Pair-Associate Learning with Modulated Spike-Time Dependent Plasticity

Nooraini Yusoff, André Grüning, and Scott Notley

Department of Computing, Faculty of Engineering and Physical Sciences,
University of Surrey, Guildford, Surrey GU2 7XH, UK
{n.yusoff,a.gruning,s.notley}@surrey.ac.uk
http://www.surrey.ac.uk

Abstract. We propose an associative learning model using reward modulated spike-time dependent plasticity in reinforcement learning paradigm. The task of learning is to associate a stimulus pair, known as the *predictor – choice* pair, to a target response. In our model, a generic architecture of neural network has been used, with minimal assumption about the network dynamics. We demonstrate that stimulus-stimulus-response association can be implemented in a stochastic way within a noisy setting. The network has rich dynamics resulting from its recurrent connectivity and background activity. The algorithm can learn temporal sequence detection and solve temporal XOR problem.

Keywords: Spiking neural networks, Associative learning, Spike-time dependent plasticity, Reinforcement learning.

1 Introduction

Numerous experimental findings have emphasised the importance of temporal correlations between pre- and postsynaptic spikes on the efficacy of synaptic changes. The Hebbian-based temporal synaptic plasticity known as spike-time dependent plasticity (STDP) essentially says that a synapse is strengthened if a presynaptic neuron fires before its postsynaptic neuron and is supressed if the presynaptic neuron fires after the postsynaptic neuron [1], [9]. There are so-called *third signals* (for example neurotransmitter concentrations) that are used as mediators relating the synaptic plasticity mechanism at the cellular level and its contribution to the adaptive changes at the behavioural level [2], [3]. Dopamine (DA) has been identified as one such signal, and plays a role in reward acquisition mechanisms [4]. It has been found that DA contributes to enhancing the long-term potentiation (LTP) and long-term depression (LTD) of synapses [3]. The release of the dopamine causes an increase in the delivery of one of the protein subunits to the cell membrane, consequently enhancing responsiveness to other neurotransmitters.

In the context of STDP based learning, the causal relationships between pre- and postsynaptic neurons are reinforced only when there is a reward. By having the selective synapse reinforcement, potentiation leads to reduced variability of

A.E.P. Villa et al. (Eds.): ICANN 2012, Part I, LNCS 7552, pp. 137–144, 2012.

the output and depression through negative reward leads to increased variability of the networks behaviour [5]. Among popular works of learning with reward modulated STDP is [6]. In the original reinforcement learning proposed by [6], learning only involves association of a stimulus to a response group. The network has fixed synaptic transmission delay of 1 ms. The experiment demonstrated plausible and promising result in learning with modulated STDP implementation. The work has proven that, reinforcement signal known as the dopamine signal can selectively enhance synaptic changes proposed by the standard STDP rule. In the reported experiment, reinforcement to different target response is implemented in batch. Initially the network was rewarded for the first response group, then after successive trials, the reward was changed to the second group. This somehow offers a challenge for learning with multiple input-output mappings with correlated spike train as competition between outputs is higher.

Inspired by the algorithm proposed in [6], we explore the ability of reward modulated STDP in tasks that require stimulus-stimulus-response association. Reinforcement of paired stimuli (i.e. $predictor-choice$ pair) to a target response is based on a reward signal derived from a reward policy whose parameter is the firing rate of a response group. The reward signal, modelled after the role of dopamine signal in the brain, enhances the amount of potentiantion (or depression) caused by STDP. We expand Izhikevich's experiment [6] by presenting input (i.e. stimulus pair) to a network randomly in a system with multiple outputs. We also implement depression of synaptic weights through negative rewarding and network with synaptic transmission delay parameter that provides more richer temporal dynamics. The results reveal the practicality of our learning rule in training a stochastic network to associate delayed paired stimuli with a response in tasks with multiple input-output mappings. Furthermore, the network can also learn temporal sequences within appropriate range of ISI.

2 Neural Network Dynamics

The proposed network model is a recurrent spiking network consisting of 800 excitatory and 200 inhibitory spiking neurons. The connectivity between neurons is random and sparse. Each synaptic connection, from neuron i to j, is defined by two parameters: a weight w_{ij} and a synaptic transmission delay d_{ij}. In our model, the delay is a random integer between 1 to 20 ms. Neurons are divided into subpopulations of stimulus groups (S), response groups (R), non-selective neurons (NS) and inhibitory pool (IH). For clarity in discussions, we divide the network into two modules; Response module and Input module (see Fig. 1).

In the Response module, each excitatory response group, e.g. $R{+}_m$, is connected to its inhibitory pool, e.g. $R{-}_m$. The inhibitory pool provides inhibition to its competitor group(s) through negative synaptic connections. The synaptic strength from an inhibitory pool of a response group to excitatory neurons in its competitor is set to -4.0 (not plastic). Each excitatory neuron in the response module has 50% of postsynaptic neurons from its inhibitory pool, and 50% of postsynaptic neurons consisting of neurons from the same excitatory response group and/or excitatory neurons in the input module.

For synaptic connections in the Input module, each excitatory neuron has random connections to 100 neurons from the whole population (of 1000 neurons, $p = 0.1$), and each inhibitory neuron in this module is connected to 100 excitatory neurons from the whole population. For all experiments described in this paper, the number of neurons in each $R+$, $R-$ and $S+$ is 100, 50 and 50 respectively.

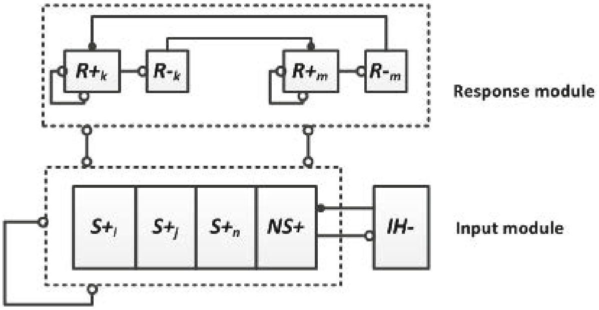

Fig. 1. Recurrent spiking network with subpopulations of stimulus groups (S), response groups (R: R_+ and R_-), non-selective neurons (NS) and inhibitory pool (IH). Lines end with open circle show excitatory connection, and lines end with solid circle indicate inhibitory connection. (Please see text for details).

2.1 The Spiking Network Model

The spiking properties of each neuron are modelled as in 1-3 as proposed by [7], [8]:

$$v' = 0.04v^2 + 5v + 140 - u + I \tag{1}$$

$$u' = a(bv - u) \tag{2}$$

$$if\ v \geq +30\ mV,\ then\ u \leftarrow u + d, v \leftarrow c\ . \tag{3}$$

where v, u, and I describe the neuron membrane potential, the recovery variable, and the input current (and the synaptic input), respectively, while a-d are the model parameters.

After the spike reaches its peak v_{peak}= +30 mV, the membrane voltage and the recovery variable are reset according to (3). For learning initialisation, the membrane potential, v is set to -60.0 mV. The value is above the resting potential, c = -65.0 mV, that assumes some initial activity prior to learning. In our model, like in Izhikevich's original, all excitatory neurons are regular spiking (RS) type and all inhibitory are fast spiking (FS) type neurons (details of neuron spiking properties can be found in [8]).

3 Synaptic Plasticity

Following the standard STDP rule as suggested in [6], for each fired neuron, we find the last spike timing of each of its presynaptic neurons. The synaptic efficacy is reinforced if the presynaptic neuron fires before its postsynaptic (+ve part of the STDP curve), and depressed otherwise (-ve part of the STDP curve). The magnitude of change is given by the following rule (4):

$$\Delta w_{stdp} = \begin{cases} A_{+} e^{\frac{-\Delta t}{\tau+}} & \text{if } \Delta t \geq 0 \\ A_{-} e^{\frac{\Delta t}{\tau-}} & \text{if } \Delta t < 0 \end{cases} \tag{4}$$

where $\Delta t = t_{post} - t_{pre}$, parameters τ_{+} (τ_{-}) are the millisecond-scale time constants, and A_{+} (A_{-}) represents the maximum of the change, Δw_{stdp} , when Δt is approaching 0. In our model the choice of the parameters is as follows: $\tau_{+} = \tau_{-} = 20$ ms, $A_{+} = 0.1$, and $A_{-} = 0.15$.

For every time step of 10 ms, weight update is applied to excitatory-excitatory and excitatory-inhibitory synapses whilst inhibitory-to-excitatory synapses are kept fixed. The weight update rule [5], [6] holds:

$$\Delta w(t) = [\alpha + r(t)]\, z(t) \ . \tag{5}$$

The change of the synaptic weight $\Delta w(t)$ is dependent on a reward signal $r(t)$, derived from (6), and an eligibility trace $z(t)$, where $z_{ij}(t)$ is the sum of weight changes w_{ij} of presynaptic neuron i to postsynaptic neuron j, proposed by STDP. α is an activity-independent increase of synaptic weight. Assuming F_i as the intensity of firings of a desired response group R_i, in a time interval, and F_j is the highest firing rate of non-target groups, $i \neq j$, the derivation of reinforcement signal $r(t)$ from the reward policy $\Theta(F)$, is given by:

$$\Theta(F) = r(t) = \begin{cases} r(t-1) + 0.5 & \text{if } F_i \geq 2F_j \\ 1 - F_j/F_i & \text{if } F_j < F_i < 2F_j \\ -0.1 & \text{if } F_i < F_j \end{cases} \tag{6}$$

Every millisecond, r decreases by $0.995 * r$, and for every synaptic weight update, z decreases by $0.99 * z$. To avoid infinite growth of weights and change of weight sign, weights are kept to be in the range between 0 to 4 mV. [1]

3.1 Learning Protocols

From the population of 1000 neurons, we select n non-overlapped groups, S_n, of 50 excitatory neurons each. Each group represents a stimulus. Another m exclusive groups, R_m, of 100 excitatory neurons each are selected as response groups. In a learning simulation with a number of trials, in the first 100-ms window

[1] In Izhikevich paper a incoming single spike over a weight w leads to an increase of the membrane potential of w mV –hence units of connection strength are measured in mV.

time, the network only experiences background activity that we stimulate an arbitrary neuron with 20 pA current (super-threshold current). For each learning trial, in the presence of the background activity, we randomly select a stimulus pair, e.g. $predictor = S_0$ and $choice = S_1$, with a target response, e.g. A and stimulate all 50 neurons within each stimulus with a super-threshold current, i.e. 20 pA at time t_n and t_{n+ISI} for a *predictor* S_i and *choice* S_j, respectively. The inter-stimulus interval (ISI) is an experimental parameter in the range of 10 - 50 ms. From the onset of a *choice* pattern, within a 20-ms time window, we count the number of spikes in the response groups, A and B. The response group with highest number of spikes is considered as the winner. The next trial starts after a delay of 100 ms after the offset of each response interval. We reward the network based on the number of spikes in A and B within the 20-ms interval following the reward policy in 6. Every learning task is repeated with 10 different network simulations that each simulation takes 20 mins.

There are 2 phases of synapse reinforcement. In the first phase, a reward signal is produced for the number of spikes in the target response inhibitory group within 10 ms of the response interval from the onset of a *choice*. This is to strengthen the synapses from a triggered stimulus pair to its postsynaptic neurons in the target response inhibitory group for lateral inhibition to its competitor group(s). In the second phase, the same mechanism is applied for reinforcement of the excitatory response groups but based on the number of spikes within 20 ms. In addition, winner-take-all strategy is implemented in both phases through biased random excitatory signals to the winner of response groups for each phase, if the winner = target response. The training performance is computed based on the percentage of number of correct response over number of trials averaged by 10 simulations (i.e. 10 different networks).

For a test phase, we run a simulation consisting of a number of trials for 200 ms each. The testing result shows the average of performance (i.e. correct recall rate of learned pairs) over 100 trials. For every trial, the network with the same background activity as during the training gets stimulated with a superthreshold current of 20 pA applied onto the tested *predictor* at some random time, t, in between 100-120 ms. The stimulation on *choice* group proceeds after the *predictor* group depending on the ISI. The number of spike counts within the 20-ms response interval (starts from the onset of the *choice*) is used to compute a winning response.

4 Simulation Results

We have run a series of simulations with various $predictor - choice$ pairing strategies. We began training a network with exclusive stimulus groups, i.e. $Pair - Response = \{(S_0, S_1) \rightarrow A, (S_2, S_3) \rightarrow B, (S_4, S_5) \rightarrow A, (S_6, S_7) \rightarrow B\}$ with ISI=10 ms. The averaged performance was achieved with correct recall rate of 94.08% and 99.75%, for training and testing respectively.

For network stimulation, we delivered a 1-ms pulse super threshold current to all neurons in the selected groups so that each of them fired almost immediately. At the early phase of learning, in addition to the background activity, the activation of neurons was only due to coincident firings evoked by those stimulated neurons. By frequently stimulating *predictor* – *choice* pairs, and firings that follow the pre-then-post order rule of the STDP, the synaptic connections from those 50 neurons in the paired groups to the fired postsynaptic neurons become eligible for potentiation (see Fig. 2). When there is no reward, i.e. DA = 0.0, after some period of time, the eligibility trace decays to zero, resulting in only small potentiation. In such case, as the LTD window of STDP is greater than the LTP window, the amount of the potentiation is compensated by the STDP depression mechanism. On the other hand, if there is a reward, i.e. increment (decrement) of DA value, the amount of potentiation (depression) can be enhanced. Therefore, rewarding mechanism based on conditional response reinforces connections to a target response group, A or B.

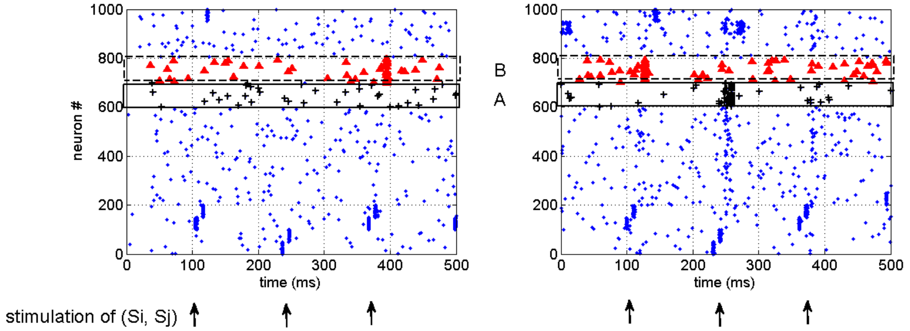

Fig. 2. An example of spike raster plot of a learning (*left*) at the early phase, and (*right*) after 150 seconds, within 500 ms time window. Neurons in the response group A (# 501-600) and group B (# 601-700) are encapsulated in the solid and dashed boxes, respectively.

4.1 Learning Temporal Sequence

In the following experiments, we investigated the non-exclusivity of pattern pairs. There were stimulus groups sharing the same *predictor* or *choice* with conflicting responses. In such condition, there were unstable patterns that could be dragged to the undesired attractor, e.g., $(S_0, S_1) \rightarrow A$ and $(S_0, S_2) \rightarrow B$. To reduce too high correlation in neural spike trains, the same stimulus group, say a *predictor*, that was to be paired with different *choices* was allowed to have some probability of non-overlapping neurons. For brevity, in group S_0 consisting of 100 neurons, 50 neurons were selected randomly to be paired with 50 neurons from group S_1 (out of 100 neurons, chosen randomly). Hence for two stimulus pairs, e.g. $(S_0, S_1) \rightarrow A$ and $(S_0, S_2) \rightarrow B$, the *predictor* S_0 may have a number of overlapping neurons with some probability.

In this paper, we report three experiments of learning with non-exlusive groups under three conditions. For condition with shared *predictor* in a learning set of $\{(S_0, S_1) \rightarrow A, (S_0, S_2) \rightarrow B\}$, training and testing achieved 86.56% and

85.40%, respectively. In learning with non-exclusivity and identical orthogonality, for stimulus pairs $\{(S_0, S_1) \rightarrow A, (S_0, S_2) \rightarrow B, (S_1, S_0) \rightarrow B\}$, the results were 76.99% for training and 73.73% for testing. We also trained the network under a learning condition with non-exclusivity and asymmetrical difference in a stimulus set of $\{(S_0, S_1) \rightarrow A, (S_0, S_2) \rightarrow B, (S_2, S_1) \rightarrow A\}$. The performance was higher when compared with learning under the other two conditions, 90.44% and 94.40%, for training and testing, respectively.

4.2 XOR Benchmark

Here we tested whether our learning approach could successfully learn the XOR problem. To perform a logic function task, we defined 4 distinct stimulus groups, S_0, S_1, S_2, and S_3. S_0 (S_1) and S_2 (S_3) represented the TRUE (FALSE) values of the first (second) stimulus, respectively. Meanwhile, the response group A represented a TRUE response and the response group B was considered a FALSE response. Therefore for XOR problem the pattern pairs were as follows: $Pair - Response = \{(S_0, S_2) \rightarrow B, (S_0, S_3) \rightarrow A, (S_1, S_2) \rightarrow A, (S_1, S_3) \rightarrow B\}$.

For this problem, all stimulus pairs are unstable due to non-exclusivity with shared *predictor* and *choice* having conflicting responses. This consequently may result high competition in learning. Nevertheless, with lateral inhibition mechanism in our proposed algorithm and appropriate ISI (i.e. 10 ms), simulation result indicates that a network with stochastic dynamics and minimal anatomical constraints can also learn temporal logic functions with good performance achieved at 81.88% and 79.53%, in training and testing respectively.

5 Conclusion

In this study, we demonstrate the ability of reward-modulated spike-time dependent plasticity in pair associate learning tasks. In the network with random connectivity, there are subpopulations of excitatory neurons that are selective to certain stimuli. The network is presented with a *predictor* stimuli followed by its paired *choice* with a certain inter-stimulus interval (ISI). The algorithm has been successfully tested for temporal sequence learning with exclusive stimulus groups as well as in a setting with overlap of patterns between stimulus groups. Furthermore, the algorithm has also been verifed its performance in solving the XOR problem. In learning with non-exclusive stimulus groups, greater influence from a *predictor* is required to facilitate discrimination of target responses due to correlation in neural spike trains. The optimal ISI for such learning condition has been found at 10 ms.

To serve a goal-directed learning, our proposed algorithm integrates STDP and firing rate. The firing rate is a parameter of a reward policy (6) that determines the adjustment value for synaptic changes proposed by STDP standard rule. The reward policy function derives a reinforcement signal (i.e. the adjustment value) based on firing rate of a response group. In [6], learning only applies positive reinforcement, for our case, the adjustment value represents the dopamine concentration variable that results in strong positive, weak positive or negative reward signals. Higher firing rate of the target response group yields stronger signal for synapse reinforcement. Therefore, rewarding mechanism is based on modulation by the dopamine variable, where the increment/decrement of its values enhances the potentiation or depression resulted by the STDP process. Furthermore, we propose a lateral inhibition mechanism to improve learning in a more competitive environment.

Learning is implemented with minimal assumption of the network dynamics. The network with random activity does not have any prior knowledge regarding the identity of learning signals. There is no need of so called 'teacher signals' as in instructive learning approaches. Input stimulation is induced at certain time only through perturbation to the network activity.

Acknowledgments. This research has been funded from the Ministry of Higher Education (Malaysia). Scott Notley (full) and Andr*é* Gr*ü*ning (part) were supported by EPSRC grant EP/I027831/1.

References

1. Bi, G.Q., Poo, M.M.: Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength and postsynaptic cell type. J. Neurosci. 18, 10464–10472 (1998)
2. Gu, Q.: Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. Neuroscience 111, 815–835 (2002)
3. Smith, W.B., Starck, S.R., Roberts, R.W., Schuman, E.M.: Dopaminergic Stimulation of Local Protein Synthesis Enhances Surface Expression of GluR1 and Synaptic Transmission in Hippocampal Neurons. Neuron 45, 765–779 (2005)
4. Wise, R.A.: Dopamine, learning and motivation. Nat. Rev. Neurosci. 5, 483–494 (2004)
5. Florian, R.V.: Reinforcement learning through modulation of spike-timing dependent synaptic plasticity. Neural Comput. 6, 1468–1502 (2007)
6. Izhikevich, E.M.: Solving the distal reward problem through linkage of STDP and dopamine signaling. Cereb Cortex 17, 2443–2452 (2007)
7. Izhikevich, E.M.: Simple Model of Spiking Neurons. IEEE Trans. Neural Networks 14(6), 1569–1572 (2003)
8. Izhikevich, E.M.: Polychronization: Computation with Spikes. Neural Computation 18, 245–282 (2006)
9. Gerstner, W., Kistler, W.: Spiking Neuron Models: Single Neurons, Populations, Plasticity. University Press, Cambridge (2002)