Nonlinear neural simulation of error-driven learning

Jennifer Senta, Sarah Oh, Milena Rmus

December 14, 2022

1 Introduction

Over the past several decades, the field of reinforcement learning (RL) has revolutionized research in computer science, neuroscience and cognitive science. Different families of RL algorithms that capture reward-based learning have contributed to better understanding of psychological phenomena (e.g., goal-directed decision-making, (Daw, Gershman, Seymour, Dayan, & Dolan, 2011)), neural mechanisms (e.g., dopaminergic neurons, (Schultz, Dayan, & Montague, 1997)) and development of artificial models/networks trained to solve a variety of problems (e.g., action selection in grid worlds (McGovern & Barto, 2001)).

Central to all such models is the idea that positive reinforcement of a choice strengthens the probability of that choice being made in the future, i.e., that some forms of learning occur through experienced reinforcement. In particular, RL models posit that an agent holds an expectation of an outcome when making a decision, and the difference between this expected outcome and the actual outcome, referred to as the "reward prediction error" (RPE), drives learning related to the value of the choice (Sutton & Barto, 1998). However, despite the success of RL algorithms in behavioral modeling and artificial intelligence domains, it remains unclear to what extent such algorithms are mechanistically realistic as accounts of human learning. In this project, we explore how alternative modeling of RL algorithms can illuminate potential biological implementations of RL in the brain, taking into account known properties of neural communication and synaptic plasticity.

1.1 Biological realism of Reinforcement Learning

As interest in the application and development of RL algorithms has increased, cognitive scientists and neuroscientists have sought to understand whether such algorithms are biologically plausible mechanistic implementations in brain-based decision-making. Research in humans has shown that RL models can accurately model and reproduce human learning and decision-making behavior (Daw et al., 2011), but testing the implementation of the algorithm at the level of individual neurons and synapses is highly challenging. In non-human primate research, some direct neural level recording is possible, and at least one key biological instantiation of RL in the brain has been found in dopaminergic neurons of the ventral tegmental area (VTA). Specifically, the dopaminergic firing of neurons in the VTA has been found to closely track RPEs, which quantify the discrepancy between expected and observed outcomes ((Schultz et al., 1997)). The RPE signals of these neurons serve as teaching signals during the complex process of biological learning. The firing of the dopaminergic neurons involved in learning induces neural plasticity, which incrementally increases/reduces the internal value representations of choices that yield better/worse than expected outcomes. This was a major case for the biological realism of RL, and more recent results in neuroscience have further extended this finding to provide a more detailed picture of possible biological mechanisms and formalism in RL (Lowet, Zheng, Matias, Drugowitsch, & Uchida, 2020).

1.2 Motivation and approach

Despite the evidence of dopaminergic tracking of RL-consistent RPEs, and the idea that neural plasticity may support brain-based learning in a manner consistent with the RL framework, most extant neural network models of RL have leaned heavily on computational methods that leverage current state-of-the-art advances in computational technology, using implementations which do not retain the feature of biological plausibility (Williams, 1988; Bakker, 2007). For instance, many current neural network RL models rely on optimization algorithms such as gradient descent and backpropagation (Rumelhart, Hinton, & Williams, 1986), which are not biologically realistic. Meanwhile, development of artificial neural network models of RL that leverage neurobiological mechanisms including action potentials and synaptic transmission that govern local and global reward signals have stalled in comparison.

In our approach, we investigate the implementation of a neural network to model RL in a biologically realistic way using a global reward signal which learns without requiring gradient descent methods (Seung, 2003; Vasilaki, Frémaux, Urbanczik, Senn, & Gerstner, 2009). Our network uses a combination of a spiking neural network (SNN) and stochastic synaptic transmission inspired by original work from Seung (2003). Interneurons are modeled with one hidden layer and one output layer of leaky integrate-and-fire neurons. Learning takes place at the level of local synaptic release probability, modulated hedonistically by a global network reward signal which may be biologically characterized as a dispersion of dopaminergic reward signal in the brain.

2 Methods

2.1 Task: Exclusive-or (XOR)

The exclusive-or (XOR) task is a classic prediction problem used broadly in training of artificial neural networks. Specifically, the network received binary input (consisting of ones and zeros) and made a single output prediction (0 or 1). The network should return true (or 1) if the binary inputs are different; it should return false (or 0) if the binary inputs are the same. The XOR problem is frequently used as a benchmark in testing the robustness of biologically plausible (and other) neural networks, since simple perceptrons cannot solve the task, and instead require multiple layers (Minsky & Papert, 1969) given the fundamentally non-linear form of the task.

Input patterns	Correct output
0,0	0
0,1	1
1,0	1
1,1	0

Table 1: XOR logic.

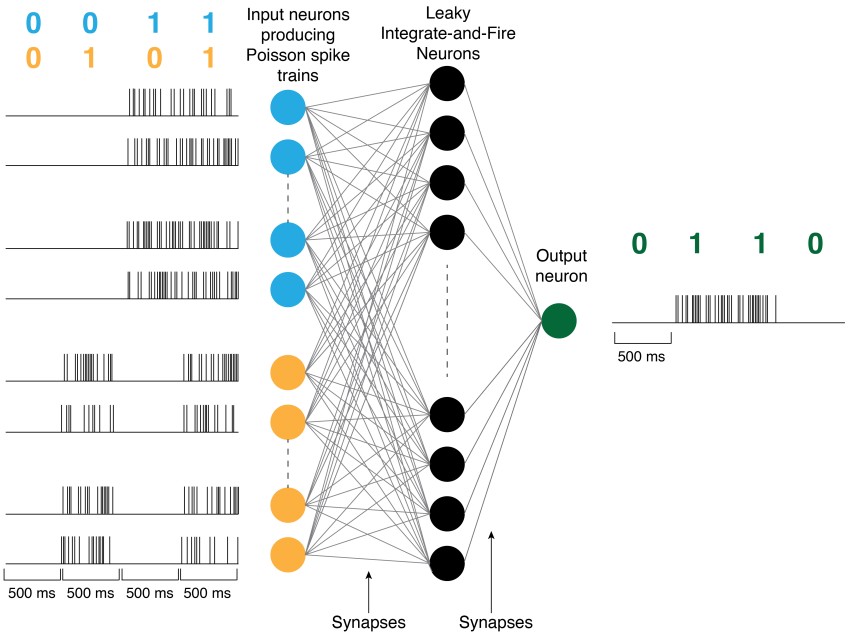

Figure 1: Structure of the neural network. First layer was the input layer, consisting of spiking neurons, which generated Poisson spike trains in response to the XOR binary patterns. Second layer consisted of LIF neurons. The connections between the neurons were modeled as stochastic hedonistic synapses which support learning. Single output neuron encoded network's prediction to input patterns. Only a subset of neurons are depicted for simplicity.

2.2 Neural network model

2.2.1 Structure

Our network consisted of three layers in total. Inputs to the network were delivered by 60 input neurons which produced Poisson spike trains in response to presented data. The resulting spike trains were densely connected

to a hidden layer of 60 interneurons, each modeled using leaky integrate-and-fire dynamics. Synapses at every connection (between every input neuron and every hidden neuron) possessed a unique probability of release and synaptic conductance strength, which together governed the transmission of signal from pre- to post-synaptic neurons. The hidden interneurons were themselves each synaptically connected to a single output leaky integrate-and-fire neuron, the spiking of which determined the final output for the network at a given timestep. (Fig: 1). Each neuron j was randomly set to be either excitatory (excitatory outgoing synapses with reversal potential $V_{ij} = 0mV$), or inhibitory (inhibitory outgoing synapses with reversal potential $V_{ij} = -70mV$).

2.3 Training

Each binary XOR input pattern (00, 01, 10, 11) was presented in a sequence for 500ms, for a minimum of 100 training iterations.

2.3.1 Dynamics

Poisson spike train neurons. Half of the input layer neurons received one sequence of zeros and ones (blue neurons/sequence in Fig. 1), and the second half of the neurons received a second sequence of zeros and ones (orange neurons/sequence in Fig. 1) forming one of the 4 XOR patterns on each presentation. The input neurons generated Poisson spike trains at a rate of 40 Hz if the value of 1 was present in the incoming stimulus:

$$\mathrm{p}_{\mathrm{spike(t)}} = \delta_t * 40\mathrm{Hz}$$

$$\mathrm{Spike}_t = \begin{cases} \mathrm{binomial}(\mathrm{p}_{\mathrm{spike(t)}}) & \text{if input digit} == 1, \\ 0 & \text{if input digit} == 0 \end{cases}$$

LIF neurons. The hidden layer interneurons were each modeled using leaky integrate-and-fire dynamics as detailed below. We simulated the time-step evolution of the dynamics using the Euler approximation method (Biswas, Chatterjee, Mukherjee, & Pal, 2013) with $\delta_t = 0.5ms$.

$$C\frac{dV_i}{dt} = -gL(V_i - V_L) - \sum G_{ij,t}(V_i - V_{ij}) + I_{tonic} \tag{1}$$

where the conductance $G_{ij,t}$ of synapse ij controlled the flow of charge from neuron j to neuron i. V_{ij} was the reversal potential of synapse ij ($V_{ij} = 0mV$ for excitatory, $-70mV$ for inhibitory). I_{tonic} was sampled from a normal distribution with mean 425 pA and standard deviation 200 pA, to simulate noisy inputs from outside the network.

When the membrane potential reached the threshold level, the interneuron spiked and entered the refractory period during which the voltage was held at a constant level. We used the following parameters in our network simulations: resting state voltage ($V_L = -74mV$), leak conductance ($g_L = 25nS$), membrane capacitance ($C = 500pF$), spiking threshold ($V_\theta = -54mV$); voltage reset after refractory period ($V_{reset} = -60mV$), refractory period time ($t = 0.001s$).

Synapse. Upon receiving spikes from pre-synaptic neurons, a synapse would release its vesicle with probability p:

$$p = \frac{1}{1 + e^{-q}} \tag{2}$$

where q was defined as the release parameter, which was updated via learning. The release probability p sigmoidally mapped the release parameter into a probability domain of [0,1].

Learning. To ensure that the network optimized vesicle releases/failures in a way that maximized the reward, each synapse maintained a recent history of its selected choices via a metric referred to as an eligibility trace $\bar{e}$ (Klopf, 1982). The eligibility trace was initialized to zero, and adjusted its value upwards with each vesicle release, and downward with each vesicle failure after input spike. When no pre-synaptic spike was received by the synapse, the eligibility trace $\bar{\mathrm{e}}$ decayed exponentially, with a time constant parameter $\tau_\mathrm{e} = 20\mathrm{ms}$ controlling the amount of time a synapse held the trace of its past releases/failures:

$$\frac{d\bar{\mathrm{e}}}{dt} = -\frac{\bar{\mathrm{e}}}{\tau_e} \tag{3}$$

The synaptic conductances G_{ij} also decayed exponentially with time constant $\tau_s = 5$ms in the absence of pre-synaptic spikes.

$$\frac{dG_{ij}}{dt} = -\frac{G_{ij}}{\tau_s} \tag{4}$$

When pre-synaptic spikes occurred, the change in $\bar{e}$ was given by:

$$\Delta\bar{e} = \begin{cases} 1-p & \text{if release,} \\ -p & \text{if failure} \end{cases} \tag{5}$$

The change in G_{ij} in response to pre-synaptic spikes was given by:

$$\Delta G_{ij} = \begin{cases} W_{ij} & \text{if release,} \\ 0 & \text{if failure} \end{cases} \tag{6}$$

where the (static) conductance amplitude parameters W_{ij} were selected from exponential distributions with mean 2.4 nS for excitatory synapses, or 45 nS for inhibitory synapses.

The global reward signal $h(t)$ was defined as the spike train of the output neuron, or its inverse, depending upon the input. Specifically, h(t) = +1 when the output neuron fired at input (0,1) or (1,0), and 0 if it did not fire at these inputs. Conversely, h(t) = -1 if the output neuron fired at input (0,0) or (1,1), and 0 if it did not fire at these inputs.

Synaptic learning via probability of vesicle release was modulated by the product of $\bar{e}$ and reward signal h:

$$\Delta q = \eta h_t \bar{e}_t \tag{7}$$

where η is the learning rate (set to 0.3).

3 Results

The results of the network after training for 100 epochs are illustrated in Fig 2. Each synapse maintained a record of actions via an eligibility trace, which in combination with the global reward signal for the network modulated the rate of vesicle release for that synapse. The synaptic conductance strength was also modulated by the value of release or failure for a given synapse, resulting in stochastically variable vesicle releases across time with greater or lesser strength of communication with adjoining neurons.

After random initialization, the network initially fired to various correct and incorrect input sequences (Fig: 2A). The global reward signal $h(t)$ was defined at each timestep in accordance with the actual single spiking output of the network, and propagated locally via direct synaptic learning updates triggered by the unadjusted reward signal. This form of learning has sometimes been referred to as "hedonistic" learning (Klopf, 1982), as each synapse directly used a single global reward signal to individually modulate its own actions, resulting in convergence of the network as a whole to correct learned behavior over time.

After at least 100 training epochs, the network converged to fire an output signal only in response to the correct input sequences (0,1) or (1,0), and suppresses spiking for all other inputs (Fig: 2B,C).

4 Discussion

These results provide just one example of a neural network which uses biologically realistic dynamics to successfully learn to perform an XOR task. In reality, the brain dynamics of input to such a learning network, and indeed the behavior of the neurons themselves, are much more complex than the model shown here. Nonetheless, the global reward signal used for learning here is compatible with a release of dopamine broadly affecting a subset of neurons, thereby inducing synaptic plasticity and acting as a teaching signal locally without need for gradient descent or other computationally intensive optimization methods which are not biologically plausible.

Although using a leaky integrate-and-fire model of each interneuron allows us to reasonably approximate broad neural behavior without excessive complexity, a more realistic neural model (e.g. Hodgkin-Huxley, (Hodgkin & Huxley, 1952), Izhikevich model (Izhikevich, 2003)) might provide more accurate timing benchmarks of biochemical updating dynamics which could be compared directly to actual biological implementation.

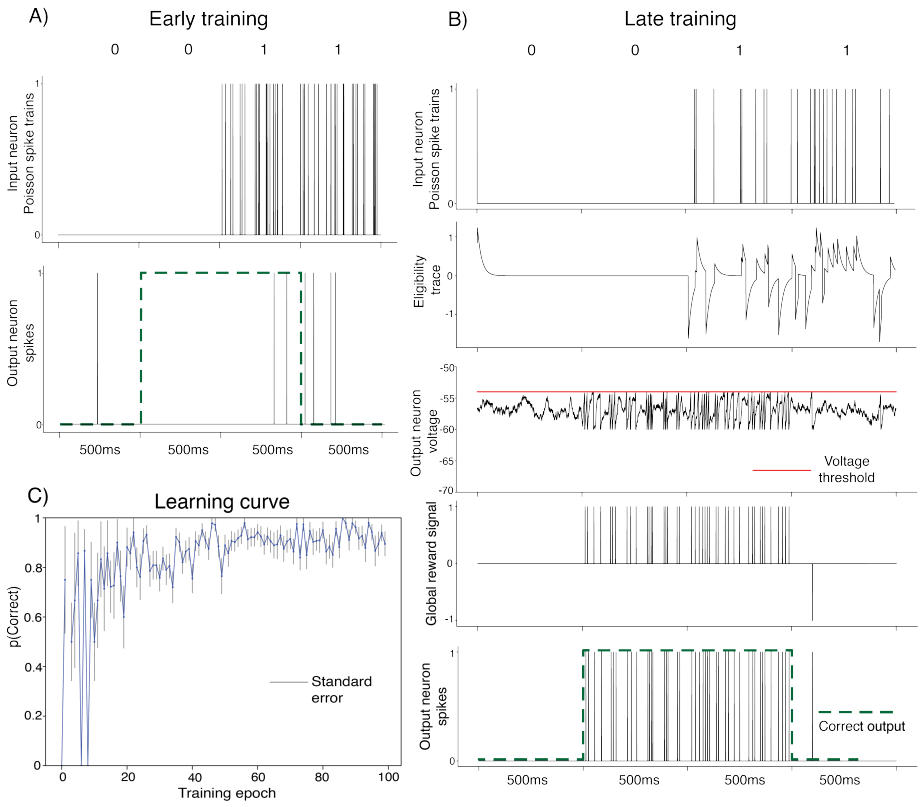

Figure 2: Results of network. A) Before giving the network a chance to learn, the network incorrectly produced spikes in response to input patterns. B) Network results at training epoch $n = 100$. Top row depicts input spike trains. Bottom row depicts network output signal after learning. Intermediate rows illustrate network dynamics of a sample neuron. Both A) and B) depict the same neuron early and later in training; this is a neuron sampled from the blue population in Fig 1. C) Over time, the network learned to reliably generate spikes when input pattern was (0,1) or (1,0), and suppress spiking when input pattern was (1,1).

One potential enhancement to the model presented here would be to further modulate synaptic plasticity using additional biological mechanisms. The stochasticity of vesicle release in the current model is modulated by the global reward signal and the eligibility trace, but other induced synaptic plasticity might also occur. For example, some literature has used calcium and other ion dynamics to control separate components of learning (Seung, 2003; Vasilaki et al., 2009), which could be an improvement here in terms of biological realism, and might enable learning of more robust task structure. Although we attempted to implement this alteration to the model, we weren't able to get this mechanism working prior to write-up.

The XOR task used here provides a well-established, valid test for an effective nonlinear reinforcement learning network. Another similar alternative would be a two-armed bandit task, wherein the network must learn the value of two choices via probabilistic reinforcement over time. Such a task is commonly used in cognitive science testing of human behavior, and our network would be similarly capable of learning this structure. However, the network as developed here would not be capable of learning more advanced sequential tasks without the incorporation of recurrent neural network connections. Current cutting-edge models of such learning might employ multiple layers and levels of recurrences within a single model, which would be both biologically plausible and capable of more robust learning.

Although it may not be the case that the algorithms used for human-created artificial intelligence will ever be truly similar to the human biology of learning, it seems clear that there are at least some aspects of commonly used RL algorithms being implemented in the brain. Here we have shown one way in which such an algorithm could be optimized via biologically plausible mechanisms like manipulation of synaptic plasticity. As the line of scientific inquiry seeking to expand learning capability in machines has seen robust growth, so too should the inquiry into the sophisticated learning mechanisms of the human brain.

References

Bakker, B. (2007). Reinforcement learning by backpropagation through an lstm model/critic. In *2007 ieee international symposium on approximate dynamic programming and reinforcement learning* (pp. 127–134).

Biswas, B., Chatterjee, S., Mukherjee, S., & Pal, S. (2013). A discussion on euler method: A review. *Electronic Journal of Mathematical Analysis and Applications*, *1*(2), 2090–2792.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, *117*(4), 500.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Transactions on neural networks*, *14*(6), 1569–1572.

Klopf, A. H. (1982). *The hedonistic neuron: A theory of memory, learning, and intelligence*. Hemisphere Pub.

Lowet, A. S., Zheng, Q., Matias, S., Drugowitsch, J., & Uchida, N. (2020). Distributional reinforcement learning in the brain. *Trends in Neurosciences*, *43*(12), 980–997.

McGovern, A., & Barto, A. G. (2001). Automatic discovery of subgoals in reinforcement learning using diverse density.

Minsky, M., & Papert, S. (1969). An introduction to computational geometry. *Cambridge tiass., HIT*, *479*, 480.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, *40*(6), 1063–1073.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: an introduction mit press. *Cambridge, MA*, *22447*.

Vasilaki, E., Frémaux, N., Urbanczik, R., Senn, W., & Gerstner, W. (2009). Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. *PLoS computational biology*, *5*(12), e1000586.

Williams, R. J. (1988). On the use of backpropagation in associative reinforcement learning. In *Icnn* (pp. 263–270).