

# Reinforcement learning in populations of spiking neurons

Robert Urbanczik & Walter Senn

**Population coding is widely regarded as an important mechanism for achieving reliable behavioral responses despite neuronal variability. However, standard reinforcement learning slows down with increasing population size, as the global reward signal becomes less and less related to the performance of any single neuron. We found that learning speeds up with increasing population size if, in addition to global reward, feedback about the population response modulates synaptic plasticity.**

The role of neuronal populations in encoding sensory stimuli has been intensively studied<sup>1,2</sup>. However, most models of reinforcement learning with spiking neurons have focused on just single neurons or small neuronal assemblies<sup>3–6</sup>. Furthermore, the following result indicates that such models do not scale well with population size. Using spike timing-dependent plasticity modulated by global reward in a large network, 100 learning trials are required to achieve an 80% probability of correctly associating a single stimulus with one of two responses, and performance does not improve with more training<sup>7</sup>. Behavioral results, in contrast, show that reinforcement learning can be reliable and fast. Macaque monkeys correctly associate one of four complex visual scenes with one of four target responses after an average of just 12 trials<sup>8</sup>.

Instead of simply broadcasting a global reward signal, as in reinforcement learning, procedures in artificial intelligence (for example, the back-propagation algorithm) use an involved machinery to generate feedback signals that are tailored to each neuron<sup>9</sup>. We found that there is a large middle ground between such biologically unrealistic procedures and standard reinforcement. Here we present a learning rule in which plasticity is modulated, not just by reward, but also by a single additional feedback signal encoding the population activity. Synapses receive the two signals via ambient neurotransmitter concentrations, leading to an on-line plasticity rule, where the learning of a first stimulus occurs concurrently with the processing of the subsequent one. Learning now speeds up with increasing population size, instead of deteriorating, as for standard reinforcement.

We studied a population of  $N$  neurons learning a decision task on the presynaptic inputs (**Supplementary Fig. 1** online). Input patterns consist of 50 spike trains (mean rate 6 Hz) of 500-ms duration. As a result of randomized projections, each of the  $N$  postsynaptic neurons receives input from roughly 40 of the 50 presynaptic spike trains. Therefore, inputs vary from one neuron to the next, but are highly correlated. Different neurons will produce different postsynaptic spike trains and aggregating these responses into a population decision must

be matched to how neurons encode information. We assumed a scoring function,  $c(Y^v)$ , that assigns a numerical value to the postsynaptic spike train  $Y^v$  evoked by the stimulus from neuron  $v$ . With this, aggregating the responses amounts to adding the postsynaptic scores to obtain a measure  $P$  of the population activity. In a pure rate code, for example,  $c(Y^v)$  is the number of spikes emitted by neuron  $v$ ,  $P$  is the total number of population spikes and the population decision might be reached by comparing  $P$  to a threshold. Here we shall focus on the simple case of a spike or no-spike code, in which  $c(Y^v)$  just registers whether the neuron fired in response to the stimulus. For notational convenience, we set  $c(Y^v) = -1$  if the neuron did not fire and  $c(Y^v) = 1$  if it did. The population decision (and thus the behavioral decision) is determined by the majority of the neurons. It equals 1 if more than half of the neurons fired in response to the stimulus and is  $-1$  otherwise. In formal terms, the population decision is the sign of  $P$ , the sum of the spike or no-spike scores.

In reinforcement learning, stochastic neuronal processing enables a population to explore different responses to a given stimulus. A reward signal  $R$  provides environmental feedback on whether the population decision was correct ( $R = 1$ ) or not ( $R = -1$ ). Plasticity is driven by this global feedback signal and by a quantity computed locally at each synapse, called eligibility trace. In our simulations, we used the escape noise neuron<sup>5</sup>, a leaky integrate and fire neuron with a stochastically fluctuating spike threshold. For this model, the synaptic eligibility trace depended on the timing of the pre- and postsynaptic spikes and on the somatic potential (**Supplementary Methods** online).

We first considered a standard reinforcement learning rule, where the strength,  $w_i^v$ , of each synapse ( $i$  is the synapse index,  $v$  is the neuron index) is changed by the equation

$$\Delta w_i^v = \eta(R - 1)E_i^v(T) \quad (1)$$

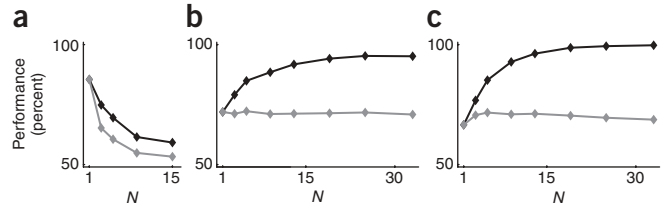
Here,  $E_i^v(T)$  is the eligibility trace, evaluated at the time  $T$  when the stimulus ends and the change occurs. The proportionality constant  $\eta$  is called the learning rate. As a result of the  $R - 1$  term, no change occurs when the population decision is correct ( $R = 1$ ). But for  $R = -1$ , the change encourages new responses by decreasing the probability that the neurons reproduce the same postsynaptic spike trains on a further trial with the same stimulus.

We determined the population performance (percentage of correct decisions) achieved with this rule for different population sizes  $N$  (**Fig. 1a**) and compared it with the performance of the individual neurons. Although the population outperformed the average single neuron, both performance measures deteriorated quickly with increasing  $N$ . The reason is rather simple. From the perspective of the single neuron, the global reward signal  $R$  is an unreliable performance measure, as the neuron may be punished for a correct response just because other neurons made a mistake and the odds of this happening increase with population size.

Department of Physiology, University of Bern, Bülhplatz 5, CH-3012 Bern, Switzerland. Correspondence should be addressed to W.S. (senn@pyl.unibe.ch).

Received 4 September 2008; accepted 23 December 2008; published online 15 February 2009; doi:10.1038/nn.2264

**Figure 1** Scaling properties of reinforcement learning. (a–c) With increasing population size  $N$ , performance deteriorates when learning is determined by just a global reward signal (a), but for individualized rewards, the population performance improves (b,c). The black line indicates the performance of the population read-out and the gray line indicates the average single neuron performance. Performance (percentage of correct decisions) was evaluated after a fixed number of learning steps; therefore, changes in performance reflect the changes in the speed of learning. Learning on the basis of just a global reward is shown in a (equation (1)). Learning with individual reward for each neuron is shown in b (equation (2)). Attenuated learning with individual reward is shown in c (equation (3)). In all cases, 30 patterns had to be learned with target responses that were equally split between the two output classes  $\pm 1$ . In each learning trial, a randomly selected pattern was presented, with the number of trials being 5,000 (a) and 2,000 (b,c). Simulation details are given in **Supplementary Methods**.



In human terms, standard reinforcement is analogous to having a class of students take an exam and then only telling them whether the majority of the class passed or failed, with the individual results being kept secret. An obvious alternative is to train the neurons individually and only use the population read-out to boost performance. For this, we assumed an individual reward signal  $r^v = \pm 1$ , indicating whether neuron  $v$  did the right thing and used

$$\Delta w_i^v = \eta(r^v - 1)E_i^v(T) \tag{2}$$

for the synaptic changes. Average single neuron performance was therefore constant, but population performance improved with increasing  $N$  up to  $95 \pm 1\%$  (Fig. 1b). Although individual reward works far better than global reward, the population boost does saturate because the neurons will tend to make the same mistakes, as they are all trying to learn the same thing.

To increase the population boost, we considered attenuating learning once the population decision was reliable and correct. Reliability is related to the population activity  $P$ , the sum of neuronal scores. If  $P$  is of large magnitude (compared with  $\sqrt{N}$ ), the sign of  $P$  (that is, the population decision) is unlikely to fluctuate as a result of noisy neural processing. If the response is also correct, then there is little need for further learning, even when some neurons responded incorrectly. We therefore introduced the attenuation factor  $a$  and considered the learning rule

$$\Delta w_i^v = \eta a(r^v - 1)E_i^v(T) \tag{3}$$

with  $a = \exp(-P^2/N)$  if  $R = 1$ . Otherwise,  $a = 1$  for a wrong population decision. This rule no longer tries to force all neurons to respond correctly to all stimuli. Hence, any given neuron can specialize on a certain subset of the stimuli, leading to a division of labor. Now,

perfect performance is approached with increasing population size (Fig. 1c). Equation (3) can be understood as a gradient descent rule (Supplementary Methods).

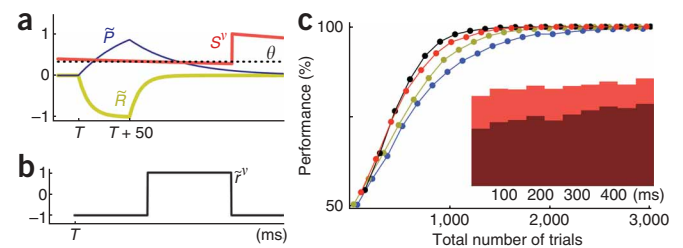
It might seem that attenuated learning requires a biophysically implausible number of feedback signals. However, just two feedback signals, global reward and the population activity, are needed if each neuron keeps a memory of its past activity. For example, if most neurons in the population spiked erroneously ( $P > 0, R = -1$ ), then a neuron,  $v$ , that stayed silent ( $c(Y^v) = -1$ ) did the right thing. So its individual reward is  $r^v = 1$ . More generally, the individual reward is

$$r^v = \text{sign}(R P c(Y^v)) \tag{4}$$

On the basis of this, we now present a fully on-line version of equation (3), which explicitly models the delivery of the feedback. In contrast with the above episodic rules, which pretend that synaptic changes are instantly triggered by stimulus endings, the on-line version takes into account the interaction between delays in feedback delivery and ongoing neuronal activity.

Taking a cue from experimental work on the modulation of synaptic plasticity<sup>10,11</sup>, we assumed that  $R$  and  $P$  are encoded by changes in ambient neurotransmitter levels. For reward, plasticity is modulated by a concentration variable,  $\bar{R}$ , which is essentially a low pass-filtered version of the instantaneous signal  $R$ . In the absence of reinforcement, the concentration of the neurotransmitter signaling reward, for example, dopamine, is maintained at a homeostatic level ( $\bar{R} = 0$ ), where a baseline release rate is balanced by an exponential decay. Reinforcement causes a transient change in the release rate, resulting in transient deviations of the concentration from its homeostatic level ( $\bar{R} \neq 0$ ; Fig. 2a). Feedback about the population activity  $P$  is provided similarly, via a second neurotransmitter with the concentration variable  $\bar{P}$ . For  $\bar{P}$ , we also assumed that the change in release rate (and thus  $\bar{P}$  itself) is

**Figure 2** Mechanism and performance of the on-line rule. (a,b) Examples are shown of on-line signals with synaptic read-out for a case in which, in the trial ending at time  $T$ , most neurons fired,  $P > 0$  (which was the wrong decision,  $R = -1$ ), and in which the example neuron  $v$  fired ( $c(Y^v) = 1$  and thus  $r^v = -1$ ). (a) The concentration variables  $\bar{P}$  and  $\bar{R}$  track the changes in neurotransmitter release rates, thus encoding  $P$  and  $R$  after the trial. In particular,  $\bar{P} > 0$  and  $\bar{R} < 0$  after time  $T$ , but their values start to decay at  $T + 50$  (when the release rates return to baseline). Because the example neuron was active during the trial, its memory variable  $s^v$  is still above the threshold  $\theta$  between  $T$  and  $T + 80$ , correctly reflecting  $c(Y^v) = 1$ . After that,  $s^v$  falls below threshold, as the neuron fired early in the trial. Note that we also allow a further spike in the subsequent trial, leading to the step increase in  $s^v$  at  $T + 150$ . (b) The on-line approximation  $\bar{r}^v$  to the individual reward  $r^v$  is correct initially, but changes when  $s^v$  falls below threshold at around  $T + 80$ . (c) Learning curves for on-line attenuated learning (equation (6)). Performance on the same task as in Figure 1 is shown in red. For comparison, the results obtained with equation (3) are shown in black. The performance of the on-line rule when the 30 patterns no longer all had the same 500-ms duration but were of different lengths (randomly chosen between 300 and 700 ms) is shown in yellow. The results when, in contrast with our standard assumption, the reward information itself only became available 100 ms after each trial ended are shown in blue. Inset, distribution of postsynaptic spike times after learning on the basis of the responses of all neurons to all of the patterns. The x axis denotes time elapsed from start of trial and dark red highlights the contributions from the patterns where the goal is to spike. Results in the panel are for  $N = 67$ .



attenuated with increasing magnitude of  $P$  (**Supplementary Methods**). For the memory mechanism, enabling each neuron to determine its score  $c(Y^v)$ , we assumed a calcium-like variable  $s^v$ . When the neuron does not fire,  $s^v$  decays exponentially with a time constant  $\tau_M$  of 500 ms. It is updated to  $s^v(t) = 1$  if there is a postsynaptic spike at time  $t$ . Therefore, the value of  $s^v$  is directly related to the time elapsed since the last spike.

Now, synapses can read out an approximation  $\tilde{r}^v$  to the individual reward signal  $r^v$  in equation (4) using

$$\tilde{r}^v = \text{sign}(\tilde{R}\tilde{P}(s^v - \theta)) \quad (5)$$

For a good choice of the threshold  $\theta$  (and the time constant  $\tau_M$ ), the value of  $\tilde{r}^v(t)$  is equal to  $r^v$  immediately after the trial. But, as a result of ongoing activity,  $\tilde{r}^v$  can change with time (**Fig. 2b**). To address this, we introduced an effective learning rate,  $\tilde{\eta}$ , modulated by the magnitude of the reward variable, and used a plasticity rule,

$$\dot{w}_i^v = \tilde{\eta}(t)\tilde{a}(t)(\tilde{r}^v(t) - 1)E_i^v(t) \quad (6)$$

with  $\tilde{\eta}(t) = \eta|\tilde{R}(t)|$ . The modulation of  $\tilde{\eta}$  confines the update to a time window because  $\tilde{R}$  is already close to 0 some 100 ms after the end of each trial. Thus, synapses can further use the momentary value  $E_i^v(t)$  of the eligibility trace rather than memorizing  $E_i^v(T)$ , the value that the trace had at the end of the trial. Because we assumed that  $\tilde{P}$  is attenuated when the population majority is large, setting  $\tilde{a} = |\tilde{P}|$  for a correct response ( $\tilde{R} > 0$ ) and  $\tilde{a} = 1$  for  $\tilde{R} < 0$  achieves the attenuation of learning.

The performance of equation (6) was very similar to that of the episodic version (**Fig. 2c**). After learning, postsynaptic spike timing was distributed quite uniformly (**Fig. 2c**), showing that the entire temporal extent of a stimulus is taken into account by the population decision. The on-line procedure also performed well with stimuli of variable lengths, although  $\tau_M$  was no longer precisely matched to stimulus duration (**Fig. 2c**). To check robustness further, we simulated delayed reinforcement, assuming that the reward information itself only becomes available 100 ms after each trial has ended (effectively

shifting the yellow curve in **Fig. 2a** to the right by 20% of stimulus duration). Reward onset now occurred during the subsequent trial, but perfect performance was approached nevertheless, albeit with a slow down in learning (**Fig. 2c**). To focus on robustness, we used the same parameter values for the three tasks in the on-line procedure.

As a specific neuronal model, we have used the escape noise neuron<sup>5</sup>, but our approach could readily be adapted to other reinforcement procedures<sup>3,4,6</sup> at the single neuron level. Indeed, for episodic learning, the neurons could even be tempotrons<sup>12</sup>. In absolute terms, population performance will obviously depend on the specifics of the neuronal model and plasticity rule, but with regard to the scaling of learning performance with population size, we expect our findings to be generic.

*Note: Supplementary information is available on the Nature Neuroscience website.*

#### ACKNOWLEDGMENTS

We thank W. Gerstner for helpful comments on the manuscript.

#### AUTHOR CONTRIBUTIONS

R.U. designed and executed the research in close interaction with W.S. The manuscript was written jointly.

Published online at <http://www.nature.com/natureneuroscience/>

Reprints and permissions information is available online at <http://www.nature.com/reprintsandpermissions/>

1. Pouget, A., Dayan, P. & Zemel, R. *Nat. Rev. Neurosci.* **1**, 125–132 (2000).
2. Averbeck, B.B., Latham, P.E. & Pouget, A. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
3. Seung, H.S. *Neuron* **40**, 1063–1073 (2003).
4. Fiete, I.R. & Seung, H.S. *Phys. Rev. Lett.* **97**, 048104 (2006).
5. Pfister, J.P., Toyozumi, T., Barber, D. & Gerstner, W. *Neural Comput.* **18**, 1318–1348 (2006).
6. Florian, R.V. *Neural Comput.* **19**, 1468–1502 (2007).
7. Izhikevich, E.M. *Cereb. Cortex* **17**, 2443–2452 (2007).
8. Wirth, S. *et al. Science* **300**, 1578–1581 (2003).
9. Rumelhart, D.E., Hinton, G.E. & Williams, R.J.D. *Parallel Distributed Processing*, Vol. 1 (eds. Rumelhart, E. & McClelland, J.L.) (MIT Press, Cambridge, Massachusetts, 1986).
10. Fehring, R.C. & Lorenzon, N.M. *Can. J. Exp. Psychol.* **53**, 45–61 (1999).
11. Centonze, D., Gubellini, P., Pisani, A., Bernardi, G. & Calabresi, P. *Rev. Neurosci.* **14**, 207–216 (2003).
12. Gütig, R. & Sompolinsky, H. *Nat. Neurosci.* **9**, 420–428 (2006).