A Biology-Inspired Algorithm for Neural Function Approximation in Reinforcement Learning

Raphael Holca-Lamarre
Bernstein Center for Computational Neuroscience
Technische Universität Berlin
10961 Berlin
raphael@bccn-berlin.de

Jörg Lücke
Dept. of Medical Physics and Acoustics
Universität Oldenburg
26129 Oldenburg
joerg.luecke@uni-oldenburg.de

Klaus Obermayer
Bernstein Center for Computational Neuroscience
Technische Universität Berlin
10961 Berlin
klaus.obermayer@mailbox.tu-berlin.de

Abstract

In animals, the ventral tegmental area (VTA) contains dopamine (DA)-releasing neurons whose activity reflects reward prediction errors [1]. Activity in the projections from the VTA to sensory cortices is known to trigger long-term changes in sensory representations [2] and to be required for reward-motivated discrimination learning [3]. These findings suggest that activity in the VTA, signalling reward prediction errors, refines sensory representations, yielding improved sensory discrimination abilities. In the present work, we develop a learning algorithm for function approximation in reinforcement learning inspired from this speculative function of the VTA. We extend a Hebbian neural model of representation learning [4] to include the effects of the VTA. In line with the observed effect of DA on synaptic plasticity, activation of the VTA in our algorithm is instantiated as a transient modification in the network's learning rate. The network is subjected to a classification task and is rewarded for taking correct classification decisions. The network is allowed to take exploratory classification decisions. Based on its exploratory behaviour, the network predicts at each trial whether or not it expects a reward. The difference between the predicted and received reward makes up a reward prediction error; this error activates the VTA. We perform parameter exploration to determine the optimal value of VTA activation depending on the reward prediction error. We find that the VTA activation values that are optimal with respect to classification performances in the model are in close qualitative agreement with those observed in animals. During training, VTA activation refines synaptic weights with respect to the classification task and significantly improves the network's performance on the task. The classification performance of our algorithm on the MNIST dataset compares favourably to gradient-based neural network of similar architecture. The algorithm presented in this work provides a possible model for reward-motivated discrimination learning in animals.

1 Introduction

1.1 Neural Function Approximation

In many real-world reinforcement learning scenarios, state spaces are either very large or continuous, which prohibits their exhaustive exploration. In these cases, function approximation can be used to generalise between similar states. In recent years, deep neural networks have successfully been used to perform such function approximation, yielding impressive results on problems involving large state spaces [5, 6]. These networks typically rely on error back-propagation to learn the parameters of the function approximation. Although error back-propagation gives outstanding performances in practice, it is remote from learning mechanisms in animal brains. Indeed, error back-propagation requires large, fully labelled datasets which is unrealistic for biological systems; it also remains unclear how back-propagation could be implemented in biological neural networks. As we better understand the learning algorithms implemented in biological neural systems, it becomes possible to reverse engineer these algorithms in artificial systems. In this work, we draw inspiration from a speculative dopamine-mediated learning mechanism in animals to develop a neural function approximation algorithm for reinforcement learning.

1.2 Reinforcement-Based Learning and Dopamine

Human studies have shown that reinforcement signals are necessary in some discrimination tasks for subjects to achieve maximal performances [7]. These reinforcement signals activate neural circuits in sensory cortices [8] which appear to control synaptic plasticity [9, 10]. In mammals, dopamine (DA)-releasing neurons of the ventral tegmental area (VTA) are thought to encode such reinforcement signals [1]. Activation of VTA projections to sensory cortices is known to trigger long-term plastic changes in the response properties of neurons [2]. These projections are also necessary for reward-motivated discrimination learning [3]. Taken together, these findings suggest that DAergic neurons of the VTA participate in reinforcement-based discrimination learning. Here, we replicate this hypothetical role of the VTA in guiding synaptic plasticity in a learning algorithm for neural function approximation.

1.3 Dopamine and Synaptic Plasticity

At the level of single synapses, DA facilitates the induction of long-term plasticity [10, 11]. At the level of receptive fields, pairing a stimulus with activation of the VTA leads to long-term synaptic modifications resulting in an increase in the responses of neurons to the paired stimulus [2]. For these reasons, we model VTA activation in our algorithm as a transient multiplication of the learning rate (i.e., a transient facilitation of synaptic plasticity). It should be noted that, in some conditions, DA-pairing protocols may also lead to decreases in the responses of neurons to stimuli [12], presumably due to synaptic depression. In our model, this corresponds to a negative multiplication of the learning rate, leading to weight decay.

2 Dopamine-Inspired Neural Function Approximation

We model the effects of the VTA on synaptic plasticity in a Hebbian-learning neural network [4]. We train the network on a classification task of images of the MNIST dataset [13]. These input images provide stimuli of intermediate complexity and high-dimensionality, akin to natural sensory stimuli. Although not a typical reinforcement learning framework, the MNIST classification task can be seen as a special case of a Markov decision process where previous action selections have no effect on future rewards and where complete reinforcement signal for a given action selection is given immediately after the selection (i.e., without temporal credit assignment problem).

Our neural network contains three layers: an input layer (784 neurons), a hidden layer (variable number of neurons), and an output layer (10 neurons). The pixel values of input images activate neurons in the input layer. Activity then propagates through the network in the following steps:

Feedforward Inhibition. Feedforward inhibition normalises the activation of input neurons by the sum of their activations:

$$y_d = (A - D)\frac{\tilde{y}_d}{\sum_{d'} \tilde{y}_{d'}} + 1 \tag{1}$$

where A is a normalisation constant, D is the number of input neurons and $\tilde{y}$ is the activation of individual input neurons before the normalisation.

Input Integration. Hidden neurons integrate their input through a weighted sum:

$$I_c = \sum_d y_d \log(W_{cd}) \tag{2}$$

where I is the integrated input of hidden neurons, y is the activation of input neurons, and W is the weight matrix.

Lateral Inhibition. The integrated input is fed through a softmax function representing lateral inhibition:

$$s_c = \frac{\exp(I_c/\tau)}{\sum_{c'} \exp(I_{c'}/\tau)} \tag{3}$$

where s is the activation of hidden neurons after the inhibition and τ is the temperature parameter of the softmax function (corresponding to the strength of lateral inhibition). The input integration and and lateral inhibition steps are then repeated for the classification layer.

Reward Delivery. The classification neuron with the greatest activation value is taken as the classification decision of the network. The network is rewarded for taking correct classification decisions (+*rew*) and not rewarded for incorrect decisions (-*rew*). Rewards take binary values (0,1) and have in themselves no effect on the network.

Exploration and Reward Expectation The network is allowed to take either exploitative or exploratory classification decisions. Exploratory decisions are taken by injecting additive noise in the activation of hidden neurons. If the network takes an exploitative decision it is said to expect a reward (+*exp*); if it takes an exploratory decision it is said not to expect a reward (-*exp*). Reward expectations are binary (0,1) and have in themselves no effect on the network.

VTA Activation The difference between the expected and delivered rewards gives rise to a reward prediction error. There are four reward prediction error scenarios (figure 1): the network may (1) expect a reward and get rewarded (+*exp* +*rew*), (2) expect a reward but not get rewarded (+*exp* -*rew*), (3) not expect a reward but get rewarded (-*exp* +*rew*), and (4) not expect a reward and not get rewarded (-*exp* -*rew*).

Hebbian Learning. Hebbian learning takes place between the input and hidden neurons, and between the hidden and output neurons:

$$\Delta W_{cd} = \epsilon \cdot \text{VTA} \cdot (s_c y_d - s_c W_{cd}) \tag{4}$$

where ϵ is the learning rate and VTA models the effect of DAergic neurons in transiently promoting synaptic plasticity. The VTA variable takes a value determined by the reward prediction error scenario. We find these VTA values for the four reward prediction error scenarios through 4-dimensional grid search optimisation with respect to classification performance.

Thus, training takes place as follow: at presentation of an input image, the activation of input neurons propagates through the network; in exploratory trials, noise is added to the activation of hidden neurons, affecting the classification decision; when the network makes an exploitative (explorative) decision, the network does (does not) expect a reward; reward is delivered; the current trial will fall into one of four reward prediction error scenarios; the VTA variable takes one of four different values based on the reward prediction error; hebbian learning takes place with a learning rate modified by the VTA variable.

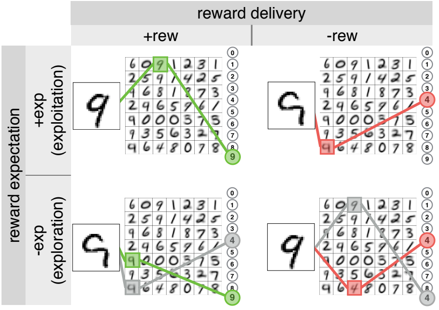

Figure 1: Four reward prediction error scenarios. In each scenario, a network is depicted with the current input, 49 hidden neurons, and 10 output neurons. Highlights indicate the most active neuron in each layer. In explorative scenarios (bottom row), grey highlights indicate neurons activated prior to noise injection (i.e., the exploitative equivalent). Each of the four scenarios leads to a distinct value of the VTA variable in equation (4).

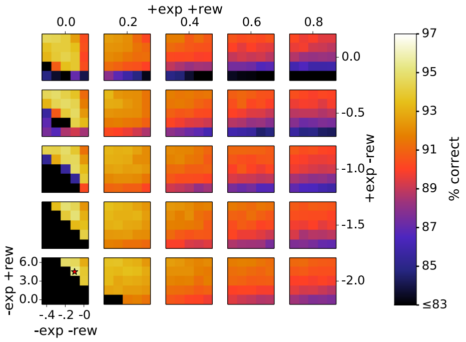

Figure 2: 4-dimensional grid search for VTA values. Optimal parameter set with respect to classification performance is indicate with a star.

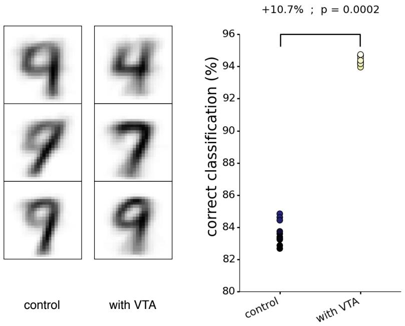

Figure 3: Changes in weights and performance. Left: examples of changes in the weights of three hidden neurons due to VTA activation. Right: increases in performance (network with 49 hidden units).

Critical Period In animals, the timeline of cortical plasticity can be broadly divided into two phases: the critical period and adulthood. The critical period is a brief phase of enhanced plasticity taking place shortly after birth. During this time, sensory representations are highly sensitive to sensory experience and rapidly adjust to the statistical structure of an animal's surrounding [14–17]. After the end of the critical period, sensory representations loose much of their adaptability [18]. Nonetheless, representations can still be modified in some specific circumstances, for instance in the case of reward-motivated learning [19]. We model these two distinct phases in our training procedure. In a first phase, akin to the critical period, learning is purely statistical: the networks exhibits no exploratory behaviour, does not receive rewards, and the VTA variable is always set to 1. During this phase, the network is essentially a clustering algorithm. In a second phase, akin to reward-motivated learning in adult amimals, the network behaves as described earlier: it exhibits exploratory behaviour, makes reward predictions, receives rewards, and the VTA is correspondingly activated. Dividing training in these two phases allows us to specifically assess the effect of reward-motivated learning on classification performance.

3 Results

We perform grid search optimisation to identify the optimal VTA values in the four reward prediction error scenarios with respect to classification performance (figure 2). We use the best identified parameters to modify the learning rate of the network during reward-mediated learning.

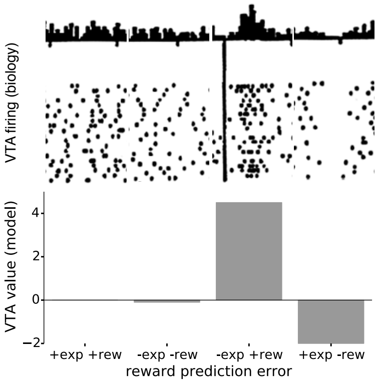

Figure 4: Optimal VTA values in the model (bottom) are in close qualitative agreement with those observed in biology (top, adapted from [1]).

We compare the weights and classification performances of the network before and after reward-mediated learning (that is, after only statistical learning and after statistical and reinforcement learning). VTA activation during rewards-motivated learning refines the network's weights with respect to the task at hand (figure 3, left). These modifications in weights in turn yield significant improvements in classification performances: in a network with 49 hidden neurons trained on a 10-class MNIST digit classification task, we obtain absolute improvements of 10.71 ± 0.07% (figure 3, right).

4 Discussion

Reward-mediated learning in our model gives rise to refinements in synaptic weights and to gains in classification performance (figure 3). These changes derive from modifications in the learning rate for individual inputs based on a reward prediction error. The optimal values of these modifications in learning rate–the values of the VTA variable in (4)–were found through 4-dimensional grid search (figure 2). The VTA values found can be divided in three categories: approximatively zero (for *-exp -rew* and *+exp +rew*), positive (for *-exp +rew*) and negative (for *+exp -rew*). These values, and more generally the learning mechanism in our algorithm, can be explained as follow.

First, the scenarios with VTA values of approximately zero are those where the network made a correct reward prediction (either *-exp -rew* or *+exp +rew*). In these cases, multiplying the learning rate by zero means that the weights remain unchanged. This can be interpreted as the network having provided a correct output and hence its behaviour should remain the same on subsequent similar trials.

Second, the positive VTA value is used when the network makes and explorative decision and receives a surprising reward (*-exp +rew*). In this case, due to noise injection, a hidden neuron other than the one most similar to the current input will win the softmax-mediated neural competition. This will lead to an exploratory classification decision. If this classification turns out to be correct, the positive multiplication of the learning rate will move the weights of the activated hidden neuron towards the current input (in the high-dimensional input space). This change in weights will make it more likely that this neuron is activated again with similar inputs in the future.

Finally, the negative VTA value is used when the networks makes an incorrect exploitative classification decision (*+exp -rew*). In this case, the negation of the learning will move the weights of the activated hidden neuron away from the current input, making it less likely that this neuron is activated by similar input on future trials.

The values found for VTA activation in the model are in close qualitative agreement with those reported in animal experiments [1] (figure 4). Our model thus provides a functional interpretation of findings from in experimental neuroscience.

training method	error rate
error back-propagation [20]	4.7%
dopamine-inspired	3.3%

Table 1: Comparison of error rates on the MNIST classification task for neural networks with 300 hidden units.

The improvements in performance we observe in our model appear similar to those reported in experiments on feedback-mediated learning. For instance, it was shown that providing performance feedback is required for human subjects to reach maximal performances on a visual orientation discrimination task [7]. In our model, classification performance following statistical training saturates after a few dozen episodes. After this point, simple stimulus presentation does not lead to performance gain. After this point, performance feedback (or rewards) need to be provided in order for further improvements to take place. Our algorithm may thus model the effect of performance feedback in discrimination learning observed in humans.

The investigated algorithm respects several biological constraints. Perhaps the most important one, however, is that synaptic plasticity relies exclusively on synaptically-local information (pre- and post-synaptic activity and local synaptic weight) as well as on a single learning signal broadcasted to all synapses (the VTA variable). This signal models DA release in sensory cortices by projections of the VTA.

Finally, our algorithm is sound from a functional perspective and compares favourably with error back-propagation-based neural networks of similar architecture (table 1). In a network with 300 hidden neurons trained on the MNIST dataset, we achieve a mean error rate of 3.31 ± 0.04%, with lowest error rate of 3.05%. In comparison, a similar network with 300 hidden neurons trained through error back-propagation achieves an error rate of 4.7% [20]. It should be noted, however, that significantly better results on the MNIST classification task can be obtained with other algorithms, such as SVMs, or with neural networks containing more hidden layers, for instance in the case of convolutional neural networks [20]. At the moment, our network is made of a single hidden layer; future work will be necessary to adapt the algorithm to support the use of multiple hidden layers.

Further future work will take two directions. From the perspective of neuroscience, we wish to test the network with stimuli used in animal experiments (e.g., orientation discrimination of Gabor filters) so that modifications in synaptic weights and in receptive fields can be directly compared with those reported in animal studies. From the perspective of machine learning, we plan to extend our work to more traditional reinforcement learning scenarios with continuous-valued and temporally-delayed rewards (for instance, the back-gammon framework).

References

1. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275,** 1593–1599 (1997).
2. Bao, S., Chan, V. T. & Merzenich, M. M. Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature* **412,** 79–83 (2001).
3. Kudoh, M. & Shibuki, K. Sound sequence discrimination learning motivated by reward requires dopaminergic D2 receptor activation in the rat auditory cortex. *Learning & Memory* **13,** 690–698 (2006).
4. Keck, C., Savin, C. & Lücke, J. Feedforward Inhibition and Synaptic Scaling–Two Sides of the Same Coin? *PLoS Comput. Biol.* **8,** e1002432 (2012).
5. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518,** 529–533 (2015).
6. Branavan, S., Silver, D. & Barzilay, R. *Non-linear monte-carlo search in civilization ii* in *AAAI Press/International Joint Conferences on Artificial Intelligence* (2011).
7. Seitz, A. R., Nanez, J. E., Holloway, S., Tsushima, Y. & Watanabe, T. Two cases requiring external reinforcement in perceptual learning. *Journal of vision* **6,** 9 (2006).
8. Pi, H.-J. *et al.* Cortical interneurons that specialize in disinhibitory control. *Nature* **503,** 521–524 (2013).
9. Letzkus, J. J. *et al.* A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* **480,** 331–335 (2011).

10. Bissière, S., Humeau, Y. & Lüthi, A. Dopamine gates LTP induction in lateral amygdala by suppressing feedforward inhibition. *Nature neuroscience* **6,** 587–592 (2003).

11. Blond, O., Crépel, F. & Otani, S. Long-term potentiation in rat prefrontal slices facilitated by phased application of dopamine. *European journal of pharmacology* **438,** 115–116 (2002).

12. Bao, S., Chan, V. T., Zhang, L. I. & Merzenich, M. M. Suppression of cortical representation through backward conditioning. *Proceedings of the National Academy of Sciences* **100,** 1405–1408 (2003).

13. LeCun, Y. & Cortes, C. The MNIST database of handwritten digits. <`http://yann.lecun.com/exdb/mnist/`>.

14. Sengpiel, F., Stawinski, P. & Bonhoeffer, T. Influence of experience on orientation maps in cat visual cortex. *Nat. Neurosci.* **2,** 727–732 (1999).

15. De Villers-Sidani, E., Chang, E. F., Bao, S. & Merzenich, M. M. Critical period window for spectral tuning defined in the primary auditory cortex (A1) in the rat. *J. Neurosci.* **27,** 180–189 (2007).

16. Han, Y. K., Köver, H., Insanally, M. N., Semerdjian, J. H. & Bao, S. Early experience impairs perceptual discrimination. *Nat. Neurosci.* **10,** 1191–1197 (2007).

17. Barkat, T. R., Polley, D. B. & Hensch, T. K. A critical period for auditory thalamocortical connectivity. *Nat. Neurosci.* **14,** 1189–1194 (2011).

18. Hensch, T. K. Critical period regulation. *Annu. Rev. Neurosci.* **27,** 549–579 (2004).

19. Recanzone, G. a., Schreiner, C. & Merzenich, M. M. Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *J. Neurosci.* **13,** 87–103 (1993).

20. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86,** 2278–2324 (1998).