Learning How to Select an Action: A Computational Model

Berat Denizdurduran, Neslihan Serap Sengor

Istanbul Technical University, Electronics and Communications Engineering Department, Maslak, 34469, Istanbul, Turkey
{denizdurdu,sengorn}@itu.edu.tr

Abstract. Neurophysiological experimental results suggest that basal ganglia plays crucial role in action selection while dopamine modifies this process. There are computational models based on these experimental results for action selection. This work focuses on modification of action selection by dopamine release and a computational model capable of adapting it behaviour with parameter change is proposed. In the model, a dynamical system is considered for action selection and adaptation of action selection process is realized by reinforcement learning. The ability of the proposed dynamical system is investigated by bifurcation analysis. Based on the results of this bifurcation analysis, effect of reinforcement learning on action selection is discussed. The model is implemented on mobile robot and foraging task is realized where an exploration in an unfamiliar environment with training in the world is accomplished. Thus, this work fulfills its aim of showing the efficiency of brain-inspired computational models in controlling intelligent agents. [1]

Keywords: Basal Ganglia Circuits, Action Selection, Reinforcement Learning, Bifurcation Analysis, Cognitive Robotics.

1 Introduction

Basal ganglia (BG) circuits are involved in a wide range of brain functions, such as perception, learning, memory forming, besides being effective in motor functions. Their functions in cognitive processes as action selection (AS), goal-directed behavior and selective attention have been studied throughly in recent years [1, 2]. Existence of at least five different loops of cortex-basal ganglia-thalamus has been suggested in [3]. Each loop has a different role in cognitive tasks, we consider one of the prefrontal loops which plays a role in AS. To understand the mechanism giving rise to cognitive processes, one has to consider neurotransmitter systems as neurotransmitters have constitutive effect on cognitive processes. Amongst eight different dopaminergic pathway, nigrostriatal pathway modulates AS process in BG [4] .

[1] This work is supported by TUBITAK Project No: 111E264.

We aim to model the modulatory effect of dopamine on AS with a computational model which can establish sufficient functionality to exhibit relevant behaviour in embodied robotics. The computational model has been developed at the system level [5] as in the well-known work of Prescott et. al. [6]. In [6], saliencies that exist in a tight competition effectively switch the behaviour of the computational BG model realizing the AS mechanism in the embodied architecture. The saliencies which control the behavior of the mobile robot are generated in the motivational and sensory sub-systems as a priori coefficients in [6]. Although the work in [6] is important as it shows that the biologically plausible models of the BG can be used for the control of physical devices, it is only a first step as it lacks the ability of determining saliencies by a learning process. We focus especially on this aspect and proposed a model [5] capable of generating an adaptive process where the parameters of the model modify the behaviour of BG circuit taking part in AS. So, [5] gives a computational model where the model parameters W_c corresponding to the saliences are rearranged while mobile robot realizes the foraging task. Here, we will extend the model in [5] by adapting parameter W_r corresponding to dopamine to mimic the modulatory effect of this neurotransmitter. Also, direct pathway along with indirect pathway is included here and including direct pathway allows us to explain the dopamine release on Str with bifurcation analysis of another parameter W_d. Based on the bifurcation diagrams for these parameters, we are able to explain not only the importance of dopamine on determining the saliencies but also differences between each salience circumstances.

2 Selecting an Action

Brain inspired computational models have an important role in understanding the behaviour of the human beings. One of the interesting questions of contemporary neuroscience is how primates make appropriate decisions at the right time, literally defined as *Action Selection*. There are a number of computational models evaluating the neurophysiologic work on BG and related neural substructures taking part in AS [5–7]. Recently, it is claimed that in the learning process of signal detection, process of BG circuits are somehow implicated exacly same as Reinforcement Learning (RL) [2]. So, difference between our work and [2,6,8] is to consider the BG circuits and RL together. The function of BG model can be described as an effectively switchboard mechanism [6] when different possibilities exist. These possible actions differ from each other with their saliencies.

The major input station in BG is striatum (Str) and this subcortical area is divided into subsections where two dopamine receptors are effective. One of these receptors is D1 subtype and these regions project primarily to the output nuclei of BG to inhibit them. These output nuclei, Substantia Nigra pars Reticulate and Globus Pallidus internal (SNr/GPi), in turn inhibits the cortex (Ctx) through thalamus (Thl). This pathway is called direct pathway and is a main part of selection mechanism. Whereas in the indirect pathway, D2 receptors are effective and they have inhibitory effect on Globus pallidus external (GPe) to disinhibit

subthalamic nucleus (Stn) through GPi and it is claimed that indirect pathway antagonizes the direct pathway by suppressing unwanted movements [9]. Based on this discussion, a representation of BG circuit with its connection to related neural substructures are given as follow by difference equations:

$$\begin{aligned} Ctx(k+1) &= f(\lambda Ctx(k) + Thl(k) + W_c I(k)) \\ Str(k+1) &= W_r f(Ctx(k)) \\ GPe(k+1) &= f(-Str(k)) \\ Stn(k+1) &= f(Ctx(k) - GPe(k)) \\ GPi(k+1) &= W_d f(Stn(k) - Str(k)) \\ Thl(k+1) &= f(Ctx(k) - GPi(k)) \\ f(x) &= 0.5(tanh(3(x-0.45)) + 1) \end{aligned} \tag{1}$$

The dimensions of these vectors are determined by the number of actions to be selected. $I(k)$ represents the input and W_c denotes the efficiency of this input. There are two more bifurcation parameters in the model where W_r denotes the effect of dopamine release on Str and W_d denotes the correlation between the direct and indirect pathways in the circuit. These parameters are effective in selection mechanism, so bifurcation analysis, considering these coefficients, will be given in Section 3.1. When there exist three actions to be selected, three dimensional salience-space comprises different subspaces where W_c, W_r and W_d reshapes this space. Once we select a priori coefficients, a classification of three subregions can be seen in Fig. 1a-c.

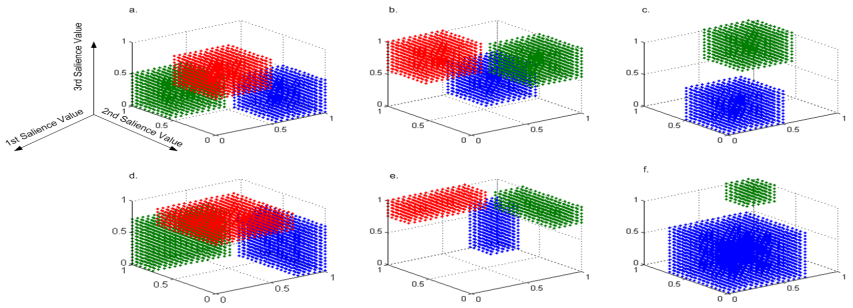

Fig. 1. Salience-space can be reshaped by significant parameters. In a., b. and c. $W_c = 0.8, W_r = 0.3, W_d = 1$ and in d, e. and f. $W_c = 0.6, W_r = 0.3, W_d = 1$. The upper figures illustrate a desired space configuration, while the bottom figures denote a space configuration where to make a choice for an action is weak. When all the figures located at each row are brought together, they constitute the whole space.

In Fig. 1a there are subregions of salience-space where only one of each salience determine the action to be selected. This means, model selects an appropriate action if the relative salience value is in one of these regions. In Fig.

1b, two salience combinations out of three determine the action to be selected. The all or none selection regions is given in Fig. 1c. As mentioned above, the coefficients W_c, W_r, W_d have a role to reshape these subregions (Fig. 1d-f). This phenomenon will be explained further in Section 3.1. As it can be followed in Fig. 1b and 1c, in some regions, system selects more than one action, simultaneously. In this case the system is not be able to decide on an appropriate action. So, to avoid this undecisive situation, RL has been implemented to the model and once the learning is completed, system selects the right decision at the right time in all circumstances.

3 Learning to Select an Action

Nonlinear dynamical system approach for modeling the neural structures gives us the possibility to control the system with bifurcation parameters. Therefore, in this work, neurocomputational model for AS circuit given with Eq.1 is investigated using bifurcation analysis and this investigation confirmed that the given model can be modified to obtain appropriate behaviour. To realize this modification, RL is utilized and with RL the parameters of the system corresponding to bifurcation parameters are updated to model learning to select appropriate actions in an unfamiliar environment. Thus the modulatory effect of dopamine on Str is explained by bifurcation analysis. The actions selected are observed in Ctx, so though dopamine effects Str, the overall effect of AS process is investigated by the signal obtained at the Ctx component of the system given by Eq.1.

3.1 Bifurcation Analysis

The model satisfies the following fold (saddle-node, tangent) bifurcation (FB) conditions where one stable and one unstable fixed points exist in the generated map at the same time and then they disappear [10]. For n-dimensional dynamical systems, following conditions are required, Jacobian matrix A_0 at FB has an eigenvalue $\mu_1 = 1$, n_{stable} eigenvalues with $|\mu_i| < 1$ and $n_{unstable}$ eigenvalues with $|\mu_i| > 1$, with $n_{stable} + n_{unstable} + 1 = n$. We claim that the system given in Eq.1 undergoes two FB consecutively and this gives rise to the switchboard like mechanism in AS circuit. These two consecutive FB occuring according to the bifurcation parameter W_c, which corresponds to the efficiency of input value in our context, is given in Fig. 2a.

There exist one stable fixed point which is near "0" and then it disappears and later reappears around "1" while the bifurcation parameter $W_c \in [0, 1]$. As the bifurcation parameter reaches FB when $W_c \tilde{=} 0.06$, the stable fixed point at "0" collapses with unstable one and disappears while another stable fixed point around "0.9" borns. Thus there exist a region where unstable fixed point is observed. Between two FB point, the eigenvalue conditions are satisfied as it can be seen in Fig. 2b. The stable fixed point near "0" means that system cannot select an action, on the other hand the stable fixed point at "0.9" means that the

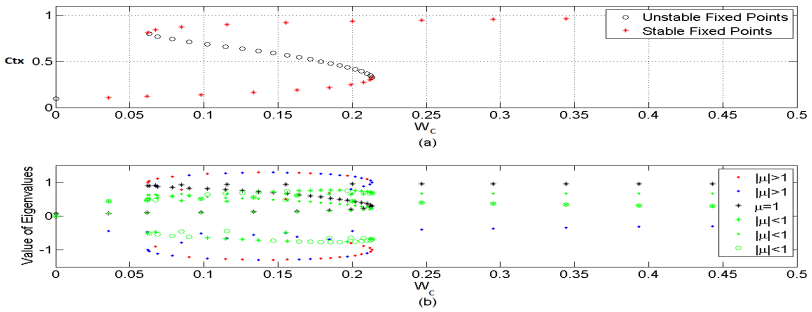

Fig. 2. a. The circles denote the unstable fixed point, dots denote the stable ones. System has two different stable fixed points around the FB. One of them corresponds to not-selecting an action (near "0"), the other one correponds to selecting an action (near "1"). The FB represents the bi-stable phase portrait in the system. b. Change of value of the eigenvalues with parameter W_c.

system selects an action. Around the bifurcation point there are two domains of attraction between "0.06" and "0.22". When we fix the parameters in this region, the proposed model decides to select/or not select an action depending on the initial conditions. This explains how the system given by Eq. 1 accomplishes AS according to the input value W_c. So to observe the effect of input value W_c while changing W_r, the two parameter bifurcation diagram given in Fig. 3a is obtained.

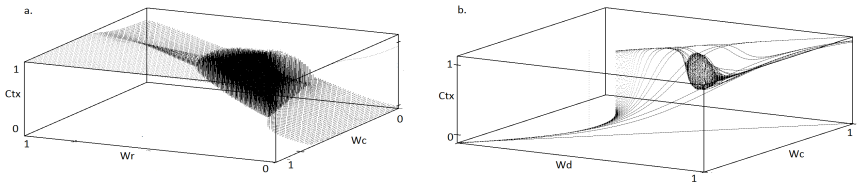

Fig. 3. a. Two parameter bifurcation diagram for W_r and W_c. b. Two parameter bifurcation diagram for W_d and W_c.

Here while $W_r < 0.23$ and $W_c \in [0, 1]$ there is one stable fixed point and while keeping $W_c \in [0, 1]$ but changing $W_r \in [0.23, 0.51]$ non convergent solutions (quasi-periodic) begin till another stable fixed point appears. This quasi-periodic solutions mean that system cannot decide which action to select. How the system's salience-space is rehaped can be explained with this bifurcation analysis. As it can be seen in Fig. 1, with different parameter values, different salience-space configuration is formed. The ability of the system to select an action is increased or decreased with these parameters. This analysis explains

the dopamine effect on Str: when the stable fixed point is located around "0" even the input value is high enough, system cannot select an action (This is shown in Fig. 1, where the area of the selection is decreased in Fig. 1d, and the non-selection area is increased in Fig. 1f), on the other hand when the dopamine release is increased, system selects the action in all circumstances. There exists another bifurcation parameter, W_d, which corresponds to correlation of the direct and indirect pathways in the BG circuit. The bifurcation analysis for W_d shows that the model's behaviour becomes unstable when $W_d \in [0.67, 0.75]$ (Fig. 3b). If this connection increases constantly, system possibly selects an undesired action and it means that indirect pathway cannot antogonizes the direct pathway to select an appropriate action.

3.2 Reinforcement Learning

Goal-Directed behaviour comprises two different concepts, AS and learning. Since the information process in BG points the relationship between AS and RL, temporal difference (TD) learning has been implemented into the model. To explain the architecture how RL effects the AS, the following pseudocode for the model is given:

Algorithm 1 Calculate W_r

Require: $0 < W_c < 1 \wedge W_d < 0.67$
Ensure: $\mu = f_x(0,0) = 1$
  $\forall |\mu_{stable}| < 1 \wedge \forall |\mu_{unstable}| > 1$
  $n_{stable} + n_{unstable} + 1 = n$
  $Ctx \Leftarrow W_c I$
  while $Error > |0.04|$ **do**
    for $k < 30$ **do**
      $BasalGanglia(k+1) \Leftarrow f(BasalGanglia(k))$
      $N \Leftarrow Ctx(k)$
    end for
    if $N < 0.85$ **then**
      $Reward \Leftarrow 1$
    else
      $Reward \Leftarrow -1$
    end if
    $Value(:, kk+1) \Leftarrow val * I'$
    $Error(:, kk) \Leftarrow Reward + \gamma * Value(:, kk+1) - Value(:, kk)$
    $val \Leftarrow val + nu_{val} * Error(:, kk) * I$
    $W_r \Leftarrow W_r + nu_r * [Error(:, kk)]' * W_r. * f(N). * N$
    $kk \Leftarrow kk + 1$
  end while

In TD, step size of the learning depends on the discount factor γ and the ultimate goal is to reach "0" value for the error in expectation through the RL, as it can be followed in Fig. 4a.

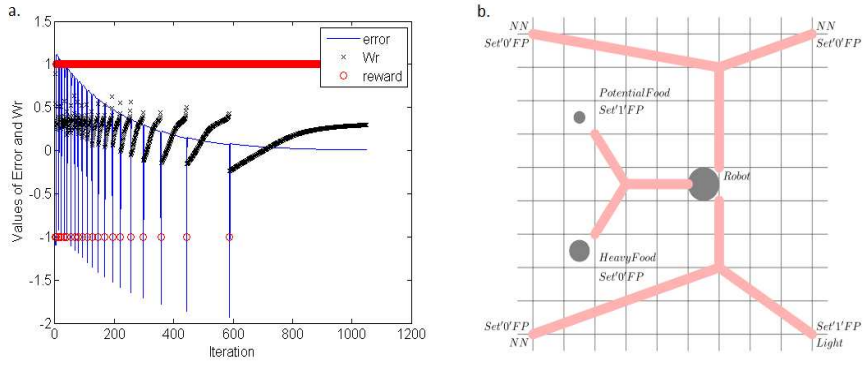

Fig. 4. a. In TD, system evaluates an error in expectation every step time which depends on reward value. Once the error in expectation reaches approximately "0", W_r (dopamine effect on Str) is fixed at the ($W_r = 0.28$). b. Circles illustrate the robot, heavy food and potential food, respectively in descending order. NN means 'Not a Nest' and the light illustrates the nest position. 'Set "0" FP' and 'Set "1" FP' show the which fixed point is occured in corresponding situation.

4 An Application: Foraging Task

Based on the discussion carried out on bifurcation analysis above, AS system and RL are considered together to solve the foraging task. In a previous work [5], there were three saliencies which defined the wheels, gripper and light/distance sensors activation. Robot sought the food during the exploration and was able to avoid the obstacles. The complete architecture of the task can be found in [5]. Here, we also considered changing the parameter W_r together with input values W_c corresponding to the salience values. When robot finds a cylinder in front of it, it picks up and calculates the weight by evaluating gripper aperture range. If it finds a heavy cylinder, robot immediately dismisses the cylinder. In this case, the nonlinear system determining the robot actions is at the fixed point near "0". On the other hand, when the fixed point reaches "1" robot finds one of the potential foods. The same idea is used for recognizing the nest position. If robot finds a light source in any corner, system's fixed point reaches "1" otherwise it is "0". It is worth remarking that the nest position and food recognition are individual saliences. The illustration of how the robot moves during the task can be found in Fig. 4b.

The actions denoted in Fig. 4b starts after the learning of the salience recognition is completed [5]. First, system has to understand the differences between the saliencies and then be able to recognize the importance of individial circumstances. In other words, the robot learns when exactly it has to move, recognize the cylinders, i.e., it has to stop and try to pick it up whatever the weight is and finally recognizes the nest to deposit the food in any corner. After this learning process, we considered the bifurcation analysis to learn to track the differences

between the potential food/heavy food and nest position, till the robot is able to recognize the food and is able to match the light to the nest position.

5 Discussion and Conclusion

Nonlinear dynamical system approach for modeling the neural structures gives us the possibility of understanding the phenomenon modeled by bifurcation analysis. Therefore, in this work, a neurocomputational model for AS circuit is investigated considering bifurcations and it is shown that training to adapt the choices of a mobile robot is possible by RL. This approach can be further used to establish a framework for understanding the cause of physiological diseases related with BG circuits. The level of the dopamine has a role in different physiological disorders as it is well-known that loss of dopamine level in BG circuit causes Parkinson's disease which means difficulty in accomplishing an action as in (Fig. 1d-f), on the other hand excessive dopamine level in BG circuit causes Huntington's disease which means choosing more than one action at a time.

References

1. Redgrave, P., Prescott, T.J., Gurney, K.: The basal ganglia: A vertebrate solution to the selection problem? Neuroscience, 89, 1009–1023 (1999).
2. Dayan, P., Daw, N.,D.: Decision theory, reinforcement learning and brain. Cognitive, Affective & Behavioral Neuroscience. 8 (4), 429–453, (2008)
3. Alexander, G.E., Cruther, M.D., DeLong, M.R.: Basal ganglia-thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. Progress in Brain Research, 85, 119–146 (1990).
4. Haber, S.N.: The reward circuit: Linking primate anatomy and human imaging. Neuropsychopharmacology Reviews, 35, 4–26 (2010).
5. Denizdurduran, B., Sengor, N.S.: A Realization of Goal-Directed Behavior, Implementing a Robot Model Based on Cortico-Striato-Thalamic Circuits, Proceedings of The 4^{th} International Conference on Agents and Artificial Intelligence, 289–294, (2012)
6. Prescott, T.J., Montes-Gonzales, F.M., Gurney, K., Humpries, M.D., Redgrave,P.: A Robot Model of the Basal Ganglia: Behaviour and Intrinsic Processing. Neural Networks, 1–31, (2006)
7. Houk, J.C., Bastianen, C., Fansler, D., Fishbach, A., Fraser, D., Reber, P.J., Roy, S.A, Simo, L.S.: Action selection and refinement in subcortical loops through basal ganglia and cerebellum. Phil. Trans. R. Soc. B, 29 vol. 362 no.1485, 1573–1583 (2007)
8. Schultz, W., Dayan, P., Montague, P.R.: A Neural Substrate of Prediction and Reward. Science 275, 1593–1599 (1997)
9. Frank, M.J.: Computational models of motivated action selection in corticostriatal circuits. Current opinion in Neurobiology, 21, 381–386, (2011)
10. Kuznetsov, Y.A.: Elements of Bifurcation Theory, In Marsden, J. E. and Sirovich, L. editors, Second Edition, Applied Mathematical Sciences 112, (1998)