Reinforcement Learning and Dimensionality Reduction: a model in Computational Neuroscience

Nishal Shah, Frédéric Alexandre

Reinforcement Learning and Dimensionality Reduction: a model in Computational Neuroscience

Nishal Shah, and Frédéric Alexandre

***Abstract*—Basal Ganglia, a group of sub-cortical neuronal nuclei in the brain, are commonly described as the neuronal substratum to Reinforcement Learning. Since the seminal work by Schultz [1], a huge amount of work has been done to deepen that analogy, from functional and anatomic points of view. Nevertheless, a noteworthy architectural hint has been hardly explored: the outstanding reduction of dimensionality from the input to the output of the basal ganglia. Bar-Gad et al. [2] have suggested that this transformation could correspond to a Principal Component Analysis but did not explore the full functional consequences of this hypothesis. In this paper, we propose to study this mechanism within a model more realistic from a computational neuroscience point of view. Particularly, we show its feasibility when the loop is closed, in the framework of Action Selection.**

I. Introduction

THE goal of computational neuroscience is to study, by the means of models, the link between structure and function in the nervous system. To that end, progresses in the better understanding of information flows in the brain and in the mastering of neuronal computing properties have to be linked. Such an approach is considered here in the case of Reinforcement Learning.

A. Overview

Computational Neuroscience has studied a lot cortical properties to establish learning principles at the macroscopic scale (e.g. [3]). In summary, the part of the cortex posterior to its central sulcus represents its sensory pole and is characterized by its self-organizing properties. For example, Self-Organizing Maps as proposed by Kohonen [4] are able to build, in an unsupervised learning process, topological maps displaying sensory information in a way similar to cortical representation in the sensory pole. Statistic methods in Machine Learning like the K-means have also been related to this kind of adaptive processing.

The part of the cortex anterior to its central sulcus (also called the frontal cortex) represents its motor pole and is studied to model motor activities, for example in autonomous robotics [5] and more generally for the temporal organization of behavior. Lastly, many sensorimotor tasks have been modeled through the association of both poles (see e.g. [6] for the visuomotor case).

Manuscript received February 3, 2011.

F. Alexandre is with the French National Institute in Computer Science and Control (INRIA), Centre of Research INRIA of Nancy, 615 rue du Jardin Botanique, CS 20101, F-54603 Villers-les-Nancy (corresponding author e-mail: frederic.alexandre@inria.fr).

N. Shah, was doing a training period at the Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), associated to the CNRS and the universities of Nancy, France.

A special attention has been given to the cortex in modeling activities certainly because it is one of the largest neuronal structures in the brain but also because Neuropsychology describes it as the centre for the most advanced cognitive functions. As far as Reinforcement Learning [7] and Action Selection are concerned (in short, Action Selection is the task of selecting the action maximizing the expectation of reward, given the current perception and the knowledge of the consequences of the actions on the outer world), they are certainly among the most advanced cognitive functions and the cortex could be thought as having all the information to tackle the tasks, considering also that the posterior and the frontal cortex contain specific areas for the interoceptive representation of the body and hence of rewards [8]. Nevertheless, the cortex is also characterized by its mainly local connectivity (each cortical neuron is only connected to 10^3-10^4 cortical neurons, among the 10^9 potential targets), which makes a global competition before decision very difficult inside that structure. Moreover, cortical learning mainly corresponds to stable sensorimotor learning [9], very different from the very dynamic and changing nature of representations in reinforcement learning [7], though some regions of the frontal cortex are also described with very dynamic and volatile representations related to planning of actions [10]. Peter Redgrave and his colleagues propose to solve that dilemma [11], postulating that the Basal Ganglia (BG), a set of sub-cortical interconnected nuclei, build in a loop that they constitute with the cortex and the thalamus, the physiological substratum associated with the cortex for action selection tasks, particularly performing reinforcement learning.

B. Basal Ganglia

Basal Ganglia are described in [11] as an "adaptive switch" performing action selection motivated by the evaluation of the predicted reward, through two loops they belong to. These loops allow for a direct analogy with the Actor-Critic architecture [12], one of the fundamental algorithms in reinforcement learning, where the Actor selects the best action from the current perceptions and acquired knowledge and the Critic predicts the expected reward from the same elements. Errors of prediction are exploited to update both agents [7].

The basal loop (Cortex-BG-Thalamus-Cortex) stands for the Actor. This main loop receives information from almost all regions (posterior and frontal) of the cortex, in the input layer of the BG: the Striatum, a large neuronal structure containing in primates up to 10^7 neurons. The Sub-Thalamic Nucleus (STN) is another (smaller) input layer of the BG but

will not be considered here, for the sake of simplicity. The output layer of the BG is composed of two structures GPi/SNr, that we will not differentiate here for the same reason. At rest, this inhibitory output layer has a tonic activity on its targets: nuclei of the Thalamus that project onto the frontal cortex. Thus, the motor pole of the cortex is, by default, inhibited and only a selective inhibition in the output structure of the BG will accordingly disinhibit the thalamus, allowing for the triggering of the corresponding action in the motor cortex. Particularly, the output structure of the BG (GPi/SNr) can be inhibited by its inhibitory input structure (the Striatum), through their direct connectivity in the main basal loop. This selection of action is made from current sensorimotor information brought by the cortex and from the prediction of reward brought by the other loop of the BG, standing for the critic.

The striato-nigral loop in the BG [13] reciprocally links the Striatum and the Substantia Nigra pars compacta (SNc) and stands for the Critic. SNc is one of the few cerebral structures containing dopaminergic neurons (the dopamine is a modulatory neurotransmitter, the action of which is related to reinforcement effects). In a schematic way, it can be said that SNc receives from the Striatum (and other cerebral structures) information that allows it to relate the sensorimotor situation to the level of reward. On that basis, it can predict the reward to come and, when the prediction fails, it can deliver dopamine to modulate the activity in the striatum, thus modulating the actor. From the seminal work by Schultz [1], it has been proposed that dopamine encode the error of prediction of reward, thus relating this mechanism to the Temporal Difference algorithm [14].

This functional sketch underlines the analogy between the two loops constituting the BG and the Actor-Critic architecture for Reinforcement Learning. Many researches have been carried out to make that analogy more precise or to modify it. Concerning the basal loop, the main question is about the criteria for action selection, allowing to selectively disinhibit one output unit from input data. Beyond the direct link between the input layer (the Striatum) and the output layer (GPi/SNr), other interconnected nuclei belonging to the BG (like STN mentioned above, or GPe) make possible other pathways, like an indirect [15] and a hyperdirect [16] pathway. How interactions between those pathways can lead to a more efficient and realistic selection of action is an open question today. Another important question is about the representation of information along the basal loop. On the one hand, information is described as segregated in territories specific to the different levels of action selection (strategy, planning and execution) and the corresponding abilities (motivation, working memory and action) [10] and displayed in a topological way in channels conserved along the loop [17]. On the other hand, the very small size of the output layer is underlined (10^5 neurons in primates: ten thousands time smaller than the cortical input!) and this funneling effect leads to conclude that a strong reduction of dimensionality takes place from the input to the output layer of the BG [2].

Concerning the striato-nigral loop (the critic), ongoing researches mainly look for a better understanding of the temporal behavior of the loop [18] and its link to respondent conditioning [19]. In this paper, we will concentrate on the main basal loop (the actor) and its supposed mechanism of dimensionality reduction.

II. Dimensionality Reduction

A. In Artificial Neural Networks

More generally, reduction of information is a filtering mechanism, well-known in the domain of automatic processing of information. It can be obtained by reducing the number of data, for example by a clustering mechanism, summarizing a set of data by a representative prototype [4] or by reducing the dimensionality of data, as it is the case with Principal Component Analysis (PCA).

Both mechanisms have been implemented with artificial neural networks. Concerning PCA [20], it is known for a long time that the hebbian rule (Eq. 1), applied to weight modification between an input layer X of dimension m and a unique output neuron y (Eq. 2), will extract in the weight vector W a direction aligned to the first principal component of the input space.

$$\Delta W_{ij} = \alpha(x_i y_j) \tag{1}$$

$$\Delta W = \alpha X y \tag{2}$$

where α is a small positive real, the incrementation step.

Nevertheless, this learning rule is also known for being divergent, which makes difficult the extraction of this direction. A classical way to prevent the rule from diverging is to normalize it, for example by dividing by the norm of the weight vector. But in this case, the calculus is no more local, which can be annoying in a neuromimetic framework. That is why E. Oja has proposed to linearize the normalization, approximating it by the first term of the corresponding Taylor expansion [21], which has also the advantage of making the calculus local (Eq. 3).

$$\Delta W = \alpha(Xy - Wy^2) \tag{3}$$

This learning rule is stable and converges (if α is chosen sufficiently small) toward a weight vector corresponding to the direction of the first principal component of the input space, for a unique output neuron. Subsequent studies have shown the possibility to extract several principal components, by displaying several neurons in the output layer Y of dimension n, endowed with an inhibitory lateral connectivity, a weight matrix A. Output neurons in Y are linearly evaluated as the weighted sum of forward and lateral activities (Eq. 4).

$$y_i = \sum_{j=1}^{m} W_{ij} x_j + \sum_{j=1}^{n} A_{ij} y_j \tag{4}$$

These studies share the principle of using an anti-hebbian rule between the output neurons [22][23], decorrelating the activations of the output units (Eq. 5).

$$\Delta A_{ij} = -\alpha y_i y_j, i \neq j \quad (5)$$

The principal components can be extracted successively, by incrementally adding neurons in the output layer or by defining the A matrix as a lower triangular matrix with a null diagonal, laying down consequently a hierarchical relation between the output neurons [22][24][25]. Foldiak has also shown [26] that using from the beginning the full output layer with a full lateral weight matrix engenders the principal sub-space with the corresponding dimension (but does not yield individual principal component directions). Let us lastly mention that these networks are generally made of linear neurons, in order to reproduce PCA, which is a linear operation. Nevertheless, some models explore non-linear versions of neuronal functioning rules [22] in order to implement some kinds of non-linear PCA related to higher order statistics [27].

B. In the Basal Ganglia

Surprisingly enough, the funneling effect in the BG (namely, the strong reduction from the cortex to the Striatum and from the Striatum to GPi/SNr) has been hardly exploited in modeling activities. Bar-Gad and his colleagues [2] are among the only ones that have proposed that a kind of PCA could be the principle of transformation of information between these layers. One of their strong arguments is that classical models of selection of action require a strong lateral competition between neurons, along the direct basal pathway (Cortex-Striatum-GPi/SNr), whereas electrophysiological observations [28] report very weak lateral weights in the basal part of this pathway. Yet, if a PCA-like processing is postulated in the pathway, its evolution will tend to decorrelate neuronal activities and to decrease the lateral (inhibitory) weights down to zero.

The RDDR model (Reinforcement Driven Dimensionality Reduction) proposed in [2] is a model of the direct basal pathway operating a PCA. It is directly inspired from the APEX model presented in [24], including forward weights updated by the Oja rule and a hierarchy of neurons in the output layer, with a lower triangular matrix of lateral weights, learned by an anti-hebbian rule also adapted from the Oja rule.

The main originality of the RDDR model is to propose that the learning rule associated to the forward weights could be modulated by the reinforcement associated to the current situation. This is a simple but efficient view of the modulatory role of the dopaminergic pathway carried by the striato-nigral loop, onto the main basal loop. Accordingly, the forward weights are updated as in Eq. 6.

$$\Delta W_{ij} = \alpha r(x_i.y_j - W_{ij}.y_i^2) \quad (6)$$

where r is the reinforcement associated to the current example X. The lateral weights are updated as in Eq. 7.

$$\Delta A_{ij} = -\alpha(y_i y_j + A_{ij} y_i^2), i > j, A_{ii} = 0 \quad (7)$$

The RDDR model has been evaluated mainly for its ability to perform a PCA, conditionally to the level of reinforcement. In the experiments [2], simple and artificial stimuli are built, corresponding to 8x8 matrices where only one line and/or one column are set to 1, the rest of the matrix being set to zero. The goal of the network is to learn to build a reduced representation of the input matrix, according to the delivery of a reward. In a first stage, the reward will be associated to the presence of a line in the matrix; in a subsequent stage, it will be associated to the presence of a column. An output with 16 neurons is sufficient to tackle both cases; if only one case is considered, 8 neurons are sufficient.

Several observations are drawn in [2] about the behavior of the PCA mechanism modulated by reinforcement. First, interestingly enough, it is shown that during convergence, lateral inhibitory weights converge up to zero. Simultaneously, the correlations between output units become null. This is very consistent with the observations by Jaeger [28] mentioned above. When the rewarding rule changes, these values will suddenly increase and will go back to zero after a new period of learning, as a new representation is learned. Secondly, to better evaluate information representation and considering that the units in the model are linear, the authors propose to project back the output toward an artificial layer, with the same size as the input and with the inverse matrix of weights. This operation allows to artificially reconstruct the original information and to check that it was conserved.

To sum up, the RDDR model has been mainly built and evaluated for its ability to implement an original mechanism: a PCA transformation modulated by a reinforcement signal. Our purpose is to see if this original mechanism is still valid in a more realistic framework, from a computational neuroscience point of view. More precisely, this has been done by:

A. Using Dynamic Neural Fields with non-linearity and leak, instead of simple linear neurons
B. Adding a sensorimotor cortical axis, allowing to preactivate eligible actions
C. Closing the basal loop, with a feed-back toward the motor cortex
D. Defining a more ecological learning protocol
E. Sending the reward as a result of action
F. Adding an exploration mechanism

These extensions are described in the next section.

III. ADAPTING RDDR TO A BIO-INSPIRED FRAMEWORK

Our goal is to incorporate the mechanism proposed in [2] in a network consistent with the main loops of the cerebral system and to feed it with more ecological stimuli. Accordingly, we have extended the RDDR mechanism to the following characteristics:

A. Dynamic Neural Fields

The RDDR model relies on very simple models of linear neurons, evaluating at each cycle their new state as a weighted sum of their inputs (no memory of the previous state). We have chosen to use the formalism classically used in bio-inspired models: Dynamic Neural Fields (DNF) [29][30]. In DNF, the activation state *u* is controlled by a differential equation (cf Eq. 8 for its discretized version, actually used for the simulations) with a leak, a non-linearity represented by the function f and the parameter $0<\delta<1$ ensuring a contracting dynamics. In this equation, k is the index iterating inside the neural field (i.e. representing lateral connectivity) and j is an index on the input structures of the neural field (here the feed-forward flow). h represents the base activity or the noise and will be used later.

$$u_i(t+1) = f\left[u_i(t) + \delta\left[-u_i(t) + \sum_j W_{ij}x_j + \sum_k A_{ik}u_k + h\right]\right] \quad (8)$$

B. Sensorimotor cortical axis

It is reported in [9] that selection of action is not performed on the set of all possible actions but on a restricted set of actions, suggested by the perceptive scene and preactivated in the motor cortex, through the associative parietal cortex. This principle is also consistent with the theory of enaction and the principle of affordance [31]. We have chosen to allow for such a mechanism, by adding a very simple sensorimotor cortical axis. It is composed of, in the posterior cortex, a sensory (here visual) area and an associative (parietal) area and of, in the frontal cortex, a motor map.

During the reinforcement learning stages in the basal loop, an associative learning stores in the associative map the links to all the actions that have been associated with a given perception. Later, when the same stimulus is perceived in the sensory map, the corresponding set of potential actions will be preactivated in the motor map, through the associative map. Initially, these potential actions will be kept below the triggering threshold by the tonic inhibition of GPi/SNr (waiting for one of them to be disinhibited for action) but their preactivation will be sufficient to activate the Striatum.

Let us also underline that adding this mechanism is mainly motivated by our will to better stick to the biological reality. At the moment, its only effect on the mechanism of reduction of dimensionality is to reduce the combinatorial of activation, which is not of great interest considering the small size of the data that we manipulate here. Moreover, this new mechanism will make more difficult the change of policy for the association of a perception to a new action. This will partly motivate the new exploration mechanism described in paragraph III. *F.* below.

C. Closing the basal loop

The main modification that we have brought to the original RDDR study is to implement the whole basal loop (Cortex-Striatum-GPi/SNr-Thalamus-Cortex) to see if the reduction of dimensionality emerges in this more natural and dynamic closed loop. Particularly, since we use DNF, evaluating units iteratively, it was of primary importance to observe how the selection of action was progressively emerging from the stimulus and the corresponding preactivated actions.

Interestingly, the motor area is at the centre of the more complete network that we have designed (cf Fig. 1), at the crossroad of the basal loop and the sensorimotor cortical axis. Each unit in the motor area updates its activity from its leak, the associative cortical input and the feed-back of the basal loop, sent by GPi/SNr through the Thalamus. This updated activity is sent (together with the sensory activity) to the Striatum to be projected onto the current principal components until one unit in the motor area goes over the threshold and triggers its action and inhibits the others.

Based on biological data, units in the Striatum, the Thalamus and the motor cortex are non-linear and a fixed lateral inhibition is set in the motor area.

D. Learning protocol

We have developed a learning protocol not more complex than the one used in [2] but more ecological from an operant conditioning point of view. Five different stimuli (e.g. colors) can be proposed to the subject who can answer by triggering five different actions (e.g. arm movements to reach five different buttons). The hidden rewarding rule associates each stimulus to a unique action. Two rules were implemented. The first one associates stimulus i with action i; the second one inverts action 2 and 3 and action 4 and 5. Changing from rule 1 to rule 2 (and vice versa) can be done without notice.

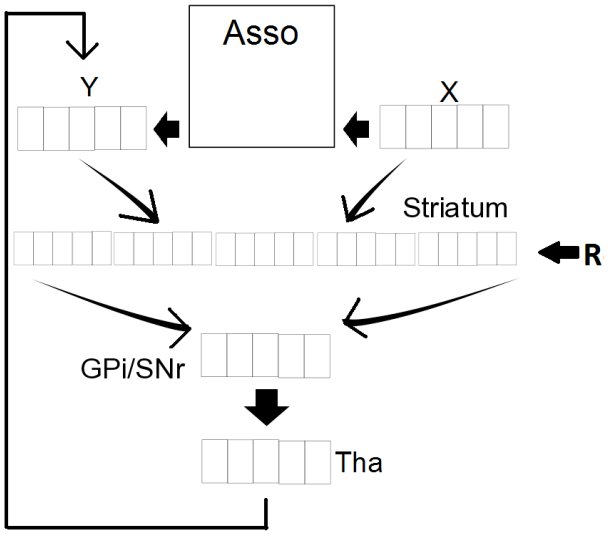

Fig 1.: General architecture of the network. X: sensory cortex; Asso: associative cortex; Y: motor cortex; Tha: Ventro-median nuclei of the thalamus; R: reward brought by SNc.

E. Rewarding mechanism

In previous experiments [2], a reward was systematically associated with each stimulus. Here, to be more consistent with an ecological protocol, the reward is only given when the action is selected and triggered, as a result of its supposed impact in the environment. Consequently, a stimulus is initially proposed and elicits possible actions. Both information will activate the BG and iterate in the basal loop until one action is selected. Depending on the current rewarding rule, the value of the reward will be 0 or 1 and the

W weight matrix will be updated accordingly. If a new rewarding rule appears, a new action can be chosen without pre-activation, thanks to the exploration mechanism described below.

F. Exploration mechanism

The DNF equation (Eq. 8) includes a term for noise and base activity. Such fluctuations are a strong characteristic of neuronal systems and can yield spontaneous activity and non-deterministic response. This is particularly interesting in operant conditioning, where an exploration mechanism is often very useful [7]. Also, in our present case, the principle of preactivation of action must be counterbalanced to be able to elicit new actions when needed. We have implemented such a mechanism in the GPi/SNr layer, for the critical choice of the action to be disinhibited: a zero-mean normally distributed random variable is added to the evaluation of each unit, before selecting the one with the highest activity in the non-linear evaluation of the Thalamus nucleus which will in turn contribute to the activity of the motor cortex.

IV. SIMULATION RESULTS

We report here preliminary results obtained with this architecture.

A. Structure of the network

The brain structures involved in the considered task are modeled mathematically using matrices and vectors. More precisely (cf Fig.1), the motor cortex Y is represented by a real valued vector of dimension 5, the sensory cortex X by a vector of dimension 5 which takes binary values (1 when the perception is present and 0 when it is not present), the associative parietal cortex by a matrix Asso (a binary 5*5 matrix), the BG input, the Striatum, by a real valued vector of dimensionality 25, the BG output, GPi/SNr, by a real valued vector of dimensionality 5 and the Ventro-Median nucleus of the Thalamus Tha by a binary vector of dimensionality 5. Non-linearities in Y and Striatum are obtained by a *tanh* function shifted to positive values.

B. Behavior of the network

The task of the network is to replicate the experiment described in section III.*D* above, where the rewarding rule can be modified without notice and the subject has to discover again the new rule. On this basis, many training examples are given to the network and it is observed how the network learns to associate actions to perceptions as a function of the rewarding rule.

To sum up, we observed that after some thousands of training samples, the relation was learnt and the correct actions generally triggered, except when the exploration mechanism was choosing another action.

To more precisely measure the behavior of the network, two kinds of evaluation were carried out. Firstly, as reported in Table 1, we measured the angles between the computed and the desired principal direction. Indeed, due to non-linearities and closed loop, it is no more possible to decompress the network back to its original representation, but it is possible to analytically compute the theoretical principal directions and to compare them with the directions extracted by the rows of the W matrix (corresponding to the vectors created by the weights linking the input vector to each neuron in GPi/SNr). They are supposed to extract the principal components and to be orthogonal one with the other. For the first rewarding rule, we calculated that vectors analytically and compared the vectors extracted by learning to that desired vectors. As shown in Table 1 below, it was observed that the angles between the computed and the desired weight vectors were small, the remaining value being due to the exploration mechanism. It is also observed that the angles between the principal directions are generally close to 90°.

	W_1	W_2	W_3	W_4	W_5
W_1*	**13.8**	**76.3**	**91.3**	**90.1**	**91.0**
W_2*	**90.7**	**26.4**	**65.1**	**82.2**	**91.3**
W_3*	**90.0**	**89.9**	**15.8**	**104.3**	**96.7**
W_4*	**90.1**	**90.3**	**90.3**	**1.4**	**91.3**
W_5*	**89.9**	**90.7**	**90.1**	**89.8**	**0.77**

Table 1: Angles (in degrees) of W achieved (columnwise) versus ideal W desired (rowwise), for the first rewarding rule.

Secondly, it is also important to check that the elements in the A matrix (the inhibitory lateral connections in S) converge to zero, as the number of trials increases. These inhibitory connections are responsible for making the directions in the W matrix orthogonal and, reciprocally, their null values indicate that orthogonality is achieved in W. Here again, as depicted in Fig. 2, the norm of A is not exactly 0 (but very small) due to the exploration tendency.

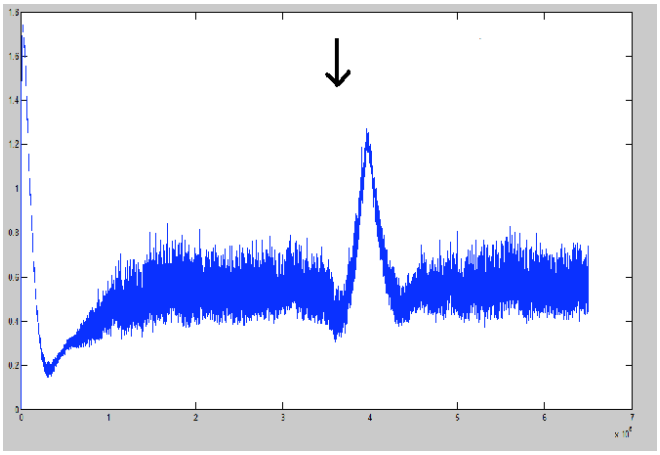

Fig. 2: Evolution of the norm of matrix A (inhibitory lateral weights) with learning. Changing the rewarding rule (shown by the arrow) leads to an increase of the norm and subsequent learning to its decrease. The norm is not completely null because of the exploration mechanism.

Now, when the rewarding rule is modified, this will imply a sudden increase in the inhibitory weights A, as shown in Fig. 2. Subsequent learning will extract another reduced representation with other orthogonal principal directions and, accordingly, will decrease again the weights in the A

matrix. When unseen action–perception associations are proposed, the exploration tendency makes the subject try other buttons, discover reward associated with the new actions and start learning it.

V. Discussion

A. Main results

The RDDR model has raised a new research assumption, hypothetizing that a reduction of dimensionality similar to a PCA could take place in the direct basal loop. An original learning rule modulating the extraction of the principal components by the signal of reinforcement has been proposed and its capacity to build that representation of information has been assessed in [2]. In the present paper, we extend this model to more realistic formalism and dynamics of neuronal computation, more realistic structure of network and more realistic learning protocol. Despite these heavy modifications, it is shown here that the reduction of dimensionality modulated by reinforcement still operates efficiently and that the selection of action is still of good quality. We also think that this more complete system is a better substratum to collaborate with neuroscientists towards a better understanding of its dynamics of information representation, as we are currently doing.

B. Perspectives

Ongoing work also corresponds to reduce the problems due to the exploration mechanism and to propose to design exploration as a function of the cumulated number of trials without reward. This can be interpreted as the tendency to randomly try new buttons if not much reward has been obtained for a long time. Inversely, if reward is regularly obtained, this tendency of exploration decreases.

Our current perspectives of work are twofold, both oriented toward an increased biological inspiration for our system. On the one hand, comparison with biology will be deeper if the size of the network is larger. Particularly, using more units in the layers could give rise to self-organization and topological phenomena, as observed in most of the concerned neuronal structures [13].

On the other hand, it can be observed that most of the efforts reported here are related to the Actor part of the architecture. Developing a more realistic Critic is also of major importance to improve the system. The major questions to be explored correspond to bring a more precise view of the dopaminergic influence on the system and to better articulate this operant conditioning to its respondent counterpart [19].

References

[1] W. Schultz, P. Dayan, and R.R. Montague. A neural substrate of prediction and reward. *Science* 275: 1593-1599, 1997.

[2] I. Bar-Gad, G. Morris, and H. Bergman, H. Information processing, dimensionality, reduction and reinforcement in the basal ganglia. *Progr. Neurobiol.*71:439-477, 2003.

[3] J. Fix, N. Rougier, and F. Alexandre. From physiological principles to computational models of the cortex, *Journal of Physiology - Paris*, 101, 1-3, pp. 32-39, 2007.

[4] T. Kohonen, *Self-Organizing Maps*, New-York: Springer-Verlag, 1997.

[5] H. Frezza-Buet, and F. Alexandre, From a biological to a computational model for the autonomous behavior of an animat, Information Sciences, 144(1-4), p. 1-43, Jul. 2002.

[6] J. Fix. A computational approach to the control of voluntary saccadic eye movements. *International Conference on Cognitive Neurodynamics*, ICCN-2007. (2007).

[7] R.S. Sutton, and A.G. Barto, (1998). *Reinforcement Learning: An Introduction*. The MIT Press Cambridge, MA.

[8] A.D. Craig. How do you feel - now? The anterior insula and human awareness, *Nat. Rev. Neurosci*. 10, pp. 59-70, 2009.

[9] M.A. Goodale, and G.K. Humphrey. "The objects of action and perception", *Cognition* 67, pp. 181-207, 1998.

[10] P. Cisek. (2005) "Neural representations of motor plans, desired trajectories, and controlled objects". *Cognitive Processing*. 6: 15-24.

[11] P. Redgrave, T.J. Prescott, and K. Gurney. (1999), The basal ganglia: a vertebrate solution to the selection problem?, *Neuroscience*, 89:1009-1023.

[12] D. Joel, Y. Niv, and E. Ruppin. (2002) - Actor-critic models of the basal ganglia: New anatomical and computational perspectives - *Neural Networks* 15, 535-547.

[13] S.N. Haber, J.L. Fudge, and N.R. McFarland, NR (2000). Striatonigrostriatal Pathways in Primates Form an Ascending Spiral from the Shell to the Dorsolateral Striatum; *The Journal of Neuroscience*, 20(6): 2369–2382.

[14] R.S. Sutton, 1988, Learning to Predict by the Method of Temporal Differences, *Machine Learning*, 3, pp. 9-448.

[15] K. Gurney, T. J. Prescott and P. Redgrave (2001). "A computational model of action selection in the basal ganglia. I. A new functional anatomy." *Biol. Cybern*. 84: 401-410.

[16] A. Leblois, T. Boraud, W. Meissner, H. Bergman, and D. Hansel. (2006). Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *J. Neurosci*. 26, 3567-3583.

[17] G.E. Alexander, M.R. DeLong, and P.L. Strick. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex, *Ann. Rev. Neurosci*, 9:357-81.

[18] N.D. Daw, and K. Doya. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16:199–204.

[19] R.C. O'Reilly, M.J. Frank, T.E. Hazy, and B. Watz. (2007). PVLV: The Primary Value and Learned Value Pavlovian Learning Algorithm. *Behavioral Neuroscience*, 121, 31-49.

[20] K.I. Diamantaras, and S.Y. Kung. (1996). *Principal Component Neural Networks: Theory and Applications*. Toronto: Wiley.

[21] E. Oja, (1982). A simplified neuron model as a principal component analyzer. *J.Math.Biol*., 15, 267-273.

[22] A. Carlson, (1990). Anti-Hebbian learning in a nonlinear neural network, *Biol. Cybern*., vol. 64, pp. 171–176.

[23] P.J. Zufiria, and J.A. Berzal, (2007) Analysis of Hebbian Models with Lateral Weight Connections ; F. Sandoval et al. (Eds.): *International Workshop on Artificial Neural Networks*, IWANN 2007, LNCS 4507, pp. 39–46, Springer-Verlag.

[24] S.Y. Kung, and K.I. Diamantaras, (1990). A neural network learning algorithm for adaptive principal component extraction (APEX). *Proc. IEEE Int. Conf. Acoustics Speech Signal Process*. 2, 861--864.

[25] J. Rubner, and K. Schulten (1990) Development of feature detectors by self-organization: A network model. *Biol Cybern* 62:193-199.

[26] P. Földiák, (1989) Adaptive network for optimal linear feature extraction, *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, Washington DC., vol. 1, pp. 401-405 (IEEE Press, New York))

[27] K.I. Diamantaras, Neural Networks and Principal Component Analysis, Chapter 8 of *Handbook of Neural Network Signal Processing*. Yu Hen Hu and Jeng-Neng Hwang, Editors. CRC Press, New York, 2002.

[28] D. Jaeger, H. Kita, and C.J. Wilson, 1994. Surround inhibition among projection neurons is weak or nonexistent in the rat neostriatum, *Journal of Neurophysiology*, 72, 2555–2558

[29] S. Amari, Dynamics of pattern formation in lateral-inhibition type neural fields, *Biological Cybernetics* 27 (2) (1977) 77–87.

[30] W. Erlhagen, and G. Schoener, Dynamic field theory of movement preparation, *Psychol Rev* 109 (3) (2002) 545–72.

[31] J.J. Gibson, (1979). The Ecological Approach to Visual Perception. Boston : Houghton Mifflin.