Characterizing a Brain-based Value-function Approximator

Patrick Connor and Thomas Trappenberg

Department of Computer Science, Dalhousie University
patrick.connor@dal.ca, tt@cs.dal.ca

Abstract. The field of Reinforcement Learning (RL) in machine learning relates significantly to the domains of classical and instrumental conditioning in psychology, which give an understanding of biology's approach to RL. In recent years, there has been a thrust to correlate some machine learning RL algorithms with brain structure and function, a benefit to both fields. Our focus has been on one such structure, the striatum, from which we have built a general model. In machine learning terms, this model is equivalent to a value-function approximator (VFA) that learns according to Temporal Difference error. In keeping with a biological approach to RL, the present work seeks to evaluate the robustness of this striatum-based VFA using biological criteria. We selected five classical conditioning tests to expose the learning accuracy and efficiency of the VFA for simple state-value associations. Manually setting the VFA's many parameters to reasonable values, we characterize it by varying each parameter independently and repeatedly running the tests. The results show that this VFA is both capable of performing the selected tests and is quite robust to changes in parameters. Test results also reveal aspects of how this VFA encodes reward value.

1 Introduction

Over the last several decades, our understanding of RL has been advanced by psychology and neuroscience through classical/instrumental conditioning experiments and brain signal recording studies (fMRI, electrophysiological recording, etc.). Over the same period, the machine learning field has been investigating potential RL algorithms. There has been some convergence of these fields, notably the discovery that the activity of a group of dopamine neurons in the brain resembles the Temporal Difference (TD) error in TD learning [1]. One research focus in machine learning RL is the mapping of expected future reward value to states (state-value mapping) from as little experience (state-value sampling) as possible. Living things clearly grapple with this problem, continually updating their beliefs about expected rewards from their limited experience. Indeed, the field of classical conditioning, which relies heavily on animal behavioural experiments, has explored a variety of reward-learning scenarios. The obvious need to acquire value for a rewarding state and the need to generalize this to similar circumstances is well recognized by both psychology and machine learning.

What is interesting, however, is that there appear to be other useful reward-learning strategies expressed in classical conditioning phenomena that have not yet translated into machine learning RL. Just as generalization improves learning efficiency by spreading learned value to nearby states, the classical conditioning phenomena of "latent inhibition" and "unovershadowing" appear to improve learning efficiency in their own right.

At the heart of classical conditioning experiments is the presentation of a stimulus (eg. a light, tone, etc.) or combination of stimuli followed by a reward outcome (reward, punishment, or none). When a stimulus is repeatedly presented and there is no change in the reward outcome, latent inhibition [2] sets in, reducing the associability of the stimulus when the change in reward outcome eventually occurs. This promotes association to novel stimuli, which seems appropriate since novel stimuli are more likely to predict a new outcome than familiar stimuli. Latent inhibition *saves the additional experience* otherwise needed to make this distinction clear. Recovery from overshadowing, or "unovershadowing" [3] is one of a family of similar strategies. First, overshadowing is the process of presenting a compound stimulus followed by, say, reward ($S_{AB} \rightarrow R$). Although the compound will learn the full reward value, its constituent stimuli (S_A and S_B) tested separately will also increase in value, where the most salient stimulus (say S_B) gains the most value. In unovershadowing, the most salient stimulus is presented but not rewarded ($S_B \rightarrow 0$) and will naturally lose some of its value. What is surprising, however, is that the absent stimulus (S_A) concurrently increases in value. This allows the animal to not only learn that S_B is less rewarding than it predicted but, by process of elimination, learns that S_A is more rewarding than it predicted. Unovershadowing *saves the need to present and reward S_A explicitly* to increase its value, taking advantage of implicit logic. Whether it is generalization, latent inhibition, or unovershadowing, learning the value-function from *fewer experiences* will assist the animal in making rewarding choices sooner.

These and other RL strategies are found in classical conditioning experiments, where subjects maintain an internal value-function, indicating reward-value based on the rate of their response (eg. lever presses). Since these biological strategies appear beneficial, a machine learning RL system based on RL structures in the brain may prove effective. After a brief review of our brain-based model that does value-function approximation [4], the present work characterizes this VFA to determine its robustness and effectiveness in several classical conditioning tests that are especially relevant to VFAs.

2 Striatal model

The striatum, the input stage of the basal ganglia (BG) brain structure, is a key candidate region on which to base a VFA. The striatum is a convergence point for inputs from all over the brain (specifically, the neocortex [5]), spanning signals of sensation to abstract thought. The majority of striatal neurons project to one another (via axon collaterals) and to other BG nuclei. The synaptic strengths

(i.e. weights) of these projection neurons are modulated by dopamine signals [6] (or the lack thereof), where dopamine neuron activity has been linked to the teaching signal of TD learning [1] mentioned earlier. In addition, several neural recording studies suggest that reward-value is encoded in the striatum [7][8][9], although it is not the only area of the brain that has been implicated in the representation of reward-value [10][11][12].

Our striatal model [4] is shown in Figure 1. The excitatory external input represents a real-world feature (eg. colour wavelength, tonal pitch, etc.) by providing a Gaussian activation profile surrounding a specific feature value (eg. Green, 530 nm). This emulates the "tuning curve" input to the striatum from the neocortex. The model is composed as a one-layer, one-dimensional neural network of striatal projection neurons, each excited by a subset of the external inputs and inhibited by a subset of the other projection neurons, as is the case in the striatum (see [5] and [13]). Each neuron is part of either the direct or indirect pathway, the main information processing routes through the BG, where D1 and D2 are their dominant dopamine receptor subtypes respectively. These pathways tend to behave in an opposite sense, where one increases BG output activity while the other decreases it. The output of the model, $V(S)$, becomes the expected value of an external input (state/stimuli), computed as the sum of the direct pathway neuron activity minus the sum of the indirect pathway neuron activity. Finally, the teaching signal can be formulated in the same way as TD error, but, for the simple one-step prediction tasks used in this work, it is only necessary to use the reward prediction error (RPE), the actual reward minus the expected reward ($RPE = R - V(S)$). A more formal description of the model is provided in Appendix A.

An important novel element in our model is the inclusion of modifiable lateral inhibitory connections. Because of these, the neurons compete, partially suppressing one another. Given an arbitrary combination of external inputs, an associated subset of neurons will become more active than the others because their external input weights correlate most with the external input. Many neurons will also be inactive, suppressed below their base activation threshold.

3 Tests, Measures, and Variables

Conventionally, to evaluate a VFA, one might seek to prove that the VFA's state-values converge for arbitrary state-value maps or seek to test performance on a particular RL task (eg. random walk). Instead, we seek to know how effectively this striatum-based VFA employs certain RL strategies found in classical conditioning to *update the value-function*. This approach puts value-function update strategy first, after which agent actions can be included and convergence proofs and specific RL task comparisons pursued. Also, using classical conditioning tests helps to ascertain whether or not the striatum is responsible for this behaviour. There are a great variety of classical conditioning tests, but to be practical, we limit this to five: two to evaluate state-value mapping *accuracy* and three to evaluate state-value learning *efficiency*. The striatum-based VFA was

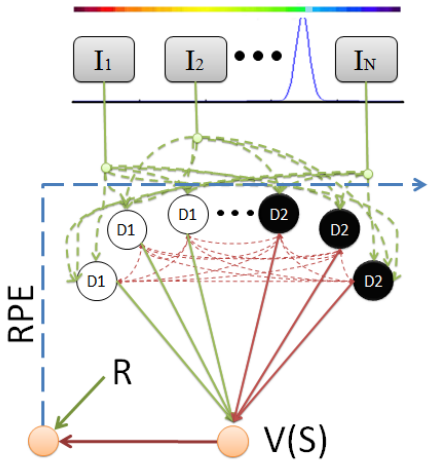

Fig. 1. Diagram of the striatal model. External input is shaped as a Gaussian activity profile surrounding a feature value. Probabilistic inputs (finely dashed lines) from external and lateral sources are excitatory (green) and inhibitory (red) respectively, while the modulatory RPE signal can be either (blue). The direct and indirect pathways are expressed in the two populations of neurons, D1 and D2 respectively, whose activities are accumulated to compute the expected value of the input state/stimulus, $V(S)$.

integrated into simulations of these tests, providing results in terms of measures that are defined for each, as described below. During a test, many trials are run, where one trial consists of presenting a state/stimulus (external input) together with its expected reward-value.

The entry level test for a VFA is the *acquisition* of a state-value. What is also important, however, is that other state-values outside of a reasonable generalization window (eg. Yellow in Fig. 1) are relatively unaffected. The acquisition test, then, pairs a state with a reward value, and compares the state-value to a sample of other state-values. We define the acquisition *effectiveness* measure as

$$E_A(S) = \frac{V(S) - \frac{1}{M}\Sigma_{i=1}^{M} V(S_i)}{V(S)} \tag{1}$$

and consider that acquisition is *observed* when the state-value, $V(S)$, is twice that of the other sampled state-values, $V(S_i)$, or $E_A(S) > 0.5$. Twenty trials are run for each acquisition test. Six $V(S_i)$ samples are used for the comparison.

Secondly, it is important that a VFA be able to represent a variety of state-value mappings. *Negative patterning* is the classical conditioning equivalent of the non-trivial "exclusive-OR" problem, where the subject learns to increase the value of two stimuli, S_A and S_B, while learning zero-value for the compound stimulus S_{AB}. Here, we will define the negative patterning effectiveness as the difference between the average constituent value and the compound value, nor-

malized by the average constituent value, which can be expressed as

$$E_{NP}(S_A, S_B, S_{AB}) = \frac{V(S_A) + V(S_B) - 2V(S_{AB})}{V(S_A) + V(S_B)} \tag{2}$$

Negative patterning is observed while $E_{NP}(S_A, S_B, S_{AB}) > 0$, that is, while the constituents have a higher value than the compound. One-hundred trials of interleaved presentation of the stimuli and their associated rewards are run for each test.

In practical situations, no two experiences are identical, making it critical to generalize state-value learning. *Generalization* also contributes significantly to learning efficiency, spreading learned value to nearby states under the assumption that similar states are likely to have similar expected reward value. This strategy reduces the amount of state-value sampling necessary to achieve reasonable accuracy. For this test, acquisition is performed for a single feature-value and the reward value computed for 500 equally spaced feature-values. Generalization effectiveness will describe the spread of the value as a weighted standard deviation, where feature values are weighted by their associated reward values,

$$E_G(S) = \sqrt{\frac{\Sigma_{i=1}^{N} V(S_i)(i - \Sigma_{k=1}^{N} \frac{kV(S_k)}{\Sigma_{j=1}^{N} V(S_j)})}{\Sigma_{j=1}^{N} V(S_j)}} \tag{3}$$

Generalization will be considered observed when the spread of value is at least 10% of the width of the tuning curve input.

To further enhance learning efficiency, we consider the phenomena of *latent inhibition* described earlier. Latent inhibition's reduction of associability can be achieved by simply reducing the input salience of the familiar stimulus. Then, when this reduced salience stimulus is combined with a novel stimulus and followed by reward, overshadowing will result. Thus, our test of latent inhibition becomes a test of overshadowing, where the novel stimulus (S_A) overshadows the reduced (half) salience stimulus (S_B). We define the latent inhibition effectiveness measure as

$$E_{LI}(S_A, S_B) = \frac{V(S_A) - V(S_B)}{V(S_A) + V(S_B)} \tag{4}$$

where the effect is observed when $E_{LI}(S_A, S_B) > 0$. Thirty trials are run for each test.

Finally, *unovershadowing* appears to improve learning efficiency by process of elimination as described previously. There are other similar phenomena (eg. backward blocking) that raise or lower the value of the absent stimulus, depending on the scenario. The unovershadowing effectiveness is defined as

$$E_{UO}(S_A, S_B) = -\frac{\Delta V(S_A)}{\Delta V(S_B)} \tag{5}$$

where $\Delta V(S_X)$ is the change of value of stimulus S_X from one trial to the next and observability occurs when $E_{UO}(S_A, S_B) > 0$. Here, unovershadowing

is simulated by first performing the process of overshadowing (see above) with equally salient stimuli, followed by 100 trials of S_B presentation without reward.

Ultimately, we seek to determine the robustness of the simulation of these five tests to changes in the VFA's parameters. Because the parameter space is very large and a full search unnecessary, we found initial values where all tests were observed and varied the parameters independently through their valid ranges. This process *characterizes* the VFA, showing the conditions under which the tests break down.

Besides parameters associated directly with the VFA there are others acknowledged here that are better associated with the particular RL task to be solved. To simulate input noise, Gaussian noise is added to the external input and rectified, where its standard deviation is the parameter varied in the tests. Since the intensity of stimuli and rewards may vary, the salience of inputs and rewards are multiplied by parameters varied between 0.01 and 1.

4 Results

Figures 2 and 3 represent the results for all five tests over 17 parameters. Since the VFA connectivity and initial weights are randomly initialized, each test and parameter combination was run 20 times to provide uncertainty estimates. The observability curve (upper panel) for each parameter is a summary of the more detailed effectiveness curves (lower panel). For each parameter, the observability curves from the five tests are multiplied together, giving an "intersection" of observability. So, wherever observability is zero, it means that at least one test is not observed for that parameter setting and when observability is one, all tests are observed. For example, once the lateral learning rate, β, becomes negative, unovershadowing effectiveness disappears (goes negative) and unovershadowing is no longer observed. So, the summary observability curve is zero for $\beta < 0$ because not all of the tests were observed in this range. In constrast, when $\beta > 0$, all tests are observed. The effectiveness curves, whose vertical bars denote standard deviation, are colour coded: acquisition (blue), negative patterning (green), generalization (red), latent inhibition (cyan), and unovershadowing (violet). Note that only the effectiveness of observed cases are given. Also, when part of an effectiveness curve is missing in the graph, this indicates that there were no cases of the associated parameter values where the effect was observed. In the effectiveness graphs, a black dotted vertical line indicates that parameter's setting while the other parameters were independently varied.

5 Discussion

The results show that this VFA is generally robust to changes in feature values. There are, however, regions where observability disappears within its valid parameter range. From the causes of low observability and key trends in effectiveness curves given in the results, the structural and functional details necessary to successfully reproduce these five effects are described.

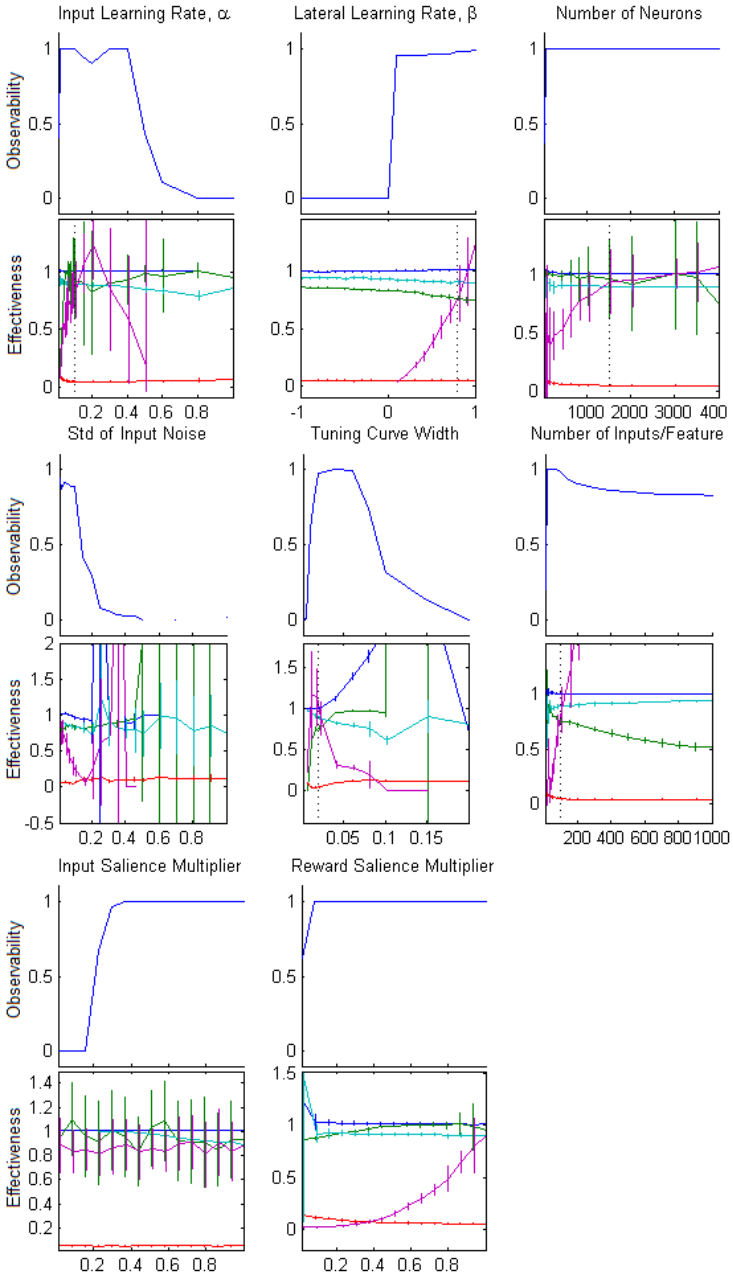

Fig. 2. Intersection of observability curves (top) with effectiveness curves (bottom) where error bars represent the standard deviation of effectiveness. Effectiveness curves are coloured according to test: acquisition (blue), negative patterning (green), generalization (red), latent inhibition (cyan), and unovershadowing (violet).

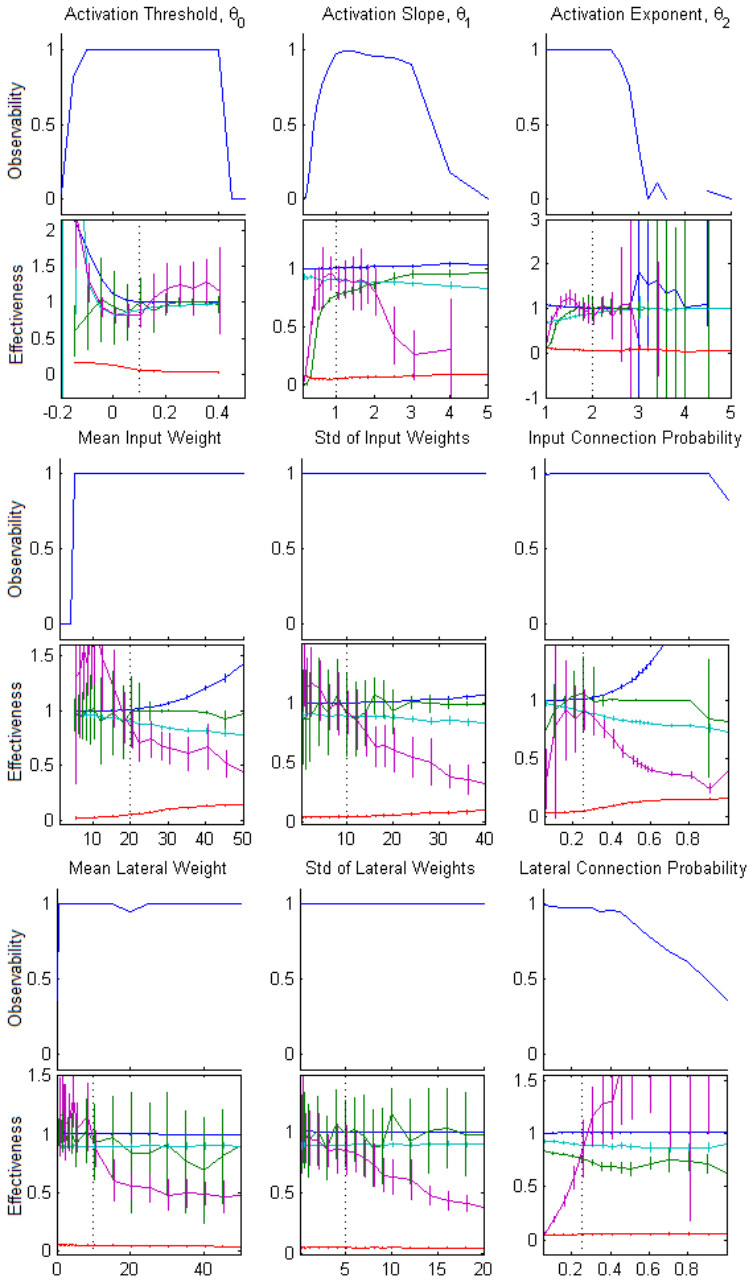

Fig. 3. Intersection of observability curves (top) with effectiveness curves (bottom) where error bars represent the standard deviation of effectiveness. Effectiveness curves are coloured according to test: acquisition (blue), negative patterning (green), generalization (red), latent inhibition (cyan), and unovershadowing (violet).

Acquisition was prevented in only three cases: high activation threshold (θ_0 in Appendix A), low input salience, and low mean input weight. As the activation threshold increases, fewer neurons are active because fewer have internal activations that exceed it. Likewise, internal activations are weak when the input salience or mean input weight is too low. Since learning only occurs in neurons that are active (see equations 8 and 9, Appendix A), neither acquisition nor any other test will learn when neurons are silent. Acquisition was otherwise robust to varying the VFA parameters. This is not surprising since the Rescorla-Wagner model [14] of classical conditioning acquires reward value in much the same way, the key ingredient being that they both learn in proportion to RPE and input salience.

As different inputs are presented to the system it becomes clear that the subset of active neurons is input specific, enabling inputs to be represented by separate populations of neurons. A lateral inhibitory network put forth by Rabinovich et al. [15] similarly showed that asymmetric lateral connectivity (implemented in the striatum-based VFA by low connection probability) led to similar input-specific patterns of activation as well. This form of activity also resembles that of sparse coarse coding [16], another value-function approximation technique that uses a state-specific subsets of elements to represent state-value. This value-encoding strategy is critical for negative patterning because it allows a compound stimulus (S_{AB}) and its constituent stimuli (S_A and S_B) to be represented in different (although overlapping) populations. Then S_A and S_B can have a strong positive value while S_{AB} holds zero value. In the results we see negative patterning sometimes failing for high lateral connection probabilities. In this scenario, we find that it becomes difficult to separately represent the constituent and compound stimuli because there is too much overlap between their active subsets.

Generalization, like acquisition, is robust, not being eliminated except when all neurons are silent. In the effectiveness curves, the generalization is always greater than or equal to the tuning curve width. As the tuning curve width is increased, a proportional increase in generalization effectiveness can be seen as well. When the generalization effectiveness is greater than the tuning curve width, closer examination reveals it to be either noise or an average increase/decrease in the state-values outside a reasonable generalization window. So, the generalization present in the VFA is actually due to the activity profile of the input rather than anything in the VFA per se. The VFA does support this means generalization, however, in that the amount of subset overlap between two feature values is proportional to the overlap between their activity profiles. This, too, accords with the approach taken by sparse coarse coding.

Again, the practical benefit of latent inhibition is its ability to reduce association of familiar, ineffectual stimuli with reward outcome. We implemented this as a test of overshadowing, where the familiar stimulus was half the salience of a novel stimulus. If reward associations were simply made in proportion to a stimulus' input salience, as is the case for the Rescorla-Wagner model (not shown), our tests should return latent inhibition effectiveness values of $\sim$0.6. However, we

see effectiveness values typically between 0.85 and 0.95, which seems to suggest that the novel stimulus really dominates the association and the familiar stimulus receives disproportionately little association. As mentioned earlier, however, this lateral inhibitory model of the striatum has competitive properties. It appears that this makes up the difference in the effectiveness measure, where the familiar (less salient) stimulus is not very competitive and is overwhelmed by background activity when presented alone.

Unovershadowing is especially affected by the lateral learning rate. A sharp increase in unovershadowing observability occurs as the lateral learning rate becomes positive. In agreement with equations 8 and 9 (Appendix A), this suggests that for unovershadowing to be observed, a neuron's lateral weights must increase when its input weights increase, and decrease when they decrease. This is unusual since, if gradient descent had been used to derive the lateral weight update equation as was done for the input weight update equation, the lateral weights would have learned in the opposite sense (i.e. would have increased when input weights decreased, etc.).

Parameters which show some of the least robustness include the standard deviation of input noise, tuning curve width, activation exponent (θ_2), and number of inputs per feature. In all of these cases, we see the neuron activity becoming too weak or too strong leading either to test incompletion or system instability, as evidenced by extreme changes in effectiveness measures. If instead of independently varying the parameters, the learning rates were also adjusted to lead to proper activity levels throughout the test, these results would show greater robustness and better reveal the effects of varying these parameters.

6 Conclusions and Future Work

We have characterized a brain-based VFA in terms of classical conditioning tests that represent RL strategies for accurate and efficient value-function updates. This approach is not limited to brain-based VFAs, but may be applied to others with the assumption that these tests represent RL strategies worth emulating. Testing the striatum-based VFA for convergence for arbitrary state-value maps and making specific RL task comparisons are worthwhile and should be investigated.

Systematically varying the VFA parameters led to both assessing the model's degree of robustness and helping to determine how the VFA is capable of successfully performing the tests. This striatum-based VFA has shown to be generally robust to changes in the parameters in these tests, supporting the notion that the striatum may be the seat of general purpose reward-value encoding in the brain. The VFA's ability to effectively demonstrate latent inhibition and unovershadowing is especially worthy of note, being emergent properties of the competitive nature and lateral learning in the VFA.

Acknowledgments

Funding for this work was supported in part by the Walter C. Sumner Foundation, CIHR, and NSERC.

Appendix A

Formally, the striatum-based neural network can be represented as:

$$\tau \frac{du(x,t)}{dt} = -u(x,t) + \int_y w^I(x,y)I(y,t)dy - \int_z w^L(x,z)r(u(z,t))dz \quad (6)$$

$$r(u) = \begin{cases} \theta_1 \left(u - \theta_0\right)^{\theta_2}, & u > \theta_0 \\ 0 & otherwise \end{cases} \quad (7)$$

where w^I and w^L are the synaptic weights connecting external input ($I(y,t)$) and lateral inputs from other neurons respectively. The activation function, $r(u)$, transforms the internal state (average membrane potential) to an instantaneous population firing rate. Parameter θ_0 is the x-intercept, θ_1 is the slope multiplier, and θ_2 is the exponent ($r(u)$ is a threshold-linear activation function when $\theta_2 = 1$). Neurons only activate if their internal state is greater than the threshold.

Learning in the model happens in two ways. Weights receiving external inputs learn according to gradient descent, minimizing the squared RPE ($J = \frac{1}{2}RPE^2$), resulting in

$$w^I(x,y) = w^I(x,y) + \alpha D(x) RPE \left[\theta_2\theta_1(u(x,t) - \theta_0)^{\theta_2 - 1} I(y,t)\right] \quad (8)$$

where α is the learning rate and $D(x) = 1$ for direct pathway neurons and -1 for indirect pathway neurons. The weights receiving lateral inputs learn in an a way that opposes the gradient,

$$w^L(x,z) = w^L(x,z) + \alpha\beta D(x) RPE \left[\theta_2\theta_1(u(x,t) - \theta_0)^{\theta_2 - 1} Q(u(z,t))\right] \quad (9)$$

where β is the relative learning rate for the lateral input connections, and $Q(u) = 1$ for $u > \theta_0$ and 0 otherwise. Just as for $r(u)$, there is no weight change for either of these learning equations when $u(x,t) < \theta_0$.

References

1. Schultz, W.: Predictive Reward Signal of Dopamine Neurons. J Neurophysiol **80**(1) (1998) 1–27
2. Lubow, R.E.: Latent inhibition. Psychological Bulletin **79** (1973) 398–407

3. Louis D. Matzel, T.R.S., Miller, R.R.: Recovery of an overshadowed association achieved by extinction of the overshadowing stimulus. Learning and Motivation **16**(4) (1985) 398–412
4. Connor, P.C., Trappenberg, T.: A striatal model of classical conditioning. In preparation (2011)
5. Wilson, C.J. In: Basal Ganglia. Fifth edn. Oxford University Press, Inc. (2004) 361–413
6. Wickens, J.R., Begg, A.J., Arbuthnott, G.W.: Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. Neuroscience **70**(1) (January 1996) 1–5
7. Hori, Y., Minamimoto, T., Kimura, M.: Neuronal encoding of reward value and direction of actions in the primate putamen. Journal of Neurophysiology **102**(6) (2009) 3530–3543
8. Lau, B., Glimcher, P.W.: Value representations in the primate striatum during matching behavior. Neuron **58**(3) (2008) 451–463
9. Samejima, K.: Representation of Action-Specific reward values in the striatum. Science **310**(5752) (2005) 1337–1340
10. Bromberg-Martin, E.S., Hikosaka, O., Nakamura, K.: Coding of task reward value in the dorsal raphe nucleus. Journal of Neuroscience **30**(18) (2010) 6262–6272
11. Gottfried, J.A.: Encoding predictive reward value in human amygdala and orbitofrontal cortex. Science **301**(5636) (2003) 1104–1107
12. Roesch, M.R.: Neuronal activity related to reward value and motivation in primate frontal cortex. Science **304**(5668) (2004) 307–310
13. Wickens, J.R., Arbuthnott, G.W., Shindou, T.: Simulation of GABA function in the basal ganglia: computational models of GABAergic mechanisms in basal ganglia function. In: Progress in Brain Research. Volume 160. Elsevier (2007) 313–329
14. Rescorla, R.A., Wagner, A.R.: A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In Black, A.H., Prokasy, W.F., eds.: Classical Conditioning II. New York: Appleton-Century-Crofts (1972)
15. M. I. Rabinovich, R. Huerta, A.V.H.D.I.A.M.S.G.L.: Dynamical coding of sensory information with competitive networks. J. Physiol. (Paris) **94** (2000) 465–471
16. Sutton, R.S.: Generalization in reinforcement learning: Successful examples using sparse coarse coding. In: Advances in Neural Information Processing Systems 8, MIT Press (1996) 1038–1044