April 27, 2006 17:29 00054

International Journal of Neural Systems, Vol. 16, No. 2 (2006) 111–124

World Scientific
www.worldscientific.com

THE ROLE OF THE BASAL GANGLIA IN EXPLORATION IN A NEURAL MODEL BASED ON REINFORCEMENT LEARNING

D. SRIDHARAN
Program in Neuroscience, Stanford University School of Medicine,
Stanford, California 94305, USA
dsridhar@stanford.edu
www.stanford.edu

P. S. PRASHANTH
Electrical and Computer Engineering, Iowa State University,
Ames, Iowa 50011, USA
prash@iastate.edu
www.iastate.edu

V. S. CHAKRAVARTHY*
Department of Biotechnology, Indian Institute of Technology — Madras,
Chennai 600036, India
schakra@ee.iitm.ac.in
www.iitm.ac.in

We present a computational model of basal ganglia as a key player in exploratory behavior. The model describes exploration of a virtual rat in a simulated water pool experiment. The virtual rat is trained using a reward-based or reinforcement learning paradigm which requires units with stochastic behavior for exploration of the system's state space. We model the Subthalamic Nucleus-Globus Pallidus externa (STN-GPe) segment of the basal ganglia as a pair of neuronal layers with oscillatory dynamics, exhibiting a variety of dynamic regimes such as chaos, traveling waves and clustering. Invoking the property of chaotic systems to *explore state-space*, we suggest that the complex exploratory dynamics of STN-GPe system in conjunction with dopamine-based reward signaling from the Substantia Nigra pars compacta (SNc) present the two key ingredients of a reinforcement learning system.

Keywords: Basal ganglia; chaotic systems; exploration; dopamine reward; reinforcement learning.

1. Introduction

The basal ganglia (BG) are a group of sub-cortical nuclei buried deep within the telencephalon. The term basal ganglia is often used to refer to a collection of nuclei including the corpus striatum (caudate nucleus and the lenticular nucleus, which includes the putamen and the globus pallidus) and other subcortical nuclei such as the subthalamic nucleus (STN), the substantia nigra (SN) (consisting of the pars compacta (SNc) and pars reticulata (SNr)), and the pedunculopontine tegmental nucleus (PPTg). The putamen and caudate are often collectively referred to as the striatum.[1]

The basal ganglia have long been afforded the role of a *gate* or a *selector* among action representations in the cortex competing for limited resources. In fact some have called these nuclei the Vertebrate Solution to the Selection Problem.[2] They have also been implicated in sequence generation,[3] and working memory.[4] Their dysfunction in motor disorders such as Parkinsons disease has been well documented.[5] In the present work we assign yet another role to the basal ganglia (specifically to the STN-GPe segment

*Corresponding author.

within the basal ganglia) — as a system that provides the exploratory drive needed in activities like navigation, foraging and the like.

The term *reinforcement* learning is, in fact, derived from animal learning studies in experimental psychology. This form of learning signifies the coupling between occurrence of an event and a response, wherein the occurrence of the event tends to increase the probability of occurrence of the response under similar external conditions if the response should result in a general improvement of the state of affairs of the animal.[6] Reinforcement learning is, thus, that form of unsupervised learning where the training signal is in the form of a global scalar known as the *reward*.

Neural network models of reinforcement learning use stochastic output units for exploration of output state space, i.e., the only way the network can know the correct response to an input is by *guessing*. The probabilistic output neurons ensure that the system thoroughly explores the space of responses to a certain input so that, the correct response when it occurs can be reinforced. Chaotic systems have been known to exhibit exhaustive exploration of their state-space. It is well-known that a network of nonlinear oscillators is intrinsically chaotic.[7] Recently a network of oscillators has been proposed as a model of motor unit recruitment in skeletal muscle. The complex dynamics of the network is used to model desynchronized activity of motor neurons in healthy muscle.[8] Oscillatory neural activity is known to exist in several structures in the brain including the basal ganglia, hippocampus, sensory cortices etc. Oscillatory dynamics in the basal ganglia have been observed at the level of the Sub-Thalamic Nucleus — Globus Pallidus network.[9,10] The STN-GPe network, depending on the patterns of the interconnections and values of the interconnecting weights, has been shown to support three general classes of sustained firing patterns: clustering, propagating waves, and repetitive spiking that may show little regularity or correlation.[10] Furthermore, it has also been demonstrated that each activity pattern can occur continuously or in discrete episodes.

The mesencephalic dopaminergic input to the basal ganglia might help modulate the activity of the STN-GPe loop by serving as a reward signal to these units in the so-called indirect pathway in the basal ganglia.[3] We hypothesize that the complex oscillations of the STN-GPe segment within the basal ganglia provide the exploratory dynamics necessary for reward-based or reinforcement learning.

The paper is organized as follows: In the following section we elaborate on the exact role of the basal ganglia in exploratory behavior. We then present a computational model of the STN-GPe segment as a network of oscillatory neurons. In the next section we evaluate this network in the context of a simulated *waterpool experiment*. In real versions of these experiments a rat has to learn the location of a submerged (invisible) platform in a pool of water based on spatial cues placed around the pool. We then discuss pathologies arising out of disruption of the dopamine reward signal due to simulated lesions of the mesencephalic dopaminergic centers or the SNc. Finally we conclude with a discussion on the unique dynamics exhibited by the oscillatory network and its significance in a biological context, and the evidence from neurophysiology underlying our model.

2. The Model

2.1. *Description of the basal ganglia model*

The general consensus regarding signal flow in basal ganglia is as follows[11]: The striatum (caudate/putamen) receives and integrates input coming from the cortex (visual/motor). The input is often in the form of competing action representations, the competition among which needs to be resolved and this resolution, we propose, takes place at the level of the striatum itself. The striatum then sends inhibitory input to a specific region in the GPi (EP) or the SNr which then disinhibit a particular region at the level of the thalamus (motor selection) or the superior colliculus (in case of visual selection) respectively to carry out the selected action. It is apparent that the most disinhibited action will eventually win the competition and be selected (Fig. 1). Further, there exist inhibitory projections from the GPe to the STN, and excitatory projections from the STN back to the GPe. The STN-GPe system is thus ideally placed to produce oscillations.[9]

We propose a simplified three layer architecture (Fig. 2) for the basal ganglia consisting of an input layer representing visual input, a hidden layer representing the oscillatory STN-GPe system, and an

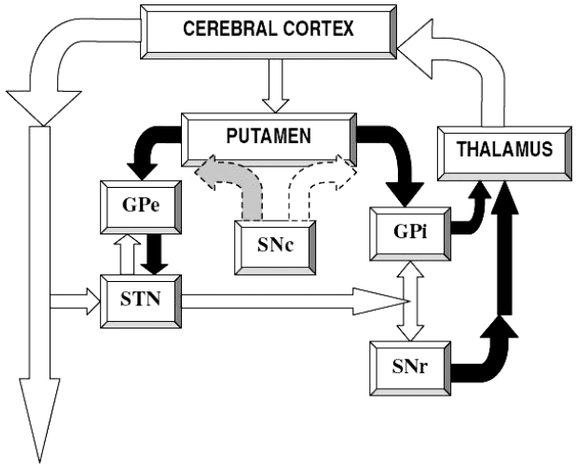


Fig. 1. Representative block diagram of the basal ganglia neuroanatomy. White arrows indicate excitatory connections, while black arrows indicate inhibitory connections. Arrows with dotted borders indicate dopaminergic excitatory (white) and inhibitory (grey) input from the SNc. The putamen is the striatal segment involved in motor activity and selection. The Direct pathway consists of the inhibitory interconnections between the putamen and the GPi. The Indirect pathway involves the STN-GPe loop which, we propose, is responsible for the exploratory behavior. The output stage of the basal ganglia, viz. the GPi (in case of the motor loop) or SNr (in case of the visual loop), decides the action modality to be selected by disinhibition of the thalamus. Dopamine secreted by the SNc modulates striatal activity, mainly by inhibiting the indirect and facilitating the direct pathways.[1] Here we propose that dopamine regulates the Striatal-GPe connection weights leading to reinforcement learning. Other pathways involving sensory/motor cortical projections are not considered in the present model.

output layer representing the selected movement of the animal for the given input. In neurophysiological terms, the input to the first layer represents the unresolved, competing visual representations arising from the cortex reaching the striatum (caudate/putamen). The activity of the hidden layer consisting of oscillatory units corresponds to the lumped activity of the STN-GPe oscillatory network. The output at the final layer corresponds to the motor output at the level of GPi (or EP in the case of the rat) to which the STN projects. It is this motor modality that is inhibited at the GPi which leads to disinhibition at the level of the thalamus which eventually leads to activity corresponding to a particular action or action sequence being initiated in the motor cortex.

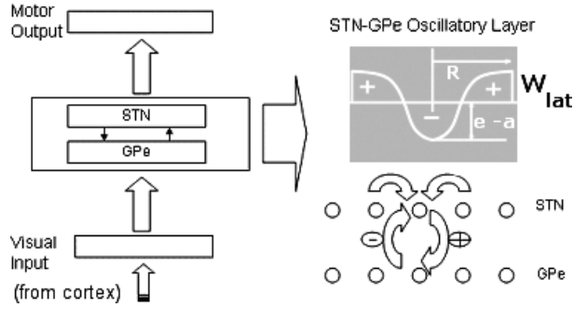


Fig. 2. Overall architecture of the network is shown on the left. The visual input from the cortex is presented to the STN-GPe layer through a set of weights. Interaction between STN and GPe layers produces complex oscillatory activity. Activity of STN layer is fed to the output layer. A detailed diagram of connectivity in the STN-GPe layer is shown on the right. The gray box depicts the lateral connectivity profile of STN neurons with an inhibitory center and excitatory surround (ref. Eq. (7) in text).

We suppose that the reward signal received from the limbic system is translated into fluctuations from baseline levels of dopamine secreted by the mesencephalic dopaminergic system comprising of the Ventral Tegmental Area (VTA) and the SNc (substantia nigra pars compacta) nucleus of the basal ganglia.[12] There is an increase in the overall level of dopamine when there is a positive reward and a corresponding decrease in case of a negative reward, with the magnitudes of these fluctuations being correlated with the magnitudes of the rewards. This reward signal (i.e., level of dopamine) is propagated as a global reinforcement signal that serves to modify the synaptic weights both among striato-pallidal (STR-GPe) and pallido-peduncular (STN-GPe to GPi) projections thereby leading to learning, i.e., the generation of a potentially rewarding motor output based on the current sensory input.

In the following sub-sections, we describe the system of equations that governs the behavior of our neural network model:

2.1.1. *The input layer*

The activity, $\mathbf{x}$, of the input layer, is projected via the weights, W^{in}, to the STN-GPe layer as input I, which is computed as:

$$I_{ij} = \sum_k W^{\text{in}}_{ij;k} x_k \tag{1}$$

2.1.2. *Oscillatory STN-GPe layer dynamics*

Unlike neurons in the input layer, which do not have lateral connections, each neuron in the STN is connected to a 2-dimensional neighborhood. Connections between STN and GPe are assumed to be one-to-one (Fig. 2). If V_{ij} represents activity of STN neuron (i, j) and S_{ij} represents activity of GPe neurons, the STN-GPe interaction may be described as:

$$\frac{\mathrm{d}u_{ij}}{\mathrm{d}t} = -u_{ij} + \sum_{i',j' \in A} W^{lat}_{iji'j'} \tanh(\lambda u_{i'j'}) - S_{ij} + D_{ij} + I_{ij} \quad (2a)$$

$$V_{ij} = \tanh(\lambda u_{ij}) \quad (2b)$$

$$\frac{\mathrm{d}S_{ij}}{\mathrm{d}t} = -S_{ij} + V_{ij} \quad (3)$$

where A is the neighborhood of lateral connections in the STN with weights $W^{lat}_{iji'j'}$ between neurons (i, j) and $(i'j')$. Equations similar to (2) and (3) have been used by Gillies *et al.*[9] to describe their model of STN-GPe interaction. An analysis of the system of equations, similar to (2) and (3) above, for a single oscillator, demonstrating the existence of a unique stable *limit cycle*, is provided as an Appendix.

A special feature of our equations, D_{ij}, represents the effect of dopamine on STN layer activity. In line with thinking that implicates dopamine in reward signaling[12] we assume that the dopamine levels available in STN (due to inputs from mesencephalic dopamine systems) determine activity levels of the oscillatory units of STN. The higher the dopamine levels, the greater the number of units that are active at any given instant. This is achieved by introducing auxiliary dynamics as described next.

Let $D(0 < D < 100)$ be the level of dopamine available in STN. When $D = 0$, the least number of STN neurons are ON; when $D = 100$ nearly all STN neurons are ON. This is achieved by varying D_{ij} as a function of the number of active STN neurons. Let e denote the discrepancy between the actual number of active units, D_a, and dopamine levels at a given instant, then:

$$D_a = \sum_{i,j}^{N} V_{ij} \quad (4a)$$

$$e = D - D_a \quad (4b)$$

where N is the number of neurons in the STN grid. This discrepancy or *error* is accumulated in E as,

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \tanh(\lambda e) \quad (5)$$

and presented as input D_{ij}, where,

$$D_{ij} = E - \frac{N}{2} \quad (6)$$

to STN neurons as in Eq. (2).

The lateral connections, $\left(W^{lat}_{iji'j'}\right)$, within STN layer are assumed to be translation invariant and are given by:

$$W^{lat}_{iji'j'} = \epsilon - ae^{\left(-{r_{lat}}^2/{\sigma_{lat}}^2\right)}, \quad \text{for } r < R$$
$$= 0, \quad \text{otherwise} \quad (7)$$

where $r_{lat} = [(i - i')^2 + (j - j')^2]^{1/2}$; a controls the depth of the Gaussian bell function and σ_{lat} its width; and R is the neighborhood size. Thus each unit has a negative center and a positive surround; the relative sizes of center and surround are determined by ϵ. Smaller ϵ implies, more negative lateral STN connections, which tends to decorrelate oscillations of STN neurons. In the absence of input from the input layer (i.e., $I_{ij} = 0$), as ϵ is varied from 0 to a, the activity of STN-GPe system exhibits three different regimes: (1) chaos, (2) traveling waves, and (3) clusters (Fig. 8). Operation of the network in the first regime, viz. chaos, is the most crucial since it is the chaotic dynamics in the STN-GPe layer that makes the network extensively explore the output space.

2.1.3. *Dopamine based reward signal*

An animal receives reward from the environment based on its actions. In our model network reward is a global, scalar signal, r, which depends on the outcome of the output of the network in the current state of the environment (betterment of the current state leads to positive reward and vice versa). The reward plays a key role in training the network. The first weight stage (between input layer and the STN-GPe layer) is trained by reinforcement learning (see Ref. 6). The reward signal is transformed into the dopamine level, D, by the SNc and transmitted to the STN layer where it controls the average activity of STN neurons.

Therefore, the reward indirectly controls learning in first stage weights, as described next. Training in the second stage weights (between the STN-GPe system and the output layer) follows a form of winner take all mechanism.

2.1.4. *Learning rule for the first stage (input layer to STN-GPe layer) weights*

The first stage layer weights are updated by Hebbian learning as follows:

$$\Delta W^{\text{in}}_{ij;k} = \eta V_{ij} x_k \tag{8}$$

where η is the learning rate. High reward situations result in higher dopamine levels in STN layer which in turn increases average value of $V_{ij}, (\langle V_{ij} \rangle)$; conversely absence of reward or even negative reward is associated with lesser values of $\langle V_{ij} \rangle$. Thus reward, via dopamine levels, D, and $\langle V_{ij} \rangle$, indirectly controls the learning rate of first stage layer weights. This is similar to the manner in which reward controls learning rate in neural network models of reinforcement learning. The only difference in our case is that learning rate modulation is achieved via dopamine levels D.

2.1.5. *Output layer activation*

Each output node represents an action choice the network has to make. When an input is presented the action corresponding to the *winning* node is executed. The response, O_m, of the output neurons is calculated as,

$$O_m = e^{\left(-\|V - W^{\text{out}}_m\|^2/\sigma^2\right)} \tag{9}$$

where V is the activity of the STN layer, and W^{out}_m is the weight vector feeding into the mth output node (from the STN layer).

$$\Delta W^{\text{out}}_{m^*} = \alpha_+ (V - W^{\text{out}}_{m^*}), \quad \text{if reward } r > 0 \tag{10a}$$

$$\Delta W^{\text{out}}_{m^*} = \alpha_- (V - W^{\text{out}}_{m^*}), \quad \text{if reward } r < 0 \tag{10b}$$

where m^* is the index of the winning neuron, $\alpha_+ (> 0)$ and $\alpha_- (< 0)$ are reward-dependent learning rates, and r is the global reward signal.

The only difference between the proposed rule and the traditional competitive learning rule is that in traditional competitive learning, the output of the hidden layer is not used directly as input for the output layer in the learning rule, but the *desired value* of the output of the hidden layer is used while learning. However, since learning is unsupervised in our model, this desired value is not known beforehand, and the hidden layer output is used *as is*. Furthermore, it has been shown that only one action modality gets selected by the basal ganglia by disinhibiting a specific region in the GPi. Hence we have found it biologically salient to use the competitive learning rule for training the output layer weights.

In summary, the network described above is solving an input-output mapping problem with global reward information. The *exploratory dynamics*, required to learn solely from reward (without an explicit teacher), is provided by the oscillatory STN-GPe layer. The activity of the STN-GPe layer is sufficiently complex only when the lateral connections are negative. Reward, and hence dopamine level, controls the activity levels of STN layer. Since the first stage weights are trained by reinforcement learning, reward indirectly controls the learning rate of the first stage weights.

3. The Simulated Waterpool Experiment

The above network is used to drive exploratory behavior in a simulated version of the so-called *water pool* experiments. In real water pool experiments, a rat is made to explore a pool of water searching for a submerged platform, which is invisible since the water is muddy. The rat attempts to navigate towards the platform with the help of landmarks placed around the pool. On reaching the platform the rat receives an intrinsic reward (relief) or an external reward administered by the experimenter. The experiment is repeated by throwing the rat at various locations in the pool.

The setup used in our simulation is depicted in Fig. 3. The large circle represents the water pool. The small segment on the right of the pool is the submerged platform. Eight landmarks are placed around the periphery of the circle with uniform spacing. The landmarks are vertical poles with different height and are assumed to be uniquely identifiable by some property other than the height, such as, color. The model rat has an angle of vision of 180 degrees. The rat is also assumed to have a position (point size) and an orientation (heading) in the water pool at any instant.

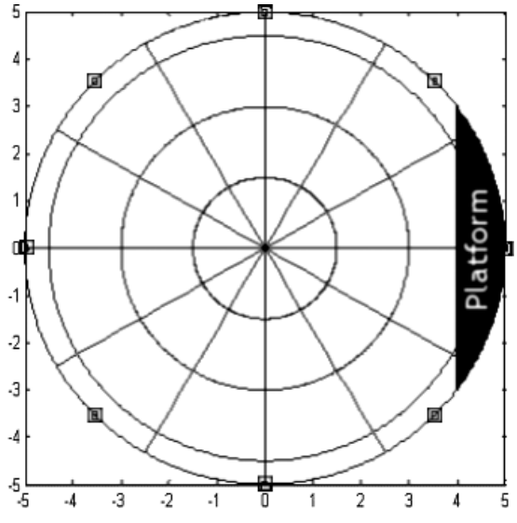


Fig. 3. The water pool experimental setup — the water-pool is in the form of a circle centered at the origin. The eight poles around the rim of the pool (at angular separations of 45°) are represented by squares. The line at x = 4 represents the edge of the platform. The platform is thus the minor segment of the circle bounded by this line. (The concentric circles (r = 30 cm, 60 cm and 90 cm) and radial lines (in steps of 30°) emanating from the center of the pool are merely shown as reference for a polar coordinate system).

From a given viewing point the rat can *see* a view containing a subset of the landmarks present around the pool. The view is encoded as a *view-matrix*, the construction of which is described below.

View-matrix construction

The virtual rat's visual field input is split into N_s sectors, which are represented by the columns of the view-matrix, I_v, of size $N_l \times N_s$, where N_l is the number of landmarks. The l^{th} element of the s^{th} column in I_v is nonzero if the l^{th} landmark is present in the s^{th} sector of the rats current view. The magnitude of that non-zero element is proportional to the size of the retinal image of that landmark. All other entries in the s^{th} column are zero since, the way the environment is arranged, typically only one landmark is present in a single sector of the view. This view-matrix, I_v, is resized into a column vector and presented as input, $\mathbf{x}$, to the network in Eq. (1).

Output representation

The view-vector is presented as an input to the 3-layer network of Fig. 2. Activity of the output layer represents the rat's motion in response to the current view. Each node in the output layer represents a displacement direction. The rat is displaced in the direction corresponding to the encoded direction of the *winning* node in the output layer. The rat moves a fixed distance of d per time step. The motion of the rat results in a change in its view and the cycle continues.

3.1. *Simulation results*

In this section we will briefly describe the experimental parameters and various outcomes of the computer simulation experiments based on the reinforcement learning framework discussed above.

Water-pool configuration: In the present computational experiment the radius of the tank is 100 cm with the platform in the shape of a minor segment of the circular tank bound by a chord, at a distance of 80 cm from the center of the circle (Fig. 3). Thus the ratio of the area of the platform to that of the tank is around 1:19.214, i.e., the platform area is only around 5% of the tank area. The rat is assumed to be able to swim a distance of 6 cm per time step and consequently would require only around 30 steps to reach the platform if perfectly trained and correctly oriented, even if it be placed at a point furthest from the platform (the end of the diameter perpendicular to the chord bounding the platform that does not lie within the platform).

Pole arrangement: There are 8 (= N_l) equally spaced poles of varying heights ranging between 100 to −100 (negative value indicates an inverted pole with respect to a reference level) placed all around the platform.

View Matrix: The rat has a field of vision of 180° (90° on either side) which is split into $N_s = 3$ sectors. Thus the view-matrix, I_v, of size 8×3, is constructed as described above. This matrix is transformed into a vector of 24 (= $N_l \times N_s$) elements and presented as input the network.

STN-GPe layer: A 10×10 grid of oscillatory neurons is chosen for the STN-GPe layer.

Output layer: The output layer consists of K (= 5) neurons which produce output by a *winner-take-all*

mechanism and map onto K distinct output states. The maximum movement deviation angle, $T_{max}/2$, is 30°, i.e., thirty degrees on either side of the current orientation.

Fluctuations in D: The dopamine level (D) providing the reward signal is assumed to be linearly proportional to the reward obtained with a maximum/minimum value for the fluctuations that is approximately 50% of the baseline value. Thus for a baseline value of, say 50, the dopamine level varies between 25 for reward, $r = -1$ and 75 for $r = +1$.

There are two phases in the network simulation, viz., training and testing.

3.1.1. *Training*

In the training phase the rat is set at random locations in the water pool at random orientations and is allowed to wander. Based on its current input vector and weight configurations, the rat wanders around the water pool, initially in an almost random fashion until it hits upon the platform by chance. When this occurs, a positive reward of +1 is provided to the rat based on the input and output of the *previous step*, so that the rat learns to select the appropriate output maneuver for each kind of visual input before entering the platform.

During its wandering in the water-pool the rat often comes into contact with the walls of the pool i.e., its trajectory often attempts to cross the pool's dimensional limits. In the simulation, at these instances, the rat is bounced off the wall and given a negative reward of −0.3, corresponding to the physical discomfort of dashing against the wall, so that it learns to actively avoid the walls. No weight update/reinforcement occurs during the wandering motion of the rat (except negative reinforcement at the walls) until the rat reaches the platform.

The rat, after several iterations learns to head directly to the platform with minimal wandering as shown, for example in Fig. 4, for extreme orientations of the rat towards the platform. A plot of the mean number of steps to platform vs. training time corresponding to one set of training trials can be found in Fig. 5a. It is clear from this figure that as training progresses, the rat learns to effectively navigate towards the platform in fewer steps on an average. The average number of bounces off the wall vs. training time is plotted for the same training sequence in Fig. 5b. Here the wall avoidance learning is apparent.

3.1.2. *Testing*

This phase involves testing the rat's movements without any update to the neural network weights, i.e., without providing any form of reinforcement, neither positive reward at the platform nor negative reward at the walls. This phase, though not biologically very realistic, is a standard practice in neural network literature to evaluate network performance.

While testing, the rat learns to efficiently avoid the walls, and once its view is within a reasonable range of the platform, heads in an almost straight line for the platform. In this phase we find interesting dynamics of the oscillatory STN-GPe layer as shown in Fig. 6. As the rat approaches the platform, the STN-GPe layer settles into a bistable state and each of the neurons enter either into a periodic alteration or a sustained maintenance of their respective outputs. Figures 6a and 6b depict both of these states of the network. On the contrary, when the rat is looking away from the platform, and exploring the other parts of the pool, the dynamics of the STN-GPe layer become chaotic, characteristic of wandering activity. The STN neurons settle into the rhythmic oscillations again only after the platform has been sighted. A few snapshots of the kind of STN-GPe layer dynamics during exploration are presented in Fig. 7.

In order to characterize the observed dynamics of the STN-GPe layer for each of the heading *toward* and heading *away* cases, two measures, viz., (1) *effective dimension* and (2) *average correlation coefficient* are computed.

Effective dimension is a measure of the effective number the degrees of freedom of the activity, $\mathbf{v}(t)$, of a system. Let λ_k and λ_{max} be the kth and the highest eigenvalues of the autocorrelation matrix of the activity, $\mathbf{v}(t)$, over a duration of interest, such that $\lambda_{max}/\lambda_{max} = \frac{1}{2}$. Then k is the effective dimension.

Average correlation coefficient is the average value of correlation between pairs of components of $\mathbf{v}(t)$, say, $v_i(t)$ and $v_j(t)$. The averaging is performed over a large number of randomly chosen pairs of components.

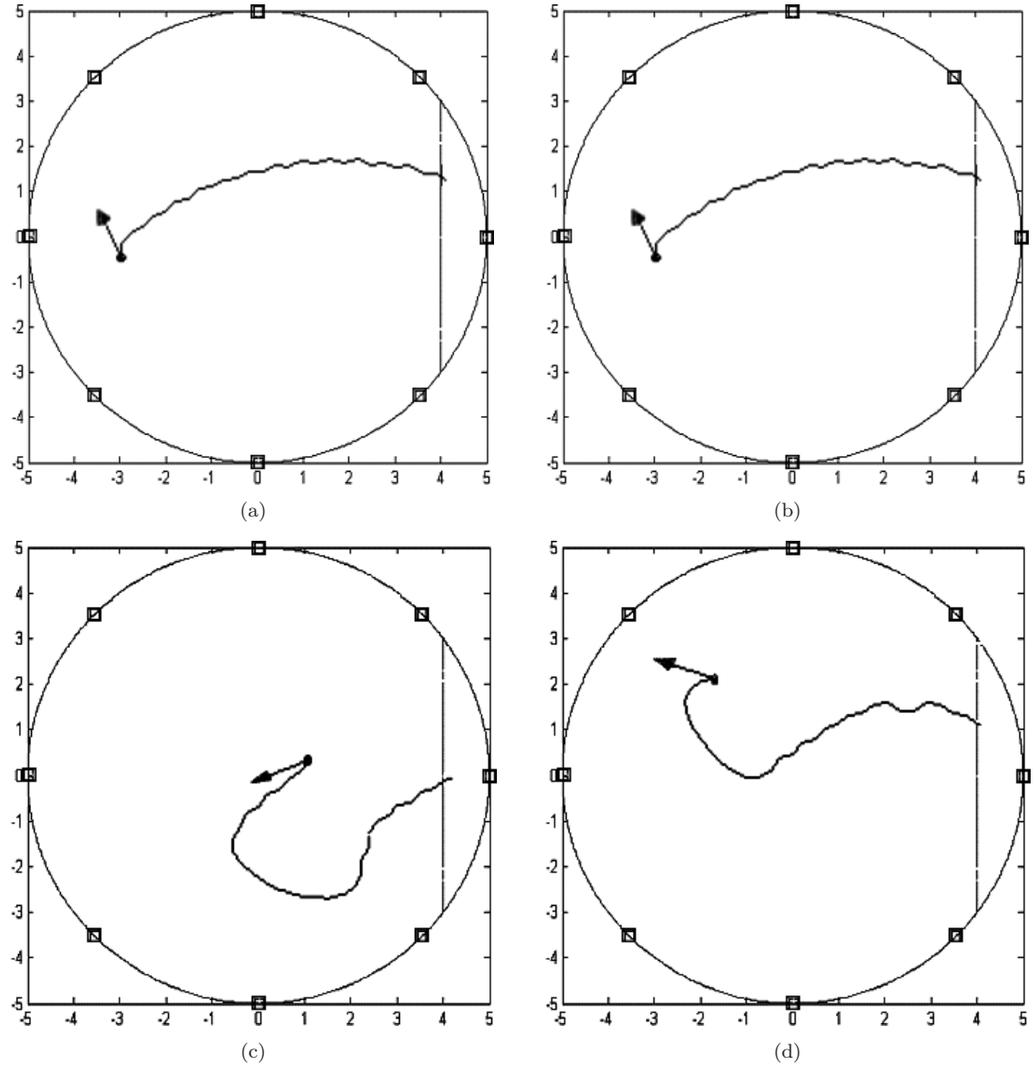


Fig. 4. Various trajectories taken by the rat for different initial locations (depicted as dark circles) and initial orientations (depicted by a short arrow emanating from initial location). Cases (a) and (b) show navigation to the platform when the platform is initially somewhat within visual range. Cases (c) and (d) indicate more difficult navigations when the platform is far removed from initial visual range.

The values of these parameters are indicated in Figs. 6 and 7 for the various dynamic regimes of the network characterizing the behavior of the rat.

4. Pathologies

Pathologies arising out of disruption of the dopamine reward signal due to lesions to the mesencephalic dopaminergic centers or the substantia nigra pars compacta (SNc) are discussed in the present section. This is simulated in the computer experiment by fixing a limit on the maximum level of dopamine available to the network which results in a spurious decrease in the magnitude of the reward signal that is fed into the network.

The network was simulated with a maximum D value of 10, and a minimum value of 0. The resulting behavior of the rat, viz. its motion is depicted in Fig. 9. It is found that the rat moves predominantly in circles and exhibits a severe turning bias. Consequently the rat fails to explore the tank properly and therefore reinforcement learning

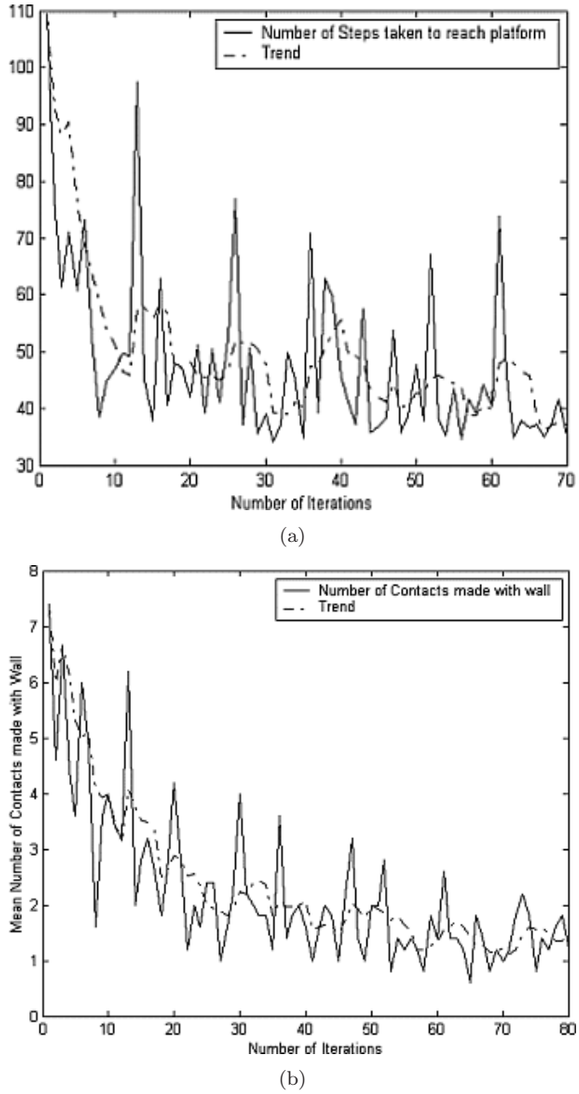


Fig. 5. (a) A plot of mean number of steps to reach the platform versus training cycle number. There is a gradual reduction in the mean number of steps to reach the platform as training progresses (averaged over several training cycles); (b) A plot of mean number of bounces off the wall before reaching the platform versus training cycle number. There is a marked reduction in the mean number of bounces off the walls as training progresses. The trends (dotted lines) are running means of the solid line values.

(which has its basis in exploration) is impaired. The history of the number of steps taken to reach the platform is depicted in Fig. 10. It can be seen that unlike the case with 50% baseline D, in the present case, the number of steps

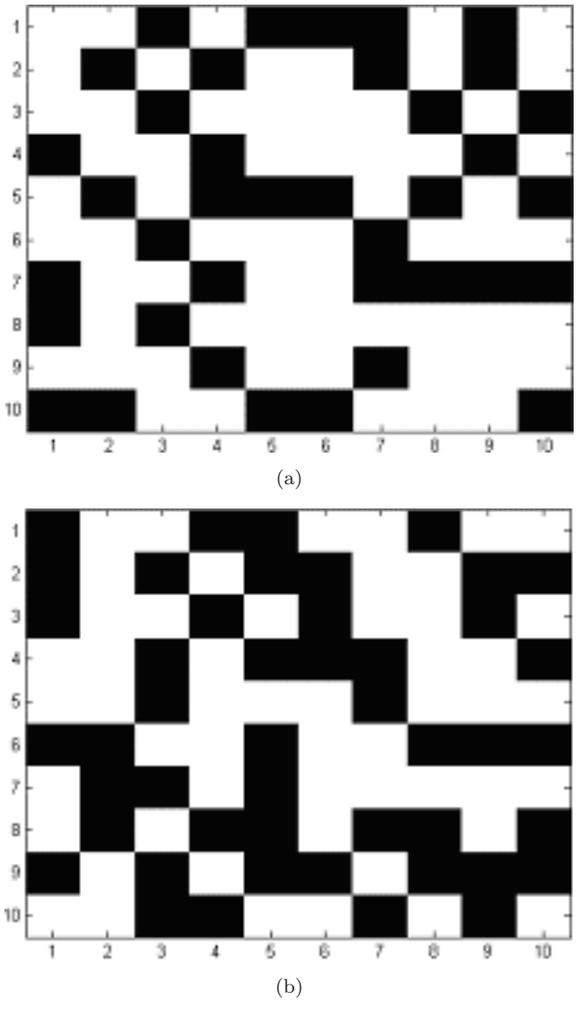
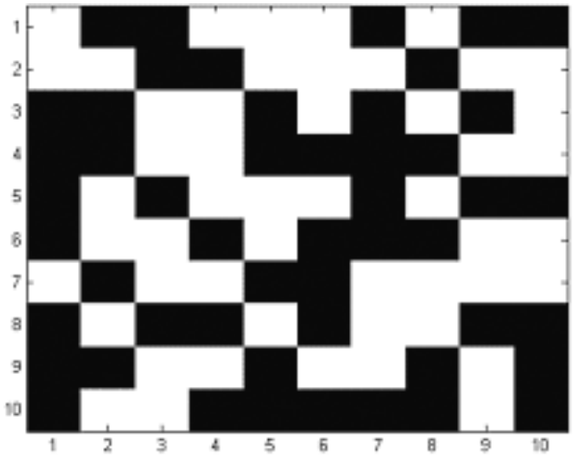


Fig. 6. The two-states characterizing the oscillatory dynamics of the hidden layer when the rat is heading toward the platform. The network periodically shuttles between these two states during the rat's platform directed motion. The effective dimension and average correlation coefficient for these states were evaluated and found to be 2 and -0.4432 respectively, which indicates that familiarity of the visual input is responsible for reducing chaos in the oscillatory layer and curtailing the exploratory dynamics.

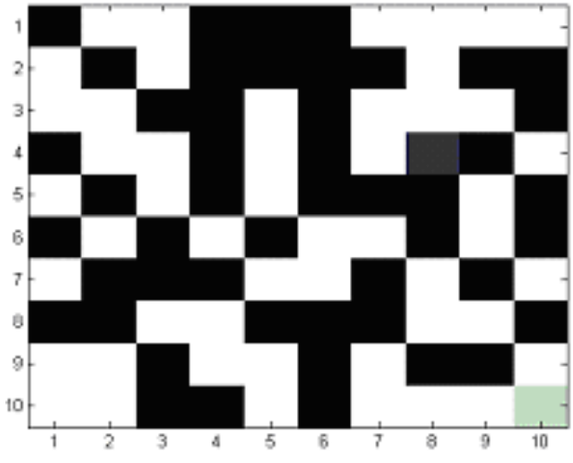
taken by the rat to reach the platform is erratic, and does not decrease with further training cycles indicating insufficient, impaired learning.

5. Discussion

In the present discussion we focus on biological constraints under which we have developed our model.



(a)

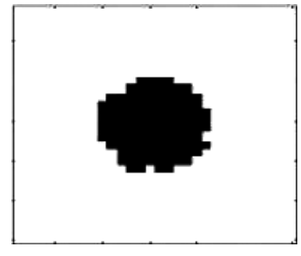


(b)

Fig. 7. Some sample hidden layer network states when the rat is looking away from the platform. The effective dimension and average correlation coefficient for these states were evaluated and found to be 6 and 0.3743 respectively, which indicates that novelty in the visual input causes chaotic dynamics in the oscillatory layer that encourages exploration.

(a) Chaos

(b) Traveling waves



(c) Compact center

Fig. 8. Three characteristic patterns of activity of the STN-GPe layer — (a) chaotic, (b) traveling waves and (c) compact center. The three activity regimes (from top to bottom) are obtained by progressively increasing ϵ (in Eq. (7)) from 0 to 2. Increasing ϵ increases the percentage of positive lateral connections in STN. In regime (c), *compact center*, the array splits into a center and a surround, with neurons in either region forming a synchronized cluster. Similar activity patterns have been observed by Terman *et al.* in their model of basal ganglia.[10]

Thus, we present the evidence from literature for the involvement of basal ganglia in exploratory behavior as well as the neurophysiological mechanisms for D-level (dopamine level) learning by the STN-GPe loop. Thence we present evidence for similar neural dynamics observed in the olfactory bulb for familiar and unfamiliar odors and correlate this observation with our oscillatory network behavior under novel vs familiar circumstances. We also attempt to justify the competitive learning rule in the output layer based on the *loser take all* learning rule of Berns and Sejnowski (see Ref. 13). Scope for further work in modeling basal ganglia pathologies is also discussed.

5.1. *Evidence from literature for basal ganglia role in exploration*

The basal ganglia have traditionally been characterized chiefly by two pathways, the direct and

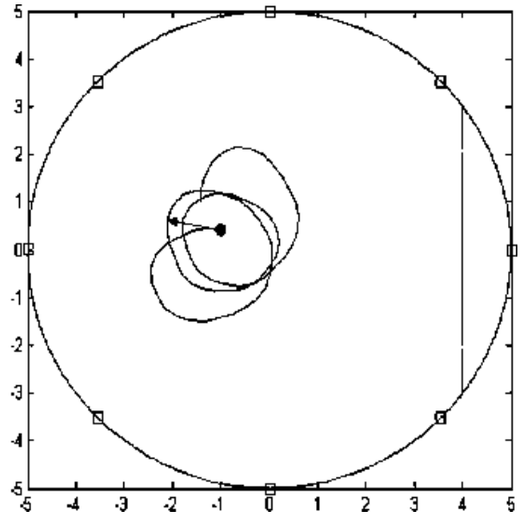


Fig. 9. Trajectory of the rat after disruption of dopamine reward signal due to lesions in the mesencephalic dopaminergic system or the SNc nucleus. The rat moves in circles with a severe turning bias leading to loss of exploration, thereby causing impairment in learning.

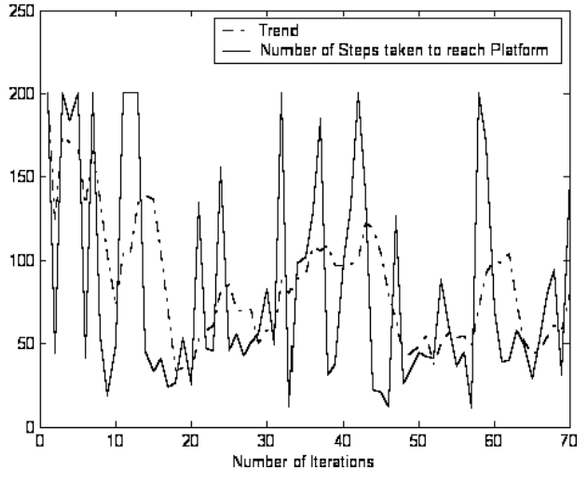


Fig. 10. A plot of mean number of steps to reach the platform versus training cycle number in the case of the pathology (of disrupted dopamine reward signal). It can be seen that there is an erratic variation and no observable reduction in the mean number of steps to reach the platform as training progresses due to impairment in learning.

indirect pathways.[1] The direct pathway is involved in the learning of potentially rewarding behaviors (by LTP of the cortico-striatal NMDA glutamate synapses via expression of the c-fos gene), selecting (providing the GO signal for) such behaviors from among many competing actions. The indirect pathway is involved in learning to abstain from potentially unrewarding responses (by LTD of the corresponding cortico-striatal synapses), thereby providing the STOP signal for behaviors eliciting no reward.[14] However as pointed out by Gillies *et al.*[9] this simple description is fast becoming obsolete in the light of new findings about the striatal pathways. In their opening note, Gillies *et al.*[9] observe that "... contemporary neuroanatomy of the basal ganglia reveals a prominent feedback system, involving the excitatory subthalamic nucleus and the inhibitory globus pallidus..."

That the cortex is not indispensable for exploratory behavior, is apparent from observations in literature. For instance the study by Grillner *et al.*[15] clearly obviates the need for the cortex in planned, goal-directed behaviors. It has been observed that even in advanced mammals like the cat, goal-directed locomotion is retained after an of the entire cerebral cortex that leaves the rest of the forebrain intact (including the basal ganglia and hypothalamus). Decorticated kittens have been found to exhibit periods of rest alternating with patterns of activity including searching for food and remembering location of food — all of which are examples of exploratory behavior. Grillner *et al.*[15] conclude that "...The neuronal substrate contained in a forebrain devoid of the cerebral cortex is thus able to produce surprisingly complex, goal-directed patterns of behavior[15] ...", thus strengthening our case for the involvement of the basal ganglia in exploratory behavior leading to reinforcement learning.

5.2. *Mechanism for the learning driven by the level of dopamine in the STN-GPe loop*

The D-level learning by the STN-GPe loop was achieved in the network model by a simplified Hopfield network model. We hypothesize the neural correlate of this network to be pallido-nigro-striatal circuit involving projections from the GPe to the SNc and back to the striatum. The detailed mechanics of this feedback learning are as follows: The STN-GPe oscillator attempts to learn the level of Dopamine signal which is fed in via

the SNc-Striosomal projections through the striatal Matrix into the GPe. The error signified by e (Eq. (4b)) is back propagated through the GPe-SNc (pallidal-nigral) projection (of which there is little documented literature), and it is hypothesized that neurons in the SNc compute the difference between the actual percentage of GPe units that are active (D_a) and the percentage that are required to be active (as given by their own activity reflecting the amount of dopamine, D), thereby feeding the error signal back into the striatum and further to the GPe-STN loop where it may modulate the percentage of active neurons according to Eqs. (2) and (3). This idea of error-back propagation is similar to that proposed by Berns and Sejnowski[3] wherein the error signal, e, computed by hypothetical projections from the striatum and the GPi to the SNc/VTA, modulates the STN-GPe synaptic weights.

5.3. *Discriminating novel from familiar odors: The case of olfactory bulb*

It has been observed using grid electrode recordings, that neurons in rabbit olfactory bulb exhibit characteristic responses to novel and familiar odors. When a familiar odor is presented to the animal, the olfactory bulb responds with a rhythmic waveform; however, when the stimulus is novel or unfamiliar, activity in the bulb exhibited chaotic wandering.[16] This is analogous to the STN-GPe layer in our model. When the rat is looking away from the platform and is searching for it, the STN-GPe layer exhibits a desynchronized pattern of activity, whereas, when the rat is heading straight towards the platform, the activity of the STN-GPe layer switches periodically between only two states (Figs. 6 and 7).

5.4. *Motor selection vis-a-vis competition at output layer*

The model of the basal ganglia proposed by Berns and Sejnowski (see Ref. 13) is based on the premise of 'action selection' and emphasizes possible timing differences between the direct and indirect pathways in a model that included only feed-forward intrinsic basal ganglia connections. An interesting feature of this model is that it incorporates a version of the dopamine hypothesis for reinforcement learning as a means for adaptively tuning the selection mechanism.[11] At the output layer, the GPi (or EP), the final action selection occurs through disinhibition of the thalamus (via inhibition of the GPi), and this is presumed to occur by a *loser take all* learning mechanism wherein the most inhibited GPi region corresponds to the action modality that finality gets selected. We therefore justify our use of the competitive learning rule for the output layer based on these observations in literature.

5.5. *Modeling pathologies*

In the present work we have studied only the pathologies arising out of the disruption of the dopamine reward signal in the basal ganglia. The loss of dopamine signal, in fact, corresponds to the condition of Parkinsons disease where there is a drop in baseline dopamine levels due to the degeneration of the SNc (Substantia Nigra pars compacta) nucleus that is involved in dopamine production. The pathologic rat in our model exhibited a sustained pattern in network output, due to a drop in baseline dopamine levels from 50% to 5%, and given that the only output of the network is a *turning angle*, this led to a severe turning bias. It may hence be fitting to consider this inability to initiate novel movements analogous to the symptoms of akinesia and bradykinesia exhibited by Parkinsonian patients. It would be interesting to observe the behavior exhibited by the rat under pathological conditions of disruption of the interconnectivity patterns within individual layers, and is a direction for future research.

In fact, it has been demonstrated by Terman *et al.*[10] that this weakened intra-pallidal inhibition within the indirect pathway can switch the dynamics of the circuit from chaotic (or irregular) to periodic (or rhythmic) in the form of traveling waves. Terman *et al.*[10] conclude that this kind of observation may be sufficient to explain the emergence of correlated oscillatory activity in the subthalamopallidal circuit after destruction of dopaminergic neurons in Parkinsons disease and in animal models of Parkinsonism. Therefore, it appears that while complex non-periodic activity is a sign of normal STN, coherence and loss of complexity marks pathological function.

In a completely different context, working with a model of motor-unit recruitment in skeletal muscle, Chakravarthy *et al.*[8] arrive at analogous conclusions. The model predicts that complex, non-periodic

activity in the a-motor units of spinal cord are a prerequisite for a skeletal muscle to reliably follow a motor command originating from higher centers. When the motor unit activity became rhythmic (due to disruption in lateral connectivity), the muscle failed to follow the motor command. Interestingly, it is known from electromyographic (EMG) recordings from individual muscle fibers, that in normal muscle activity of individual fibers is desynchronized[17]; whereas in diseased muscle (e.g., polio) the activity is highly correlated.[18]

6. Conclusion

We have presented a simplified computational model of the basal ganglia and have demonstrated its role in exploratory reinforcement learning by simulating the hidden layer oscillatory dynamics with an interconnected network of neurons representing the STN-GPe segment. We have shown that such a network is capable of learning well under external reinforcement cues by studying the behavior of a rat learning the location of a submerged platform in a circular tank of water when dropped at random locations and at random orientations in the tank and left to fend for itself.

While it is the hippocampus which has received the greatest attention as the key player in spatial exploratory behavior (e.g., maze learning in rats), we propose that the validation of our model would require similar data from basal ganglia. In the meantime we can only point to available observed data from the human STN-GPe loop,[5] and computational models constructed on the basis of these experiments.[10]

In furtherance of our of hypothesis regarding the putative role of the basal ganglia in exploration based reinforcement learning, we direct the reader's attention to a recent article by Kao *et al.*[19] These authors have shown that a basal ganglia-forebrain circuit in song-birds, the anterior forebrain pathway (AFP), contributes to motor learning by biasing motor outputs towards desired targets or by introducing stochastic variabilityn required for reinforcement learning.[19]

Acknowledgments

We thank G. Gangadhar for the fruitful discussions on the analysis of the oscillator.

Appendix A Analysis of the Oscillator

The system of equations for single oscillator are given as,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -x + v - s + I \tag{A.1}$$

$$v = \tanh(\lambda x) \tag{A.2}$$

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -s + v \tag{A.3}$$

Differentiating (A.1),

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\frac{\mathrm{d}x}{\mathrm{d}t} + \lambda\mathrm{sech}^2(\lambda x)\frac{\mathrm{d}x}{\mathrm{d}t} - \frac{\mathrm{d}s}{\mathrm{d}t} \tag{A.4}$$

Substituting (A.2) and (A.3) in (A.4),

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\frac{\mathrm{d}x}{\mathrm{d}t} + \lambda\mathrm{sech}^2(\lambda x)\frac{\mathrm{d}x}{\mathrm{d}t} - (-s + \tanh(\lambda x)) \tag{A.5}$$

Substituting (A.1) and (A.2) in (A.5),

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\frac{\mathrm{d}x}{\mathrm{d}t} + \lambda\mathrm{sech}^2(\lambda x)\frac{\mathrm{d}x}{\mathrm{d}t} - \left(\frac{\mathrm{d}x}{\mathrm{d}t} + x - v - I + \tanh(\lambda x)\right) \tag{A.6}$$

On rearranging,

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \frac{\mathrm{d}x}{\mathrm{d}t}(2 - \lambda\mathrm{sech}^2(\lambda x)) + (x - I) = 0 \tag{A.7}$$

Which is similar to Lienard's equation,

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \frac{\mathrm{d}x}{\mathrm{d}t}f(x) + g(x) = 0 \tag{A.8}$$

where $f(x) = 2 - \lambda\mathrm{sech}^2(\lambda x)$, and $g(x) = x - I$

Checking for the Lienards conditions (assume $I = 0$):

1. Both $f(x)$ and $g(x)$ are continuous and differentiable $\forall x \in \mathbb{R}$;
2. $g(-x) = -g(x) \quad \forall x \in \mathbb{R}$ (i.e., $g(x)$ is an *odd* function);
3. $f(-x) = f(x) \quad \forall x \in \mathbb{R}$ (i.e., $f(x)$ is an *even* function);
4. $g(x) > 0 \quad \forall x > 0, x \in \mathbb{R}$
5. The odd function $F(x) = \int_0^x f(u)\mathrm{d}u = 2x - \tanh(\lambda x)$ has exactly one positive zero at $x = x_0$, is negative for $0 < x < x_0$, is positive and non-decreasing for $x > x_0$, and $F(x) \to \infty$ as $x \to \infty$ (one can estimate x_0 from the graph of $F(x)$.

The system governed by Eq. (A.6) satisfies Lienard's conditions. Hence the system has a unique stable *limit cycle* surrounding the origin in the phase plane.

References

1. J. A. Obeso, M. C. Rodriguez-Oroz, M. Rodriguez, J. Arbizu and J. M. Gimenez-Amaya, The Basal Ganglia and disorders of movement: Pathophysiological mechanisms, *News Physiol. Sci.* **17** (2002) 51–55.
2. P. Redgrave, T. J. Prescott and K. Gurney, The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience* **89**(4) (1999) 1009–1023.
3. G. S. Berns and T. J. Sejnowski, A computational model of how the basal ganglia produce sequences, *J. Cogn. Neurosci.* **10**(1) (1998) 108–121.
4. J. C. Houk, J. L. Davis and D. G. Beiser (eds.), *Models of Information Processing in the Basal Ganglia* (MIT Press, Cambridge, MA, 1995)
5. M. D. Bevan, P. J. Magill, D. Terman, J. P. Bolam and C. J. Wilson, Move to the rhythm: Oscillations in the subthalamic nucleus-external globus pallidus network, *Trends Neurosci.* **25**(10) (2002) 525–531.
6. A. G. Barto, Reinforcement Learning, in *The Handbook of Brain Theory and Neural Networks*, 1st edn. (MIT Press, Cambridge, MA, 1999).
7. B. Chirikov, A universal instability of many-dimensional oscillator systems, *Phys. Rev.* **52** (1979) 263–379.
8. V. S. Chakravarthy, S. T. Thomas and N. Nair, A model for scheduling motor unit recruitment in skeletal muscle, in *Intl. Conf. on Theoretical Neurobiology, National Brain Research Center, Gurgaon* (Feb. 24–26, 2003).
9. A. Gillies, D. Willshaw and Z. Li, Subthalamic-pallidal interactions are critical in determining normal and abnormal functioning of the basal ganglia, *Proc. Biol. Sci.* **269**(1491) (2002) 545–551.
10. D. Terman, J. E. Rubin, A. C. Yew and C. J. Wilson, Activity patterns in a model for the subthalamopallidal network of the basal ganglia, *J. Neurosci.* **22**(7) (2002) 2963–2976.
11. T. J. Prescott, K. Gurney and P. Redgrave, Basal Ganglia in *The Handbook of Brain Theory and Neural Networks*, 2nd edn. (MIT Press, Cambridge, MA, 2002).
12. R. Montague, P. Dayan and T. J. Sejnowski, A Framework for mesencephalic Dopamine systems based on predictive Hebbian learning, *J. Neurosci.* **16**(5) (1996) 1936–1947.
13. G. S. Berns and T. J. Sejnowski, How the basal ganglia make decisions, in *Neurobiology of Decision Making*, (Springer-Verlag, Berlin, 1996).
14. J. W. Brown, D. Bullock and S. Grossberg, How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades, *Neural Netw.* **17**(4) (2004) 471–510.
15. S. Grillner, A. P. Gerogopoulos and L. M. Jordan, Selection and initiation of motor behaviour, in *Neurons, Networks and Motor Behaviour* (MIT Press, Cambridge, MA, 1997).
16. C. A. Skarda and W. J. Freeman, How brain makes chaos in order to make sense of the world, *Behavioral and Brain Sciences* **10** (1987) 161–195
17. J. A. R. Lenman and A. E. Ritchie, *Clinical Electromyography* (Pitman Medical Publishing Co., Kent, 1978).
18. F. Buchthal and S. Clemmesen, The electromyogram of atrophic muscles in cases of intermedullary affections, *Acta Psychiatry (Kbh)* **18** (1943) 377–387.
19. M. H. Kao, A. J. Doupe and M. S. Brainard, Contributions of an avian basal ganglia-forebrain circuit to real-time modulation of song, *Nature* **433**(7026) (2005) 638-4.