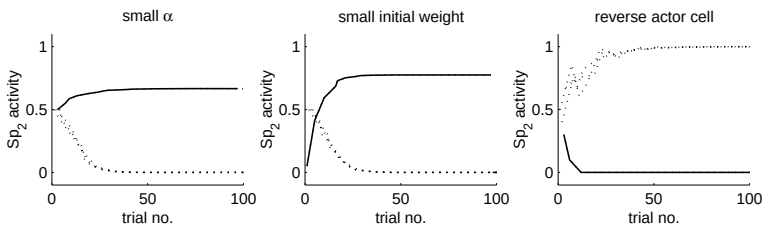

Figure 3: Some varieties of Sp activities in response to the rewarded target (solid line) and non-rewarded targets (dotted lines) in the model simulations. (a) Critic cell with small learning rate for rewarded target; (b) Critic cell with small initial weight to rewarded target; (c) Actor cell with reverse input.

3.3 Including multiple Sps

The minimal model with a single Sp and DA can be extended to include several Sps and DAs as shown in Figure 1b. Here, we only discuss the case of several Sps and one DA. Note that we have included synaptic weights $\mathbf{v}$ for the Sp-DA projections in Figure 1b. This architecture does include critics and actors in the sense that actors do not contribute to the reward prediction and can be simulated by zero weights to the DAs. There are different ways to combine the predictions (outputs) of different critic Sps. In the following we used simply a geometrical average of critic Sp predictions by setting $v_i = 1$.

We showed a typical response of R-dependent (not reversed) critic Sps by the minimal model in Figure 2a. A variation, however, exists in the degree of such enhancement for rewarded targets in Kawagoe et al. [2], though the response for rewarded targets was still the biggest in each block. There are several ways to achieve such responses in our model including (1) varying learning rates for some synapses, (2) varying the initial synaptic strength, (3) varying the magnitude of input, and (4) limiting the range of synaptic strength. We demonstrated the first two scenarios in Figure 3 (a&b) with a setting similar to the first block in Figure 1a except that we let $r = 2$ and included a second Sp (critic node with a unit projection to DA). The learning rate of the synapse for the target in Sp_2 was smaller than other synapses of Sp_1 and Sp_2 in Figure 1a, whereas the initial value of the same synapse in Sp_2 was smaller than others in Figure 3b. Both modifications result in a weakened activity ($y < 1$).

As mentioned in Section 2, Kawagoe et al. [2] found the reversed Sp behavior which enhanced their activity only for non-rewarded targets. This behavior can be simulated with a minimal model by treating such a Sp as an actor. We demonstrate this in Figure 3c with an actor cell (no projection to DA) which receives reverse input $1 - x$. This reversed response is very interesting to consider as actor in relation to the direct and indirect pathways scheme of the basal ganglia. In this scheme, the inhibition of Sps in CD results in facilitating movement in the direct pathway, whereas the excitation of Sps results in suppressing movement in the indirect pathway. Then, it is intriguing to ask whether the reverse Sps are related to the direct pathway.

4 Discussion

Based on the actor-critic hypothesis of the basal ganglia, this paper has outlined a minimal model which can be compared to the experiments on spiny (Sp) neurons in the caudate (CD) of monkeys under various reward conditions by Kawagoe et al. [2]. We first demonstrated that our minimal model exhibits a typical reward-dependent behavior of Sps found in [2] with qualitative feature resembling dopamine (DA) neuron responses in SNc similar to that found in [3].

We also outlined some extensions to capture more details of their experimental data, necessary for a more quantitative analysis. It was shown that some overlaps in the cortical input representation lead to fluctuations in Sp responses and we also noted that various patterns of reward-dependent responses found in [2] could be realized by considering several modifications such as different learning rates and initial preferences in synaptic projections. It was also possible to show reversed Sp response by simple modifications of the minimal model.

Many important questions remain to be investigated experimentally and computationally. The relation between DA responses and physical rewards should be quantitively investigated in the on-going experiments. It should be noted that a monotone relation between Sp responses in CD and the physical rewards seems not to hold in relative-1DR in the preliminary experiments. We need more experiment data, which are under way. Also, we should explore different models of several Sps, e.g., competitive scheme between Sps. We pointed out that reverse Sp responses may work as actor particularly in the direct pathway. A further experimental and computational analysis is required to relate their responses with behavior, or saccades. This paper only treated one DA but the effects of multiple DAs (see Figure 1b) on reinforcement learning in CD should be investigated with different models of Sp-DA projections.

References

[1] J.C. Houk, J.L. Adams and A.G. Barto, A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement, in *Models of Information Processing in the Basal Ganglia*, (eds. Houk, Davis and Beiser), MIT Press, Cambridge 1995

[2] R. Kawagoe, Y. Takikawa and O. Hikosaka., Basal ganglia translate motivation into oculomotor action., to be published.

[3] W. Schultz, P. Apicella and T. Ljungberg, Response of Monkey Dopamine Neurons to Reward and Conditioned Stimuli during Successive Steps of Learning a Delayed Response Task, J. of Neuroscience, 1993, 13(3):900-913

[4] R.S. Sutton, Learning to predict by the method of temporal differences, Machine Learning, 1988, 3:9-44

[5] A.G. Barto, Adaptive Critics and the Basal Ganglia, in *Models of Information Processing in the Basal Ganglia*, (eds. Houk, Davis and Beiser), MIT Press, Cambridge 1995