Reinforcement Learning Rules for Spiking Neurons Can Learn Spatiotemporal Activity Patterns

Nicolas Frémaux and Wulfram Gerstner, EPFL, Switzerland

INTRODUCTION. The ability to learn from the outcome of past experiences is crucial for the survival of animals. How this is implemented in neurons and synapses is less clear. Recent work in neuroscience casts some light on this question.

First, in a series of landmark experiments, Schultz and colleagues have shown that the activity of dopamine neurons match very well the reward prediction error in reinforcement learning [1], suggesting dopamine as a possible neural correlate to the δ of TD learning. Second, various experiments have shown that dopamine receptors D1/D5 modulate spike-timing synaptic plasticity [2].

These findings suggest that reinforcement learning could be implemented neurally by so-called three factor rules, where the information given by a global reward signal is combined with the local pre/post spike timing information at each synapse. Several such rules have been proposed, and used for simple tasks. However, no systematical comparison of these rules has been published so far. In this research, we propose a spatiotemporal learning task and directly compare how these rules perform.

LEARNING RULES. We compare two rules that combine spike-timing dependent plasticity [3, 4] with a global reward factor (such as dopamine or other neuromodulators). The first rule [5–7] takes a standard STDP model [3, 8] and modulates the learning rate by reward. We call this the R-STDP model. The second rule [6, 9, 10] is derived from a reward maximization approach akin to policy-gradient rules in reinforcement learning [11, 12]. We call this the R-max rule.

Both rules have the same general form: the weight change is driven by the product of a synaptic eligibility trace e_{ij} trace and a global reinforcement signal $R(t)$: $\frac{dw_{ij}}{dt} = \gamma R(t)\, e_{ij}(t)$. The eligibility trace is a filtered version of a spike timing dependent kernel $\xi_{ij}(t)$:

$$\tau_e \frac{de_{ij}}{dt} = -e_{ij} + \xi_{ij}(t)$$

The rules differ in the form of the $\xi_{ij}(t)$ term, that depends on the firing times t_j^f and t_i^f of the pre- and post-synaptic neurons, respectively. In R-STDP, the factor $\xi_{ij}(t)$ is:

$$\xi_{ij}^{\text{R-STDP}}(t) = \sum_{t_i^f} \delta(t - t_i^f) A_+ \sum_{t_j^f < t_i^f} \exp\left(\frac{t_j^f - t}{\tau_+}\right) + \sum_{t_j^f} \delta(t - t_j^f) A_- \sum_{t_i^f < t_j^f} \exp\left(\frac{t_i^f - t}{\tau_-}\right)$$

For the R-max rule, $\xi_{ij}(t)$ depends in addition on the instantaneous firing rate $\rho(t)$ of our stochastic neuron:

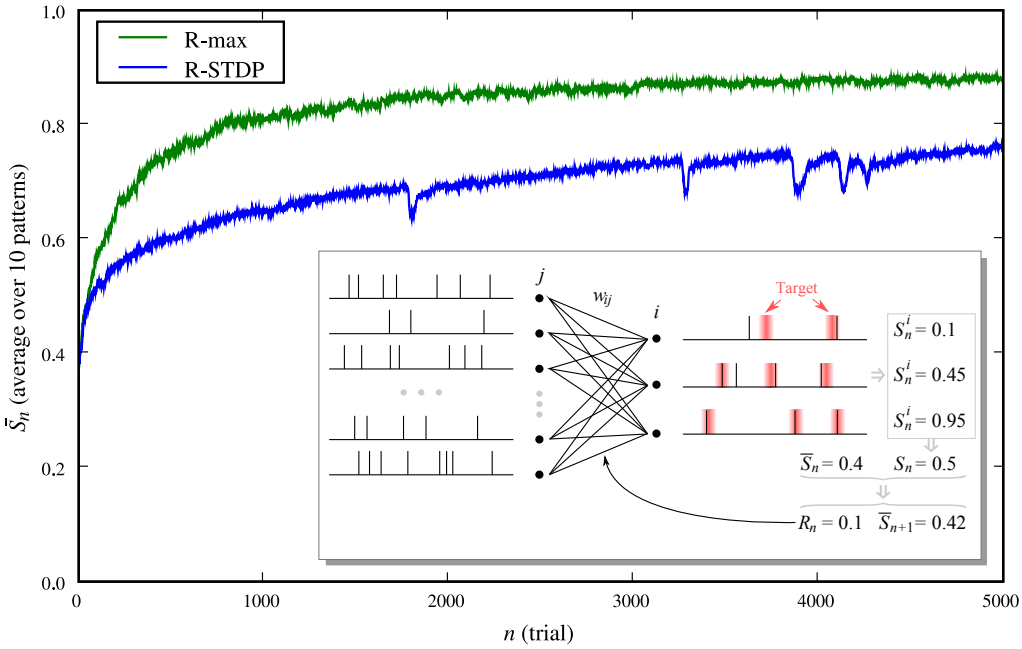
$$\xi_{ij}^{\text{R-max}}(t) = \Delta u \left[\sum_{t_i^f} \delta(t - t_i^f) - \rho_i(t)\right] \sum_{t_j^f < t} \varepsilon(t - t_j^f)$$

Here $\varepsilon(t - t_j^f)$ denotes the time course of an excitatory post-synaptic potential (EPSP).

For both learning rules, we use a spike response model (SRM) [13] as our post-synaptic neuron model. The instantaneous firing rate is then given as $\rho(t) = \rho_\theta \exp\left(\frac{u(t)-\theta}{\Delta u}\right)$, where $u(t)$ is the membrane potential.

In the limit $\Delta_u \to 0$, this becomes a deterministic neuron model, with θ as threshold.

SPIKE TRAIN LEARNING. We simulate a simple paradigm consisting of 50 input spike trains, that project, via learning synapses, to a layer of 5 noisy SRM neurons. The task is to respond to a spatiotemporal-spike pattern in the input with a target spike pattern in the output (see inset of figure). The input spike trains are generated once by homogeneous Poisson processes at 6Hz and kept fixed afterwards. The target output pattern is generated by presenting the input pattern to the network, with a set of random synaptic efficacies drawn uniformly in the interval $(0,1)$. Hence the task is to find the correct set of weights so as to reproduce the target. At the beginning



of the learning phase, all weights are set to 0.5. In each trial, the input pattern is presented, and the actual output pattern is compared with the target pattern [14]. This results in a score measure S_n. A reward $R_n = S_n - \overline{S}_n$ is computed, where $\overline{S}$ is a running average of the score.

The global score S_n is computed as a mean of the individual scores S_n^i of each output neuron's spike train in trial t against its target spike train. The learning is done at the end of each trial in batch mode: $\Delta w_{ij} = \eta R_n e_{ij}(T)$.

RESULTS. Our results (see main figure) show that the R-max rule perform significantly better than R-STDP in our learning paradigm. This is not very surprising as this rule is optimal for precise spike time learning. In our experiments with the R-STDP rule, we found that the ratio $A_+\tau_+/A_-\tau_-$ does not affect learning in a significant way. It follows that the negative part of the STDP window is not necessary for learning in a reward modulated context.

[1] J. Hollerman and W. Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309, 1998.

[2] Verena Pawlak and Jason N D Kerr. Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *J Neurosci*, 28(10):2435–2446, Mar 2008.

[3] W. Gerstner, R. Kempter, J. Leo van Hemmen, and H. Wagner. A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383:76–78, 1996.

[4] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postysnaptic AP and EPSP. *Science*, 275:213–215, 1997.

[5] E.M. Izhikevich. Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cerebral Cortex*, 17:2443–2452, 2007.

[6] R. V. Florian. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation*, 19: 1468–1502, 2007.

[7] Robert Legenstein, Dejan Pecevski, and Wolfgang Maass. A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLOS Comput. Biol.*, 4:e1000180, 2008.

[8] L. F. Abbott and Sacha B. Nelson. Synaptic plasticity: Taming the beast. *Nature Neuroscience*, 3:1178–1183, 2000.

[9] Jean-Pascal Pfister, Taro Toyoizumi, David Barber, and Wulfram Gerstner. Optimal Spike-Timing-Dependent Plasticity for Precise Action Potential Firing in Supervised Learning. *Neural Comp.*, 18(6):1318–1348, 2006.

[10] X. Xie and S. Seung. Learning in neural networks by reinforcement of irregular spiking. *Phys. Rev. E*, 69:41909, 2004.

[11] R.J. Williams. Simple statistical gradient-following methods for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[12] J. Baxter and P.L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15(4):319–350, 2001.

[13] W. Gerstner and W. K. Kistler. *Spiking Neuron Models*. Cambridge University Press, Cambridge UK, 2002.

[14] J.D. Victor and K.P. Purpura. Metric-space analysis of spike trains: theory, algorithms, and application. *Network: Comput. Neural Syst*, 8:127–164, 1997.