A Learning Theory for Reward-Modulated Spike-Timing-Dependent Plasticity with Application to Biofeedback

Robert Legenstein*, Dejan Pecevski*, Wolfgang Maass
Institute for Theoretical Computer Science
Graz University of Technology
A-8010 Graz, Austria
{legi,dejan,maass}@igi.tugraz.at

June 23, 2008

Abstract

Reward-modulated spike-timing-dependent plasticity (STDP) has recently emerged as a candidate for a learning rule that could explain how behaviorally relevant adaptive changes in complex networks of spiking neurons could be achieved in a self-organizing manner through local synaptic plasticity. However the capabilities and limitations of this learning rule could so far only be tested through computer simulations. This article provides tools for an analytic treatment of reward-modulated STDP, which allows us to predict under which conditions reward-modulated STDP will achieve a desired learning effect. These analytical results imply that neurons can learn through reward-modulated STDP to classify not only spatial, but also temporal firing patterns of presynaptic neurons. They also can learn to respond to specific presynaptic firing patterns with particular spike patterns. Finally, the resulting learning theory predicts that even difficult credit-assignment problems, where it is very hard to tell which synaptic weights should be modified in order to increase the global reward for the system, can be solved in a self-organizing manner through reward-modulated STDP. This yields an explanation for a fundamental experimental result on biofeedback in monkeys by Fetz and Baker. In this experiment monkeys were rewarded for increasing the firing rate of a particular neuron in the cortex, and were able to solve this extremely difficult credit assignment problem. Our model for this experiment relies on a combination of reward-modulated STDP with variable spontaneous firing activity. Hence it also provides a possible functional explanation for trial-to-trial variability, which is characteristic for cortical networks of neurons, but has no analogue in currently existing artificial computing systems. In addition our model demonstrates that reward-modulated STDP can be applied to all synapses in a large recurrent neural network without endangering the stability of the network dynamics.

*These authors contributed equally to this work.

1 Introduction

Numerous experimental studies (see [1] for a review; [2] discusses more recent in-vivo results) have shown that the efficacy of synapses changes in dependence of the time difference $\Delta t = t_{post} - t_{pre}$ between the firing times t_{pre} and t_{post} of the pre- and postsynaptic neurons. This effect is called spike-timing-dependent plasticity (STDP). But a major puzzle for understanding learning in biological organisms is the relationship between experimentally well-established rules for STDP on the microscopic level, and adaptive changes of the behavior of biological organisms on the macroscopic level. Neuromodulatory systems, which send diffuse signals related to reinforcements (rewards) and behavioral state to several large networks of neurons in the brain, have been identified as likely intermediaries that relate these two levels of plasticity. It is well-known that the consolidation of changes of synaptic weights in response to pre- and postsynaptic neuronal activity requires the presence of such third signals [3, 4]. In particular, it has been demonstrated that dopamine (which is behaviorally related to novelty and reward prediction [5]) gates plasticity at corticostriatal synapses [6, 7] and within the cortex [8]. It has also been shown that acetylcholine gates synaptic plasticity in the cortex (see for example [9] and [10]; [11] contains a nice review of the literature).

Corresponding spike-based rules for synaptic plasticity of the form

$$\frac{d}{dt} w_{ji}(t) = c_{ji}(t) d(t) \tag{1}$$

have been proposed in [12] and [13] (see Fig. 1 for an illustration of this learning rule), where w_{ji} is the weight of a synapse from neuron i to neuron j, $c_{ji}(t)$ is an eligibility trace of this synapse which collects weight changes proposed by STDP, and $d(t) = h(t) - \bar{h}$ results from a neuromodulatory signal $h(t)$ with mean value $\bar{h}$. It was shown in [12] that a number of interesting learning tasks in large networks of neurons can be accomplished with this simple rule (1). It has recently been shown that quite similar learning rules for spiking neurons arise when one applies the general framework of distributed reinforcement learning from [14] to networks of spiking neurons [15, 13], or if one maximizes the likelihood of postsynaptic firing at desired firing times [16]. However no analytical tools have been available, which make it possible to predict for what learning tasks, and under which parameter settings, reward-modulated STDP will be successful. This article provides such analytical tools, and demonstrates their applicability and significance through a variety of computer simulations. In particular, we identify conditions under which neurons can learn through reward-modulated STDP to classify temporal presynaptic firing patterns, and to respond with particular spike patterns.

We also provide a model for the remarkable operant conditioning experiments of [17] (see also [18, 19]). In the simpler ones of these experiments the spiking activity of single neurons (in area 4 of the precentral gyrus of monkey cortex) was recorded, the deviation of the current firing rate of an arbitrarily selected neuron from its average firing rate was made visible to the monkey through the displacement of an illuminated meter arm, whose rightward position corresponded to the threshold for the feeder discharge. The monkey received food rewards for increasing (or in alternating trials for decreasing) the firing rate of this neuron. The monkeys learnt quite reliably (within a few minutes) to change

the firing rate of this neuron in the currently rewarded direction.[1] Obviously the existence of learning mechanisms in the brain which are able to solve this extremely difficult credit assignment problem provides an important clue for understanding the organization of learning in the brain. We examine in this article analytically under what conditions reward-modulated STDP is able to solve such learning problem. We test the correctness of analytically derived predictions through computer simulations of biologically quite realistic recurrently connected networks of neurons, where an increase of the firing rate of one arbitrarily selected neuron within a network of 4000 neurons is reinforced through rewards (which are sent to all 142813 synapses between excitatory neurons in this recurrent network). We also provide a model for the more complex operant conditioning experiments of [17] by showing that pairs of neurons can be differentially trained through reward-modulated STDP, where one neuron is rewarded for increasing its firing rate, and simultaneously another neuron is rewarded for decreasing its firing rate. More precisely, we increased the reward signal $d(t)$ which is transmitted to all synapses between excitatory neurons in the network whenever the first neuron fired, and decreased this reward signal whenever the second neuron fired (the resulting composed reward corresponds to the displacement of the meter arm that was shown to the monkey in these more complex operant conditioning experiments).

Our theory and computer simulations also show that reward-modulated STDP can be applied to all synapses within a large network of neurons for long time periods, without endangering the stability of the network. In particular this synaptic plasticity rule keeps the network within the asynchronous irregular firing regime, which had been described in [21] as a dynamic regime that resembles spontaneous activity in the cortex. Another interesting aspect of learning with reward-modulated STDP is that it requires spontaneous firing and trial-to-trial variability within the networks of neurons where learning takes place. Hence our learning theory for this synaptic plasticity rule provides a foundation for a functional explanation of these characteristic features of cortical network of neurons that are undesirable from the perspective of most computational theories.

[1] Adjacent neurons tended to change their firing rate in the same direction, but also differential changes of directions of firing rates of pairs of neurons are reported in [17] (when these differential changes were rewarded). For example, it was shown in Fig. 9 of [17] (see also Fig. 1 in [19]) that pairs of neurons that were separated by no more than a few hundred microns could be independently trained to increase or decrease their firing rates. It was also reported in [17], and further examined in [20], that bursts of the reinforced neurons were often accompanied by activations of specific muscles. But the relationship between bursts of the recorded neurons in precentral motor cortex and muscle activations was reported to be quite complex and often dropped out after continued reinforcement of the neuron alone. Furthermore in [20] it was shown that all neurons tested in that study could be dissociated from their correlated muscle activity by differentially reinforcing simultaneous suppression of EMG activity. These results suggest that the solution of the credit assignment problem by the monkeys (to stronger activate that neuron out of billions of neurons in their precentral gyrus that was reinforced) may have been supported by large scale exploration strategies that were associated with muscle activations. But the previously mentioned results on differential reinforcements of two nearby neurons suggest that this large scale exploration strategy had to be complemented by exploration on a finer spatial scale that is difficult to explain on the basis of muscle activations (see section 2 of [19] for a detailed discussion).

2 Results

We first give a precise definition of the learning rule (1) for reward-modulated STDP. The standard rule for STDP, which specifies the change $W(\Delta t)$ of the synaptic weight of an excitatory synapse in dependence on the time difference $\Delta t = t_{post} - t_{pre}$ between the firing times t_{pre} and t_{post} of the pre- and postsynaptic neuron, is based on numerous experimental data (see [1]). It is commonly modeled by a so-called learning curve of the form

$$W(\Delta t) = \begin{cases} A_+ e^{-\Delta t/\tau_+} & , \quad \text{if } \Delta t \geq 0 \\ -A_- e^{\Delta t/\tau_-} & , \quad \text{if } \Delta t < 0 \end{cases} , \tag{2}$$

where the positive constants A_+ and A_- scale the strength of potentiation and depression respectively, and τ_+ and τ_- are positive time constants defining the width of the positive and negative learning window. The resulting weight change at time t of synapse ji for a presynaptic spike train S_i^{pre} and a postsynaptic spike train S_j^{post} is usually modeled [22] by the instantaneous application of this learning rule to all spike pairings with the second spike at time t

$$\left[\frac{d}{dt} w_{ji}(t)\right]_{STDP} = \int_0^\infty dr\, W(r) S_j^{post}(t) S_i^{pre}(t-r) + \int_0^\infty dr\, W(-r) S_j^{post}(t-r) S_i^{pre}(t). \tag{3}$$

The spike train of a neuron i which fires action potentials at times $t_i^{(1)}, t_i^{(2)}, t_i^{(3)}, \ldots$ is formalized here by a sum of Dirac delta functions $S_i(t) = \sum_n \delta(t - t_i^{(n)})$.

The model analyzed in this article is based on the assumption that positive and negative weight changes suggested by STDP for all pairs of pre- and postsynaptic spikes at synapse ji (according to the two integrals in (3)) are collected in an eligibility trace $c_{ji}(t)$ at the site of the synapse. The contribution to $c_{ij}(t)$ of all spike pairings with the second spike at time $t - s$ is modeled for $s > 0$ by a function $f_c(s)$ (see Fig. 1A); the time scale of the eligibility trace is assumed in this article to be on the order of seconds. Hence the value of the eligibility trace of synapse ji at time t is given by

$$c_{ji}(t) = \int_0^\infty ds f_c(s) \left[\frac{d}{dt} w_{ji}(t-s)\right]_{STDP} , \tag{4}$$

see Fig. 1B. The actual weight change $\frac{d}{dt} w_{ji}(t)$ at time t for reward-modulated STDP is the product $c_{ij}(t) \cdot d(t)$ of the eligibility trace with the reward signal $d(t)$ as defined by equation (1). Since this simple model can in principle lead to unbounded growth of weights, we assume that weights are clipped at the lower boundary value 0 and an upper boundary w_{max}.

The network dynamics of a simulated recurrent network of spiking neurons where all connections between excitatory neurons are subject to STDP is quite sensitive to the particular STDP-rule that is used. Therefore we have carried out our network simulations not only with the additive STDP-rule (3), whose effect can be analyzed theoretically, but also with the more complex rule proposed in [23] (which was fitted to experimental data from hippocampal neurons in culture [24]), where the magnitude of the weight change depends on the current value of the weight. An implementation of this STDP-rule (with the

parameters proposed in [23]) produced in our network simulations of the biofeedback experiment (computer simulation 1) as well as for learning pattern classification (computer simulation 4) qualitatively the same result as rule (3).

Theoretical analysis of the resulting weight changes

In this section, we derive a learning equation for reward-modulated STDP. This learning equation relates the change of a synaptic weight w_{ji} over some sufficiently long time interval T to statistical properties of the joint distribution of the reward signal $d(t)$ and pre- and postsynaptic firing times, under the assumption that the weight and correlations between pre- and postsynaptic spike times are slowly varying in time. We treat spike times as well as the reward signal $d(t)$ as stochastic variables. This mathematical framework allows us to derive the expected weight change over some time interval T (see [22]), with the expectation taken over realizations of the stochastic input- and output spike trains as well as stochastic realizations of the reward signal, denoted by the ensemble average $\langle\cdot\rangle_E$

$$\frac{\langle w_{ji}(t+T) - w_{ji}(t)\rangle_E}{T} = \frac{1}{T}\left\langle \int_t^{t+T} \frac{d}{dt} w_{ji}(t')dt' \right\rangle_E = \left\langle \left\langle \frac{d}{dt} w_{ji}(t) \right\rangle_T \right\rangle_E , \quad (5)$$

where we used the abbreviation $\langle f(t)\rangle_T = T^{-1}\int_t^{t+T} f(t')\, dt'$. If synaptic plasticity is sufficiently slow, synaptic weights integrate a large number of small changes. In this case, the weight w_{ji} can be approximated by its average $\langle w_{ji}\rangle_E$ (it is "self-averaging", see [22]). We can thus drop the expectation on the left hand side of equation (5) and write it as $\frac{d}{dt}\langle w_{ji}(t)\rangle_T$. Using equation (1), this yields (see Methods)

$$\boxed{\begin{aligned}\frac{d}{dt}\langle w_{ji}(t)\rangle_T = &\int_0^\infty dr\, W(r) \int_0^\infty ds\, f_c(s)\, \langle D_{ji}(t,s,r)\, \nu_{ji}(t-s,r)\rangle_T \\ &+ \int_{-\infty}^0 dr\, W(r) \int_{|r|}^\infty ds\, f_c(s+r)\, \langle D_{ji}(t,s,r)\, \nu_{ji}(t-s,r)\rangle_T .\end{aligned}} \quad (6)$$

This formula contains the *reward correlation* for synapse ji

$$D_{ji}(t,s,r) = \langle d(t) |\, \text{Neuron } j \text{ spikes at } t-s, \text{ and neuron } i \text{ spikes at } t-s-r\rangle_E , \quad (7)$$

which is the average reward at time t given a presynaptic spike at time $t-s-r$ and a postsynaptic spike at time $t-s$. The joint firing rate $\nu_{ji}(t,r) = \langle S_j(t) S_i(t-r)\rangle_E$ describes correlations between spike timings of neurons j and i, i.e., it is the probability density for the event that neuron i fires an action potential at time $t-r$ and neuron j fires an action potential at time t. For synapses subject to reward-modulated STDP, changes in efficacy are obviously driven by co-occurrences of spike pairings and rewards within the time scale of the eligibility trace. Equation (6) clarifies how the expected weight change depends on how the correlations between the pre- and postsynaptic neurons correlate with the reward signal.

If one assumes for simplicity that the impact of a spike pair on the eligibility trace is always triggered by the postsynaptic spike, one gets a simpler equation (see Methods)

$$\boxed{\frac{d}{dt}\langle w_{ji}(t)\rangle_T = \int_0^\infty ds\, f_c(s) \int_{-\infty}^\infty dr\, W(r)\, \langle D_{ji}(t,s,r)\, \nu_{ji}(t-s,r)\rangle_T .} \quad (8)$$

The assumption introduces a small error for post-before-pre spike pairs, because for a reward signal that arrives at some time d_r after the pairing, the weight update will be proportional to $f_c(d_r)$ instead of $f_c(d_r + r)$. The approximation is justified if the temporal average is performed on a much longer time scale than the time scale of the learning window, the effect of each pre-post spike pair on the reward signal is delayed by an amount greater than the time scale of the learning window, and f_c changes slowly compared to the time scale of the learning window (see Methods for details). For the analyzes presented in this article, the simplified equation (8) is a good approximation for the learning dynamics. Equation (8) is a generalized version of the STDP learning equation $\frac{d}{dt}w_{ji}(t) = \int_{-\infty}^{\infty} dr\ W(r)\,\langle \nu_{ji}(t-s,r)\rangle_T$ in [22] that includes the impact of the reward correlation weighted by the eligibility function. To see the relation between standard STDP and reward-modulated STDP, consider a constant reward signal $d(t) = d_0$. Then also the reward correlation is constant and given by $D(t,s,r) = d_0$. We recover the standard STDP learning equation scaled by d_0 if the eligibility function is an instantaneous delta-pulse $f_c(s) = \delta(s)$. Furthermore, if the statistics of the reward signal $d(t)$ is time-independent and independent from the pre- and postsynaptic spike statistics of some synapse ji, then the reward correlation is given by $D_{ji}(t,s,r) = \langle d(t)\rangle_E = d_0$ for some constant d_0. Then, the weight change for synapse ji is $\frac{d}{dt}\langle w_{ji}(t)\rangle_T = d_0 \int_{-\infty}^{\infty} dr\ W(r) \int_0^{\infty} ds f_c(s)\,\langle \nu_{ji}(t-s,r)\rangle_T$. The temporal average of the joint firing rate $\langle \nu_{ji}(t-s,r)\rangle_T$ is thus filtered by the eligibility trace. We assumed in the preceding analysis that the temporal average is taken over some long time interval T. If the time scale of the eligibility trace is much smaller than this time interval T, then the weight change is approximately $\frac{d}{dt}\langle w_{ji}(t)\rangle_T \approx d_0(\int_0^{\infty} ds f_c(s)) \int_{-\infty}^{\infty} dr\ W(r)\,\langle \nu_{ji}(t,r)\rangle_T$, and the weight w_{ji} will change according to standard STDP scaled by a constant proportional to the mean reward and the integral over the eligibility function. In the remainder of this article, we will always use the smooth time-averaged weight change $\frac{d}{dt}\,\langle w_{ji}(t)\rangle_T$, but for brevity, we will drop the angular brackets and simply write $\frac{d}{dt}w_{ji}(t)$.

The learning equation (8) provides the mathematical basis for our following analyses. It allows us to determine synaptic weight changes if we can describe a learning situation in terms of reward correlations and correlations between pre- and postsynaptic spikes.

Application to models for biofeedback experiments

We now apply the preceding analysis to the biofeedback experiment of [17] that were described in the introduction. These experiments pose the challenge to explain how learning mechanisms in the brain can detect and exploit correlations between rewards and the firing activity of one or a few neurons within a large recurrent network of neurons (the credit assignment problem), without changing the overall function or dynamics of the circuit.

We show that this phenomenon can in principle be explained by reward-modulated STDP. In order to do that, we define a model for the experiment which allows us to formulate an equation for the reward signal $d(t)$. This enables us to calculate synaptic weight changes for this particular scenario. We consider as model a recurrent neural circuit where the spiking activity of one neuron k is recorded by the experimenter.[2] We assume that in the monkey brain a reward signal $d(t)$ is produced which depends on the

[2] Experiments where two neurons are recorded and reinforced were also reported in [17]. We tested this case in computer simulations (see Fig. (4)) but did not treat it explicitly in our theoretical analysis.

visual feedback (through an illuminated meter, whose pointer deflection was dependent on the current firing rate of the randomly selected neuron k) as well as previously received liquid rewards, and that this signal $d(t)$ is delivered to *all* synapses in large areas of the brain. We can formalize this scenario by defining a reward signal which depends on the spike rate of the arbitrarily selected neuron k (see Fig. 2A, B). More precisely, a reward pulse of shape $\epsilon_r(r)$ (the reward kernel) is produced with some delay d_r every time the neuron k produces an action potential

$$d(t) = \int_0^\infty dr\ S_k^{post}(t - d_r - r)\epsilon_r(r). \tag{9}$$

Note that $d(t) = h(t) - \bar{h}$ is defined in equation (1) as a signal with zero mean. In order to satisfy this constraint, we assume that the reward kernel ϵ_r has zero mass, i.e., $\bar{\epsilon}_r = \int_0^\infty dr\ \epsilon_r(r) = 0$. For the analysis, we use the linear Poisson neuron model described in Methods. The mean weight change for synapses to the reinforced neuron k is then approximately (see Methods)

$$\boxed{\frac{d}{dt}w_{ki}(t) \approx \int_0^\infty ds\ f_c(s + d_r)\epsilon_r(s) \int_{-\infty}^\infty dr\ W(r)\ \langle \nu_{ki}(t - d_r - s, r)\rangle_T\,.} \tag{10}$$

This equation describes STDP with a learning rate proportional to $\int_0^\infty ds\ f_c(s + d_r)\epsilon_r(s)$. The outcome of the learning session will strongly depend on this integral and thus on the form of the reward kernel ϵ_r. In order to reinforce high firing rates of the reinforced neuron we have chosen a reward kernel with a positive bump in the first few hundred milliseconds, and a long negative tail afterwards. Fig. 2C shows the functions f_c and ϵ_r that were used in our computer model, as well as the product of these two functions. One sees that the integral over the product is positive and according to equation (10) the synapses to the reinforced neuron are subject to STDP. This does not guarantee an increase of the firing rate of the reinforced neuron. Instead, the changes of neuronal firing will depend on the statistics of the inputs. In particular, the weights of synapses to neuron k will not increase if that neuron does not fire spontaneously. For uncorrelated Poisson input spike trains of equal rate, the firing rate of a neuron trained by STDP stabilizes at some value which depends on the input rate (see [25, 26]). However, in comparison to the low spontaneous firing rates observed in the biofeedback experiment [17], the stable firing rate under STDP can be much higher, allowing for a significant rate increase. It was shown in [17] that also low firing rates of a single neuron can be reinforced. In order to model this, we have chosen a reward kernel with a negative bump in the first few hundred milliseconds, and a long positive tail afterwards, i.e. we inverted the kernel used above to obtain a negative integral $\int_0^\infty ds\ f_c(s + d_r)\epsilon_r(s)$. According to equation (10) this leads to anti-STDP where not only inputs to the reinforced neuron which have low correlations with the output are depressed (because of the negative integral of the learning window), but also those which are causally correlated with the output. This leads to a quick firing rate decrease at the reinforced neuron.

The mean weight change of synapses to non-reinforced neurons $j \neq k$ is given by

$$\boxed{\frac{d}{dt}w_{ji}(t) \approx \int_0^\infty ds\ f_c(s) \int_{-\infty}^\infty dr\ W(r) \int_0^\infty dr' \epsilon_r(r') \left\langle \frac{\nu_{kj}(t-d_r-r', s-d_r-r')}{\nu_j(t-s)} \nu_{ji}(t-s, r) \right\rangle_T ,} \tag{11}$$

where $\nu_j(t) = \langle S_j(t) \rangle_E$ is the instantaneous firing rate of neuron j at time t. This equation indicates that a non-reinforced neuron is trained by STDP with a learning rate proportional to its correlation with the reinforced neuron given by $\nu_{kj}(t-d_r-r', s-d_r-r')/\nu_j(t-s)$. In fact, it was noted in [17] that neurons nearby the reinforced neuron tended to change their firing rate in the same direction. This observation might be explained by putative correlations of the recorded neuron with nearby neurons. On the other hand, if a neuron j is uncorrelated with the reinforced neuron k, we can decompose the joint firing rate into $\nu_{kj}(t-d_r-r', s-d_r-r') = \nu_k(t-d_r-r')\nu_j(t-s)$. In this case, the learning rate for synapse ji is approximately zero (see Methods). This ensures that most neurons in the circuit keep a constant firing rate, in spite of continuous weight changes according to reward-modulated STDP.

Altogether we see that the weights of synapses to the reinforced neuron k can only change if there is spontaneous activity in the network, so that in particular also this neuron k fires spontaneously. On the other hand the spontaneous network activity should not consist of repeating large-scale spatio-temporal firing patterns, since that would entail correlations between the firing of neuron k and other neurons j, and would lead to similar changes of synapses to these other neurons j. Apart from these requirements on the spontaneous network activity, the preceding theoretical results predict that stability of the circuit is preserved, while the neuron which is causally related to the reward signal is trained by STDP, if $\int_0^\infty ds\ f_c(s+d_r)\epsilon_r(s)$ is positive.

Computer simulation 1: Model for biofeedback experiment

We tested these theoretical predictions through computer simulations of a generic cortical microcircuit receiving a reward signal which depends on the firing of one arbitrarily chosen neuron k from the circuit (reinforced neuron). The circuit was composed of 4000 LIF neurons, with 3200 being excitatory and 800 inhibitory, interconnected randomly by 228954 conductance based synapses with short term dynamics [3]. In addition to the explicitly modeled synaptic connections, conductance noise (generated by an Ornstein-Uhlenbeck process) was injected into each neuron according to data from [27], in order to model synaptic background activity of neocortical neurons in-vivo.[4] This background noise elicited spontaneous firing in the circuit at about 4.6 Hz. Reward-modulated STDP

[3] All computer simulations were also carried out as a control with static current based synapses, see Methods and Suppl.

[4] More precisely, for 50% of the excitatory neurons the amplitude of the noise injection was reduced to 20%, and instead their connection probabilities from other excitatory neurons were chosen to be larger (see Methods and Suppl. Fig. 1 and 2 for details). The reinforced neuron had to be chosen from the latter population, since reward-modulated STDP does not work properly if the postsynaptic neuron fires too often because of directly injected noise.

was applied continuously to all synapses which had excitatory presynaptic and postsynaptic neurons, and all these synapses received the same reward signal. The reward signal was modeled according to equation (9). Fig. 2C shows one reward pulse caused by a single postsynaptic spike at time $t = 0$ with the parameters used in the experiment. For several postsynaptic spikes, the amplitude of the reward signal follows the firing rate of the reinforced neuron, see Fig. 2B.

This model was simulated for 20 minutes of biological time. Panels A, B, D of Fig. 3 show that the firing rate of the reinforced neuron increases within a few minutes (like in the experiment of [17]), while the firing rates of the other neurons remain largely unchanged. The increase of weights to the reinforced neuron shown in Fig. 3C can be explained by the correlations between its presynaptic and postsynaptic spikes shown in panel E. This panel shows that pre-before-post spike pairings (black curve) are in general more frequent than post-before-pre spike pairings. The reinforced neuron increases its rate from around 4 Hz to 12 Hz, which is comparable to the measured firing rates in [17] before and after learning.

In Fig. 9 of [17] and Fig. 1 of [19] the results of another experiment were reported where the activity of two adjacent neurons was recorded, and high firing rates of the first neuron and low firing rates of the second neuron were reinforced simultaneously. This kind of differential reinforcement resulted in an increase and decrease of the firing rates of the two neurons correspondingly. We implemented this type of reinforcement by letting the reward signal in our model depend on the spikes of the two randomly chosen neurons (we refer to these neurons as neuron A and neuron B), i.e. $d(t) = d_+^A(t) + d_-^B(t)$, where $d_+^A(t)$ is the component that positively rewards spikes of neuron A, and $d_-^B(t)$ negatively rewards spikes of neuron B. Both parts of the reward signal, $d_+^A(t)$ and $d_-^B(t)$, were defined as in equation (9) for the corresponding neuron. For $d_+^A(t)$ we used the reward kernel ϵ_r as defined in equation (29), whereas for $d_-^B(t)$ we used $\epsilon_{r-} = -\epsilon_r$ (note that the integral over ϵ_{r-} is still zero). At the middle of the simulation (simulation time $t = 10$min), we changed the direction of the reinforcements by negatively rewarding the firing of neuron A and positively rewarding the firing of neuron B (i.e., $d(t) = d_-^A(t) + d_+^B(t)$). The results are summarized in Fig. 4. With a reward signal modeled in this way, we were able to independently increase and decrease the firing rates of the two neurons according to the reinforcements, while the firing rates of the other neurons remained unchanged. Changing the type of reinforcement during the simulation from positive to negative for neuron A and from negative to positive for neuron B resulted in a corresponding shift in their firing rate change in the direction of the reinforcement.

The dynamics of a network where STDP is applied to all synapses between excitatory neurons is quite sensitive to the specific choice of the STDP-rule. The preceding theoretical analysis (see equation (10), (11)) predicts that reward-modulated STDP affects in the long run only those excitatory synapses where the firing of the postsynaptic neuron is correlated with the reward signal. In other words: the reward signal gates the effect of STDP in a recurrent network, and thereby can keep the network within a given dynamic regime. This prediction is confirmed qualitatively by the two panels of Fig. 3A, which show that even after all excitatory synapses in the recurrent network have been subject to 20 minutes (in simulated biological time) of reward-modulated STDP, the network stays within the asynchronous irregular firing regime. It is also confirmed quantitatively through Fig. 5. These figures show results for the simple additive version of STDP (according to equation

(3)). Very similar results (see Suppl. Fig. 3 and 4) arise from an application of the more complex STDP-rule proposed in [23] where the weight-change depends on the current weight value.

Rewarding spike-times

The preceding model for the biofeedback experiment of Fetz and Baker focused on learning of firing rates. In order to explore the capabilities and limitations of reward-modulated STDP in contexts where the temporal structure of spike trains matters, we investigated another reinforcement learning scenario where a neuron should learn to respond with particular temporal spike patterns. We first apply analytical methods to derive conditions under which a neuron subject to reward-modulated STDP can achieve this.

In this model, the reward signal $d(t)$ is given in dependence on how well the output spike train S_j^{post} of a neuron j matches some rather arbitrary spike train S^* (which might for example represent spike output from some other brain structure during a developmental phase). S^* is produced by a neuron μ^* that receives the same n input spike trains $S_1, \ldots, S_n$ as the trained neuron j, with some arbitrarily chosen weights $\mathbf{w}^* = (w_1^*, \ldots, w_n^*)^T$, $w_i^* \in \{0, w_{max}\}$. But in addition the neuron μ^* receives $n' - n$ further spike trains $S_{n+1}, \ldots, S_{n'}$ with weights $w_{n+1}^*, \ldots, w_{n'}^* = w_{max}$. The setup is illustrated in Fig. 6A. It provides a generic reinforcement learning scenario, when a quite arbitrary (and not perfectly realizable) spike output is reinforced, but simultaneously the performance of the learner can be evaluated clearly according to how well its weights $w_{j1}, \ldots, w_{jn}$ match those of the neuron μ^* for those n input spike trains which both of them have in common. The reward $d(t)$ at time t depends in this task on both the timing of action potentials of the trained neuron and spike times in the target spike train S^*

$$d(t) = \int_{-\infty}^{\infty} dr\ \kappa(r) S_j^{post}(t - d_r) S^*(t - d_r - r), \tag{12}$$

where the function $\kappa(r)$ with $\bar{\kappa} = \int_{-\infty}^{\infty} ds\ \kappa(s) > 0$ describes how the reward signal depends on the time difference r between a postsynaptic spike and a target spike, and $d_r > 0$ is the delay of the reward.

Our theoretical analysis (see Methods) predicts that under the assumption of constant-rate uncorrelated Poisson input statistics this reinforcement learning task can be solved by reward-modulated STDP for arbitrary initial weights if three constraints are fulfilled:

$$-\nu_{min}^{post} \bar{W} > w_{max} \bar{W}_\epsilon \tag{13}$$

$$\int_{-\infty}^{\infty} dr\ W(r)\epsilon(r)\epsilon_\kappa(r) \geq -\nu_{max}^{post} \bar{W} \int_0^{\infty} dr\ \epsilon(r)\epsilon_\kappa(r) \tag{14}$$

$$\int_{-\infty}^{\infty} dr\ W(r)\epsilon_\kappa(r) > -\bar{W}\bar{\kappa}\left[\frac{\nu^* \nu_{max}^{post}}{w_{max}} \frac{\bar{f}_c}{f_c(d_r)} + \frac{\nu^*}{w_{max}} + \nu^* + \nu_{max}^{post}\right] \tag{15}$$

The following parameters occur in these equations: ν^* is the output rate of neuron μ^*, ν_{min}^{post} is the minimal output rate, ν_{max}^{post} is the maximal output rate of the trained neuron, $\bar{f}_c = \int_0^{\infty} dr\ f_c(r)$ is the integral over the eligibility trace, $\bar{W} = \int_{-\infty}^{\infty} dr\ W(r)$ is the integral over the STDP learning curve (see equation (2)), $\epsilon_\kappa(r) = \int_{-\infty}^{\infty} dr'\ \kappa(r')\epsilon(r - r')$

is the convolution of the reward kernel with the shape of the postsynaptic potential (PSP) $\epsilon(s)$, and $\bar{W}_\epsilon = \int_{-\infty}^{\infty} dr\ \epsilon(r)W(r)$ is the integral over the PSP weighted by the learning window.

If these inequalities are fulfilled and input rates are larger than zero, then the weight vector of the trained neuron converges on average from any initial weight vector to $\mathbf{w}^*$ (i.e., it mimics the weight distribution of neuron μ^* for those n inputs which both have in common). To get an intuitive understanding of these inequalities, we first examine the idea behind constraint (13). This constraint assures that weights of synapses i with $w_i^* = 0$ decay to zero in expectation. First note that input spikes from a spike train S_i with $w_i^* = 0$ have no influence on the target spike train S^*. In the linear Poisson neuron model, this leads to weight changes similar to STDP which can be described by two terms. First, all synapses are subject to depression stemming from the negative part of the learning curve W and random pre-post spike pairs. This weight change is bounded from below by $\alpha\nu_i^{pre}\nu_{min}^{post}\bar{W}$ for some positive constant α. On the other hand, the positive influence of input spikes on postsynaptic firing leads to potentiation of the synapse bounded from above by $\alpha\nu_i^{pre}w_{max}\bar{W}_\epsilon$. Hence the weight decays to zero if $-\alpha\nu_i^{pre}\nu_{min}^{post}\bar{W} > \alpha\nu_i^{pre}w_{max}\bar{W}_\epsilon$, leading to inequality (13). For synapses i with $w_i^* = w_{max}$, there is an additional drive, since each presynaptic spike increases the probability of a closely following spike in the target spike train S^*. Therefore, the probability of a delayed reward signal after a presynaptic spike is larger. This additional drive leads to positive weight changes if inequalities (14) and (15) are fulfilled (see Methods).

Note that also for the learning of spike times spontaneous spikes (which might be regarded as "noise") are important, since they may lead to reward signals that can be exploited by the learning rule. It is obvious that in reward-modulated STDP, a silent neuron cannot recover from its silent state, since there will be no spikes which can drive STDP. But in addition, condition (13) shows that in this learning scenario, the minimal output rate ν_{min}^{post} – which increases with increasing noise – has to be larger than some positive constant, such that depression is strong enough to weaken synapses if needed. On the other hand, if the noise is too strong also synapses i with $w_i = w_{max}$ will be depressed and may not converge correctly. This can happen when the increased noise leads to a maximal postsynaptic rate ν_{max}^{post} such that constraints (14) and (15) are not satisfied anymore.

The conditions (13)-(15) also reveal how parameters of the model influence the applicability of this setup. For example, the eligibility trace enters the equations only in the form of its integral and its value at the reward delay in equation (15). In fact, the exact shape of the eligibility trace is not important. The important property of an ideal eligibility trace is that it is high at the reward delay and low at other times as expressed by the fraction in condition (15). Interestingly, the formulas also show that one has quite some freedom in choosing the form of the STDP window, as long as the reward kernel ϵ_κ is adjusted accordingly. For example, instead of a standard STDP learning window W with $W(r) \geq 0$ for $r > 0$ and $W(r) \leq 0$ for $r < 0$ and a corresponding reward kernel κ, one can use a reversed learning window W' defined by $W'(r) \equiv W(-r)$ and a reward kernel κ' such that $\epsilon_{\kappa'}(r) = \epsilon_\kappa(-r)$. If (15) is satisfied for W and κ, then it is also satisfied for W' and κ' (and in most cases also condition (14) will be satisfied). This reflects the fact that in reward modulated STDP the learning window defines the weight changes in combination with the reward signal.

For a given STDP learning window, the analysis reveals what reward kernels κ are suitable for this learning setup. From condition (15), we can deduce that the integral over κ should be small (but positive), whereas the integral $\int_{-\infty}^{\infty} dr\ W(r)\epsilon_\kappa(r)$ should be large. Hence, for a standard STDP learning window W with $W(r) \geq 0$ for $r > 0$ and $W(r) \leq 0$ for $r < 0$, the convolution $\epsilon_\kappa(r)$ of the reward kernel with the PSP should be positive for $r > 0$ and negative for $r < 0$. In the computer simulation we used a simple kernel depicted in Fig. 6B, which satisfies the aforementioned constraints. It consists of two double-exponential functions, one positive and one negative, with a zero crossing at some offset t_κ from the origin. The optimal offset t_κ is always negative and in the order of several milliseconds for usual PSP-shapes ϵ. We conclude that for successful learning in this scenario, a positive reward should be produced if the neuron spikes around the target spike or somewhat later, and a negative reward should be produced if the neuron spikes much too early.

Computer simulation 2: Learning spike times

In order to explore this learning scenario in a biologically more realistic setting, we trained a LIF neuron with conductance based synapses exhibiting short term facilitation and depression. The trained neuron and the neuron μ^* which produced the target spike train S^* both received inputs from 100 input neurons emitting spikes from a constant rate Poisson process of 15 Hz. The synapses to the trained neuron were subject to reward-modulated STDP. The weights of neuron μ^* were set to $w_i^* = w_{max}$ for $0 \leq i < 50$ and $w_i^* = 0$ for $50 \leq i < 100$. In order to simulate a non-realizable target response, neuron μ^* received 10 additional synaptic inputs (with weights set to $w_{max}/2$). During the simulations we observed a firing rate of 18.2 Hz for the trained neuron, and 25.2 Hz for the neuron μ^*. The simulations were run for 2 hours simulated biological time.

We performed 5 repetitions of the experiment, each time with different randomly generated inputs and different initial weight values for the trained neuron. In each of the 5 runs, the average synaptic weights of synapses with $w_i^* = w_{max}$ and $w_i^* = 0$ approached their target values, as shown in Fig. 7A. In order to test how closely the trained neuron reproduces the target spike train S^* after learning, we performed additional simulations where the same spike input was applied to the trained neuron before and after the learning. Then we compared the output of the trained neuron before and after learning with the output S^* of neuron μ^*. Fig. 7B shows that the trained neuron approximates the part of S^* which is accessible to it quite well. Panels C-F of Fig. 7 provide more detailed analyses of the evolution of weights during learning. The computer simulations confirmed the theoretical prediction that the neuron can learn well through reward-modulated STDP only if a certain level of noise is injected into the neuron (see preceding discussion and the Suppl. Fig. 6).

Both the theoretical results and these computer simulations demonstrate that a neuron can learn quite well through reward-modulated STDP to respond with specific spike patterns.

Computer simulation 3: Testing the analytically derived conditions

Equations (13) - (15) predict under which relationships between the parameters involved the learning of particular spike responses through reward-modulated STDP will be successful. We have tested these predictions by selecting 6 arbitrary settings of these parameters, which are listed in Table 1. In 4 cases (marked by light gray shading in Fig. 8) these conditions were not met (either for the learning of weights with target value w_{max}, or for the learning of weights with target value 0. Fig. 8 shows that the derived learning result is not achieved in exactly these 4 cases. On the other hand, the theoretically predicted weight changes (black bar) predict in all cases the actual weight changes (gray bar) that occur for the chosen simulation times (listed in the last column of Table 1) remarkably well.

Pattern discrimination with reward-modulated STDP

We examine here the question whether a neuron can learn through reward-modulated STDP to discriminate between two spike patterns P and N of its presynaptic neurons, by responding with more spikes to pattern P than to pattern N. Our analysis is based on the assumption that there exist internal rewards $d(t)$ that could guide such pattern discrimination. This reward based learning architecture is biologically more plausible than an architecture with a supervisor which provides for each input pattern a target output and thereby directly produces the desired firing behavior of the neuron (since the question becomes then how the supervisor has learnt to produce the desired spike outputs).

We consider a neuron that receives input from n presynaptic neurons. A pattern X consists of n spike trains, each of time length T, one for each presynaptic neuron. There are two patterns, P and N, which are presented in alternation to the neuron, with some reset time between presentations. For notational simplicity, we assume that each of the n presynaptic spike trains consists of exactly one spike. Hence, each pattern can be defined by a list of spike times: $P = (t_1^P, \ldots, t_n^P)$, $N = (t_1^N, \ldots, t_n^N)$, where t_i^X is the time when presynaptic neuron i spikes for pattern $X \in \{P, N\}$. A generalization to the easier case of learning to discriminate spatio-temporal presynaptic firing patterns (where some presynaptic neurons produce different numbers of spikes in different patterns) is straightforward, however the main characteristics of the learning dynamics are better accessible in this conceptually simpler setup. It had already been shown in [12] that neurons can learn through reward-modulated STDP to discriminate between different *spatial* presynaptic firing patterns. But in the light of the analysis of [28] it is still open whether neurons can learn with simple forms of reward-modulated STDP, such as the one considered in this article, to discriminate *temporal* presynaptic firing patterns.

We assume that the reward signal $d(t)$ rewards – after some delay d_r – action potentials of the trained neuron if pattern P was presented, and punishes action potentials of the neuron if pattern N was presented. More precisely, we assume that

$$d(t) = \begin{cases} \alpha^P \int_0^\infty dr\ \epsilon_r(r) S^{post}(t - d_r - r) & , \quad \text{if a pattern } P \text{ was presented} \\ \alpha^N \int_0^\infty dr\ \epsilon_r(r) S^{post}(t - d_r - r) & , \quad \text{if a pattern } N \text{ was presented} \end{cases} \tag{16}$$

with some reward kernel ϵ_r and constants $\alpha^N < 0 < \alpha^P$. The goal of this learning task is to produce many output spikes for pattern P, and few or no spikes for pattern N.

The main result of our analysis is an estimate of the expected weight change of synapse i of the trained neuron for the presentation of pattern P, followed after a sufficiently long time T' by a presentation of pattern N

$$\Delta w_i = \int_0^{T'} dt \left[\left\langle \frac{dw_i(t)}{dt} \right\rangle_{E|P} + \left\langle \frac{dw_i(t)}{dt} \right\rangle_{E|N} \right],$$

where $\langle \cdot \rangle_{E|X}$ is the expectation over the ensemble given that pattern X was presented. This weight change can be estimated as (see Methods)

$$\boxed{\Delta w_i = \int_{-\infty}^{\infty} dr W(r) \left[\nu^P (t_i^P + r) A_i^P + \nu^N (t_i^N + r) A_i^N \right],} \tag{17}$$

where $\nu^X(t)$ is the postsynaptic rate at time t for pattern X, and the constants A_i^X for $X \in \{P, N\}$ are given by

$$A_i^X = \alpha^X \int_0^{\infty} dr' \epsilon_r(r') \left[f_c(d_r + r') + \int_0^{T'} dt f_c(t - t_i^X) \nu^X(t - d_r - r') \right]. \tag{18}$$

As we will see shortly, an interesting learning effect is achieved if A_i^P is positive and A_i^N is negative. Since $f_c(r)$ is non-negative, a natural way to achieve this is to choose a positive reward kernel $\epsilon_r(r) \geq 0$ for $r > 0$ and $\epsilon_r(r) = 0$ for $r < 0$ (also, $f_c(r)$ and $\epsilon_r(r)$ must not be identical to zero for all r).

We use equation (17) to provide insight on when and how the classification of temporal spike patterns can be learnt with reward-modulated STDP. Assume for the moment that $A_i^N = -A_i^P$. We first note that it is impossible to achieve through any synaptic plasticity rule that the time integral over the membrane potential of the trained neuron has after training a larger value for input pattern P than for input pattern N. The reason is that each presynaptic neuron emits the same number of spikes in both patterns (namely one spike). This simple fact implies that it is impossible to train a linear Poisson neuron (with any learning method) to respond to pattern P with more spikes than to pattern N. But equation (17) implies that reward-modulated STDP increases the variance of the membrane potential for pattern P, and reduces the variance for pattern N. This can be seen as follows. Because of the specific form of the STDP learning curve $W(r)$, which is positive for (small) positive r, negative for (small) negative r, and zero for large r, $\Delta w_i = \int_{-\infty}^{\infty} dr W(r) \nu^P (t_i^P + r) A_i^P$ has a potentiating effect on synapse i if the postsynaptic rate for pattern P is larger (because of a higher membrane potential) shortly after the presynaptic spike at this synapse i than before that spike. This tends to further increase the membrane potential after that spike. On the other hand, since A_i^N is negative, the same situation for pattern N has a depressing effect on synapse i, which counteracts the increased membrane potential after the presynaptic spike. Dually, if the postsynaptic rate shortly after the presynaptic spike at synapse i is lower than shortly before that spike, the effect on synapse i is depressing for pattern P. This leads to a further decrease of the membrane potential after that spike. In the same situation for pattern N, the effect is potentiating, again counteracting the variation of the membrane potential. The total effect on the postsynaptic membrane potential is that the fluctuations for pattern P are increased, while the membrane potential for pattern N is flattened.

For the LIF neuron model, and most reasonable other non-linear spiking neuron models, as well as for biological neurons in-vivo and in-vitro [29, 30, 31], larger fluctuations of the membrane potential lead to more action potentials. As a result, reward-modulated STDP tends to increase the number of spikes for pattern P for these neuron models, while it tends to decrease the number of spikes for pattern N, thereby enabling a discrimination of these purely temporal presynaptic spike patterns.

Computer simulation 4: Learning pattern classification

We tested these theoretical predictions through computer simulations of a LIF neuron with conductance based synapses exhibiting short-term depression and facilitation. Both patterns, P and N, had 200 input channels, with 1 spike per channel (hence this is the extreme where *all* information lies in the timing of presynaptic spikes). The spike times were drawn from an uniform distribution over a time interval of 500ms, which was the duration of the patterns. We performed 1000 training trials where the patterns P and N were presented to the neuron in alternation. To introduce exploration for this reinforcement learning task, the neuron had injected 20% of the Ornstein-Uhlenbeck process conductance noise (see Methods for further details).

The theoretical analysis predicted that the membrane potential will have after learning a higher variance for pattern P, and a lower variance for pattern N. When in our simulation of a LIF neuron the firing of the neuron was switched off (by setting the firing threshold potential too high) we could observe the membrane potential fluctuations undisturbed by the reset mechanism after each spike (see Fig. 9C, D). The variance of the membrane potential did in fact increase for pattern P from $2.49(mV)^2$ to $5.43(mV)^2$ (panel C), and decrease for pattern N (panel D), from $2.34(mV)^2$ to $1.33(mV)^2$. The corresponding plots with the firing threshold included are given in panels E and F, showing an increased member of spikes of the LIF neuron for pattern P, and a decreased number of spikes for pattern N. Furthermore, as panels A and B in Fig. 9 show, the increased variance of the membrane potential for the positively reinforced pattern P led to a stable temporal firing pattern in response to pattern P.

We repeated the experiment 6 times, each time with different randomly generated patterns P and N, and different random initial synaptic weights of the neuron. The results in Fig. 9 G and H show that the learning of temporal pattern discrimination through reward-modulated STDP does not depend on the temporal patterns that are chosen, nor on the initial values of synaptic weights.

Computer simulation 5: Training a readout neuron with reward-modulated STDP to recognize isolated spoken digits

A longstanding open problem is how a biologically realistic neuron model can be trained in a biologically plausible manner to extract information from a generic cortical microcircuit. Previous work [32, 33, 34, 35, 36] has shown that quite a bit of salient information about recent and past inputs to the microcircuit can be extracted by a non-spiking linear readout neuron (i.e., a perceptron) that is trained by linear regression or margin maximization methods. Here we examine to what extent a LIF readout neuron with conductance based synapses (subject to biologically realistic short term synaptic plasticity) can learn

through reward-modulated STDP to extract from the response of a simulated cortical microcircuit (consisting of 540 LIF neurons), see Fig. 10A, the information which spoken digit (transformed into spike trains by a standard cochlea model) is injected into the circuit. In comparison with the preceding task in simulation 4, this task is easier because the presynaptic firing patterns that need to be discriminated differ in temporal and spatial aspects (see Fig. 10B; Suppl. Fig. 10 and 11 show the spike trains that were injected into the circuit). But this task is on the other hand more difficult, because the circuit response (which creates the presynaptic firing pattern for the readout neuron) differs also significantly for two utterances of the same digit (Fig. 10C), and even for two trials for the same utterance (Fig. 10D) because of the intrinsic noise in the circuit (which was modeled according to [27] to reflect in-vivo conditions during cortical UP-states). The results shown in Fig. 10E - H demonstrate that nevertheless this learning experiment was successful. On the other hand we were not able to achieve in this way speaker-independent word recognition, which had been achieved in [32] with a linear readout. Hence further work will be needed in order to clarify whether biologically more realistic models for readout neurons can be trained through reinforcement learning to reach the classification capabilities of perceptrons that are trained through supervised learning.

3 Methods

We first describe the simple neuron model that we used for the theoretical analysis, and then provide derivations of the equations that were discussed in the preceding section. After that we describe the models for neurons, synapses, and synaptic background activity ("noise") that we used in the computer simulations. Finally we provide technical details to each of the 5 computer simulations that we discussed in the preceding section.

Linear Poisson Neuron Model

In our theoretical analysis, we use a linear Poisson neuron model whose output spike train $S_j^{post}(t)$ is a realization of a Poisson process with the underlying instantaneous firing rate $R_j(t)$. The effect of a spike of presynaptic neuron i at time t' on the membrane potential of neuron j is modeled by an increase in the instantaneous firing rate by an amount $w_{ji}(t')\epsilon(t-t')$, where ϵ is a response kernel which models the time course of a postsynaptic potential (PSP) elicited by an input spike. Since STDP according to [12] has been experimentally confirmed only for excitatory synapses, we will consider plasticity only for excitatory connections and assume that $w_{ji} \geq 0$ for all i and $\epsilon(s) \geq 0$ for all s. Because the synaptic response is scaled by the synaptic weights, we can assume without loss of generality that the response kernel is normalized to $\int_0^\infty ds\ \epsilon(s) = 1$. In this linear model, the contributions of all inputs are summed up linearly:

$$R_j(t) = \sum_{i=1}^{n} \int_0^\infty ds\ w_{ji}(t-s)\ \epsilon(s)\ S_i(t-s)\ , \tag{19}$$

where $S_1, \ldots, S_n$ are the n presynaptic spike trains. Since the instantaneous firing rate $R(t)$ is analogous to the membrane potential of other neuron models, we occasionally refer to $R(t)$ as the "membrane potential" of the neuron.

Learning equations

In the following, we denote by $\langle x \rangle_{E|S_k^{post}(t),S_i^{pre}(t')}$ the ensemble average of a random variable x given that neuron k spikes at time t and neuron i spikes at time t'. We will also sometimes indicate the variables $Y_1, Y_2, \ldots$ over which the average of x is taken by writing $\langle x \rangle_{Y_1,Y_2,\ldots|\ldots}$.

Derivation of equation (6). Using equation (5), (1), and (4), we obtain the expected weight change between time t and $t+T$

$$
\begin{aligned}
&\frac{\langle w_{ji}(t+T) - w_{ji}(t) \rangle_E}{T} = \\
&\quad \int_0^\infty ds f_c(s) \int_0^\infty dr W(r) \left\langle \left\langle d(t) S_j^{post}(t-s) S_i^{pre}(t-s-r) \right\rangle_T \right\rangle_E + \\
&\quad \int_0^\infty ds\ f_c(s) \int_{-\infty}^0 dr\ W(r) \left\langle \left\langle d(t) S_j^{post}(t-s+r) S_i^{pre}(t-s) \right\rangle_T \right\rangle_E \\
&= \int_0^\infty dr\ W(r) \int_0^\infty ds\ f_c(s) \left\langle \left\langle d(t) S_j^{post}(t-s) S_i^{pre}(t-s-r) \right\rangle_E \right\rangle_T + \\
&\int_{-\infty}^0 dr\ W(r) \int_{|r|}^\infty ds\ f_c(s+r) \left\langle \left\langle d(t) S_j^{post}(t-s) S_i^{pre}(t-s-r) \right\rangle_E \right\rangle_T \\
&\quad = \int_0^\infty dr\ W(r) \int_0^\infty ds\ f_c(s) \left\langle D_{ji}(t,s,r)\ \nu_{ji}(t-s,r) \right\rangle_T + \\
&\qquad \int_{-\infty}^0 dr\ W(r) \int_{|r|}^\infty ds\ f_c(s+r) \left\langle D_{ji}(t,s,r)\ \nu_{ji}(t-s,r) \right\rangle_T ,
\end{aligned}
$$

with $D_{ji}(t,s,r) = \langle d(t) |$ Neuron j spikes at $t-s$, and neuron i spikes at $t-s-r \rangle_E$, and the joint firing rate $\nu_{ji}(t,r) = \langle S_j(t) S_i(t-r) \rangle_E$ describes correlations between spike timings of neurons j and i. The joint firing rate $\nu_{ji}(t-s,r)$ depends on the weight at time $t-s$. If the learning rate defined by the magnitude of $W(r)$ is small, the synaptic weights can be assumed constant on the time scale of T. Thus, the time scales of neuronal dynamics are separated from the slow time scale of learning. For slow learning, synaptic weights integrate a large number of small changes. We can then expect that averaged quantities enter the learning dynamics. In this case, we can argue that fluctuations of a weight w_{ji} about its mean are negligible and it can well be approximated by its average $\langle w_{ji} \rangle_E$ (it is "self-averaging", see [22, 37]). To ensure that average quantities enter the learning dynamics, many presynaptic and postsynaptic spikes as well as many independently delivered rewards at varying delays have to occur within T. Hence, in general, the time scale of single spike occurrences and the time scale of the eligibility trace is required to be much smaller than the time scale of learning. If time scales can be separated, we can drop the expectation on the left hand side of the last equation and write

$$
\frac{\langle w_{ji}(t+T) - w_{ji}(t) \rangle_E}{T} = \frac{w_{ji}(t+T) - w_{ji}(t)}{T} = \frac{1}{T} \int_t^{t+T} \frac{d}{dt} w_{ji}(t') dt' = \frac{d}{dt} \langle w_{ji}(t) \rangle_T .
$$

We thus obtain equation (6):

$$\begin{aligned}\frac{d}{dt}\langle w_{ji}(t)\rangle_T &= \int_0^\infty dr\, W(r) \int_0^\infty ds\, f_c(s)\, \langle D_{ji}(t,s,r)\, \nu_{ji}(t-s,r)\rangle_T \\ &+ \int_{-\infty}^0 dr\, W(r) \int_{|r|}^\infty ds\, f_c(s+r)\, \langle D_{ji}(t,s,r)\, \nu_{ji}(t-s,r)\rangle_T\,.\end{aligned}$$

Simplification of equation (6). In order to simplify this equation, we first observe that $W(r)$ is vanishing for large $|r|$. Hence we can approximate the integral over the learning window by a bounded integral $\int_{-\infty}^{\infty} dr\; W(r) \approx \int_{-T_W}^{T_W} dr\; W(r)$ for some $T_W > 0$ and $T_W \ll T$. In the analyzes of this article, we consider the case where reward is delivered with a relatively large temporal delay. To be more precise, we assume that a pre-post spike pair has an effect on the reward signal only after some minimal delay d_r and that we can write $D_{ji}(t,s,r) = d_0 + D_{ji}^{pre,post}(t,s,r)$ for some baseline reward d_0 and a part which depends on the timing of pre-post spike pairs with $D_{ji}^{pre,post}(t,s,r) = 0$ for $s < d_r$ and $d_r > T_W$. We can then approximate the second term of equation (6):

$$\begin{aligned}&\int_{-\infty}^0 dr\, W(r) \int_{|r|}^\infty ds\, f_c(s+r)\, \langle D_{ji}(t,s,r)\, \nu_{ji}(t-s,r)\rangle_T \\ &\quad\approx \int_{-T_W}^0 dr\, W(r) \int_{|r|}^\infty ds\, f_c(s+r)\, \left\langle (d_0 + D_{ji}^{pre,post}(t,s,r))\, \nu_{ji}(t-s,r)\right\rangle_T \\ &\quad\quad\approx \int_{-T_W}^0 dr\, W(r) \left[\int_0^\infty ds\, f_c(s) d_0\, \langle \nu_{ji}(t-s,r)\rangle_T \right. \\ &\quad\quad\quad \left. + \int_{|r|}^\infty ds\, f_c(s+r)\, \left\langle D_{ji}^{pre,post}(t,s,r)\, \nu_{ji}(t-s,r)\right\rangle_T\right]\end{aligned}$$

because $\langle \nu_{ji}(t-s-r,r)\rangle_T \approx \langle \nu_{ji}(t-s,r)\rangle_T$ for $r \in [-T_W, T_W]$ and $T_W \ll T$. Since $D_{ji}^{pre,post}(t,s,r) = 0$ for $s \le T_W$, the second term in the brackets is equivalent to $\int_0^\infty ds\, f_c(s+r) \left\langle D_{ji}^{pre,post}(t,s,r)\, \nu_{ji}(t-s,r)\right\rangle_T$ which in turn is approximately given by $\int_0^\infty ds\, f_c(s) \left\langle D_{ji}^{pre,post}(t,s,r)\, \nu_{ji}(t-s,r)\right\rangle_T$ if we assume that $f_c(s+r) \approx f_c(s)$ for $s \ge d_r$ and $|r| < T_W$. We can thus approximate the second term of equation (6) as

$$\begin{aligned}&\int_{-\infty}^0 dr\, W(r) \int_{|r|}^\infty ds\, f_c(s+r)\, \langle D_{ji}(t,s,r)\, \nu_{ji}(t-s,r)\rangle_T \\ &\quad\quad\approx \int_{-T_W}^0 dr\, W(r) \left[\int_0^\infty ds\, f_c(s) d_0\, \langle \nu_{ji}(t-s,r)\rangle_T \right. \\ &\quad\quad\quad \left. + \int_0^\infty ds\, f_c(s)\, \left\langle D_{ji}^{pre,post}(t,s,r)\, \nu_{ji}(t-s,r)\right\rangle_T\right] \\ &\quad\quad\quad\approx \int_{-\infty}^0 dr\, W(r) \int_0^\infty ds\, f_c(s)\, \langle D_{ji}(t,s,r)\, \nu_{ji}(t-s,r)\rangle_T\,.\end{aligned}$$

With this approximation, the first and second term of equation (6) can be combined in a single integral to obtain equation (8).

Derivations for the biofeedback experiment

We assume that a reward with the functional form ϵ_r is delivered for each postsynaptic spike with a delay d_r. The reward as time t is therefore

$$d(t) = \int_0^\infty dr\ S_k^{post}(t - d_r - r)\epsilon_r(r).$$

Weight change for the reinforced neuron (derivation of equation (10)). The reward correlation for a synapse ki afferent to the reinforced neuron is

$$\begin{aligned} D_{ki}(t,s,r) &= \langle d(t)\rangle_{E|S_k^{post}(t-s),S_i^{pre}(t-s-r)} \\ &= \int_0^\infty dr'\ \epsilon_r(r')\langle S_k^{post}(t - d_r - r')\rangle_{E|S_k^{post}(t-s),S_i^{pre}(t-s-r)} \\ &= \int_0^\infty dr'\ \epsilon_r(r')\left[\nu_k(t - d_r - r') + w_{ki}\epsilon(s + r - d_r - r')\right. \\ &\qquad\qquad \left. +\delta(s - d_r - r')\right] \\ = \int_0^\infty dr'\epsilon_r(r')\nu_k(t - d_r - r') &+ w_{ki}\int_0^\infty dr'\ \epsilon_r(r')\epsilon(s + r - d_r - r') + \epsilon_r(s - d_r). \end{aligned}$$

If we assume that the output firing rate is constant on the time scale of the reward function, the first term vanishes. We rewrite the result as

$$D_{ki}(t,s,r) = \epsilon_r(s - d_r) + w_{ki}\int_{-\infty}^\infty dr'\ \epsilon_r(s - d_r + r')\epsilon(r - r').$$

The mean weight change for weights to the reinforced neuron is therefore

$$\begin{aligned} \frac{d}{dt}w_{ki}(t) = \int_{-\infty}^\infty dr\ W(r)\left(\int_0^\infty ds\ f_c(s)\epsilon_r(s - d_r)\left\langle\nu_{ki}(t - s, r)\right\rangle_T + \right. \\ \left. w_{ki}\int_{-\infty}^\infty dr'\ \epsilon(r - r')\int_0^\infty ds\ f_c(s)\epsilon_r(s - d_r + r')\left\langle\nu_{ki}(t - s, r)\right\rangle_T\right). \quad (20) \end{aligned}$$

We show that the second term in the brackets is very small compared to the first term:

$$\begin{aligned} &w_{ki}\int_{-\infty}^\infty dr'\ \epsilon(r - r')\int_0^\infty ds\ f_c(s)\epsilon_r(s - d_r + r')\left\langle\nu_{ki}(t - s, r)\right\rangle_T = \\ &w_{ki}\int_{-\infty}^\infty dr'\ \epsilon(r - r')\int_0^\infty ds\ f_c(s - r')\epsilon_r(s - d_r)\left\langle\nu_{ki}(t - s - r', r)\right\rangle_T \approx \\ &w_{ki}\int_{-\infty}^\infty dr'\ \epsilon(r - r')\int_0^\infty ds\ f_c(s)\epsilon_r(s - d_r)\left\langle\nu_{ki}(t - s, r)\right\rangle_T. \end{aligned}$$

The last approximation is based on the assumption that $f_c(s) \approx f_c(s - r')$ and $\langle\nu_{ki}(t - r', r)\rangle_T \approx \langle\nu_{ki}(t, r)\rangle_T$ for $r' \in [-T_W - T_\epsilon, T_W]$. Here, T_W is the time scale of the learning window (see above), and T_ϵ is time scale of the PSP, i.e., we have $\epsilon(s) \approx 0$ for $s \geq T_\epsilon$. Since $\int_{-\infty}^\infty dr\ \epsilon(r) = 1$ by definition, we see that this is the first term in the brackets of equation (20) scaled by w_{ki}. For neurons with many input synapses we have $w_{ki} \ll 1$.

Thus the second term in the brackets of equation (20) is small compared to the first term. We therefore have

$$\frac{d}{dt}w_{ki}(t) \approx \int_0^\infty ds\; f_c(s+d_r)\epsilon_r(s) \int_{-\infty}^\infty dr\; W(r)\; \langle \nu_{ki}(t-d_r-s,r)\rangle_T\,.$$

Weight change for non-reinforced neurons (derivation of equation (11)). The reward correlation of a synapse ji to a non-reinforced neuron j is given by

$$\begin{aligned} D_{ji}(t,s,r) &= \langle d(t)\rangle_{E|S_j^{post}(t-s),S_i^{pre}(t-s-r)} \\ &= \int_0^\infty dr'\; \epsilon_r(r')\langle S_k^{post}(t-d_r-r')\rangle_{E|S_j^{post}(t-s),S_i^{pre}(t-s-r)}. \end{aligned}$$

We have

$$\begin{aligned} &\langle S_k^{post}(t-d_r-r')\rangle_{E|S_j^{post}(t-s),S_i^{pre}(t-s-r)} \\ &\quad= \frac{\langle S_k^{post}(t-d_r-r')S_j^{post}(t-s)\rangle_{E|S_i^{pre}(t-s-r)}}{\langle S_j^{post}(t-s)\rangle_{E|S_i^{pre}(t-s-r)}} \\ &\quad= \frac{\nu_{kj}(t-d_r-r',s-d_r-r') + w_{ki}w_{ji}\epsilon(s+r-d_r-r')\epsilon(r)}{\nu_j(t-s)+w_{ji}\epsilon(r)}, \end{aligned}$$

for which we obtain

$$D_{ji}(t,s,r) = \int_0^\infty dr'\; \epsilon_r(r')\frac{\nu_{kj}(t-d_r-r',s-d_r-r') + w_{ki}w_{ji}\epsilon(s+r-d_r-r')\epsilon(r)}{\nu_j(t-s)+w_{ji}\epsilon(r)}.$$

In analogy to the previous derivation, we assume here that the firing rate $\nu_j(t-s)$ in the denominator results from many PSPs. Hence, the single PSP $w_{ji}\epsilon(r)$ is small compared to $\nu_j(t-s)$. Similarly, we assume that with weights $w_{ki}, w_{ji} \ll 1$, the second term in the nominator is small compared to the joint firing rate $\nu_{kj}(t-d_r-r',s-d_r-r')$. We therefore approximate the reward correlation by

$$D_{ji}(t,s,r) \approx \int_0^\infty dr'\; \epsilon_r(r')\frac{\nu_{kj}(t-d_r-r',s-d_r-r')}{\nu_j(t-s)}.$$

Hence, the reward correlation of a non-reinforced neuron depends on the correlation of this neuron with the reinforced neuron. The mean weight change for a non-reinforced neuron $j \neq k$ is therefore

$$\frac{d}{dt}w_{ji}(t) \approx \int_0^\infty ds\; f_c(s) \int_{-\infty}^\infty dr\; W(r) \int_0^\infty dr' \epsilon_r(r') \left\langle \frac{\nu_{kj}(t-d_r-r',s-d_r-r')}{\nu_j(t-s)}\nu_{ji}(t-s,r)\right\rangle_T$$

This equation deserves a remark for the case that $\nu_j(t-s)$ is zero, since it appears in the denominator of the fraction. Note that in this case, both $\nu_{kj}(t-d_r-r',s-d_r-r')$ and $\nu_{ji}(t-s,r)$ are zero. In fact, if we take the limit $\nu_j(t-s)\to 0$, then both of these factors approach zero at least as fast. Hence, in the limit of $\nu_j(t-s)\to 0$, the term in the angular brackets evaluates to zero. This reflects the fact that since STDP is driven by pre- and postsynaptic spikes, there is no weight change if no postsynaptic spikes occur.

For uncorrelated neurons, equation 11 evaluates to zero. For uncorrelated neurons k, j, $\nu_{kj}(t-d_r-r', s-d_r-r')$ can be factorized into $\nu_k(t-d_r-r')\nu_j(t-s)$, and we obtain

$$\frac{d}{dt}w_{ji}(t) \approx \int_0^\infty ds\ f_c(s) \int_{-\infty}^\infty dr\ W(r) \int_0^\infty dr' \epsilon_r(r')\ \langle \nu_k(t-d_r-r')\nu_{ji}(t-s,r)\rangle_T\,.$$

This evaluates approximately to zero if the mean output rate of neuron k is constant on the time scale of the reward kernel.

Analysis of spike-timing dependent rewards (derivation of the conditions (13)-(15)).

Below, we will indicate the variables $Y_1, Y_2, \ldots$ over which the average of x is taken by writing $\langle x\rangle_{Y_1,Y_2,\ldots|\ldots}$. From equation (12), we can determine the reward correlation for synapse i

$$\begin{aligned} D_{ji}(t,s,r) &= \int_{-\infty}^\infty dr'\kappa(r') \left\langle S_j^{post}(t-d_r)S^*(t-d_r-r')\right\rangle_{E|S_j^{post}(t-s),S_i^{pre}(t-s-r)} \\ &= \int_{-\infty}^\infty dr'\kappa(r') \left[\nu_j^{post}(t-d_r) + \delta(s-d_r) + w_{ji}(s+r-d_r)\epsilon(s+r-d_r)\right] \\ &\qquad \left[\nu^*(t-d_r-r') + w_i^*\epsilon(s+r-d_r-r')\right], \quad (21) \end{aligned}$$

where $\nu_j^{post}(t) = \langle S_j^{post}(t)\rangle_E$ denotes the instantaneous firing rate of the trained neuron at time t, and $\nu^*(t) = \langle S^*(t)\rangle_E$ denotes the instantaneous rate of the target spike train at time t. Since weights are changing very slowly, we have $w_{ji}(t-s-r) \approx w_{ji}(t)$. In the following, we will drop the dependence of w_{ji} on t for brevity. For simplicity, we assume that input rates are stationary and uncorrelated. In this case (since the weights are changing slowly), also the correlations between inputs and outputs can be assumed stationary, $\nu_{ji}(t,r) = \nu_{ji}(r)$. With constant input rates, we can rewrite (21) as

$$\begin{aligned} D_{ji}(t,s,r) = \bar\kappa\nu^*\nu_j^{post} &+ \bar\kappa\nu^*\delta(s-d_r) + \bar\kappa\nu^* w_{ji}\epsilon(s+r-d_r) \\ &+ w_i^* \int_{-\infty}^\infty dr'\kappa(r')\epsilon(s+r-d_r-r') \\ &\left[\nu_j^{post}(t-d_r) + \delta(s-d_r) + w_{ji}(s+r-d_r)\epsilon(s+r-d_r)\right], \end{aligned}$$

with $\bar\kappa = \int_{-\infty}^\infty ds\ \kappa(s)$. We use this results to obtain the temporally smoothed weight change for synapse ji. With stationary correlations, we can drop the dependence of ν_{ji} on

t and write $\nu_{ji}(t,r) = \nu_{ji}(r)$. Furthermore, we define $\nu_{ji}^W(r) = \nu_{ji}(r)W(r)$ and obtain

$$\begin{aligned}
\frac{d}{dt}w_{ji}(t) &= \int_{-\infty}^{\infty} dr\; W(r)\nu_{ji}(r) \int_0^{\infty} ds\; f_c(s)\,\langle D_{ji}(t,s,r)\rangle_T \\
&= \int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r)\bar{\kappa}\left[\nu^*\nu_j^{post}\bar{f}_c + \nu^* f_c(d_r)\right. \\
&\quad \left. +\nu^* w_{ji}\int_0^{\infty} ds\; f_c(s)\epsilon(s+r-d_r)\right] + \\
&\quad \int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r) w_i^*\nu^{post}\int_{-\infty}^{\infty} dr'\kappa(r')\int_0^{\infty} ds\; f_c(s)\epsilon(s+r-d_r-r') + \\
&\quad \int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r) w_i^*\int_{-\infty}^{\infty} dr'\kappa(r') f_c(d_r)\epsilon(r-r') + \\
&\quad \int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r) w_i^*\int_{-\infty}^{\infty} dr'\kappa(r') w_{ji}\int_0^{\infty} ds\; f_c(s)\epsilon(s+r-d_r)\epsilon(s+r-d_r-r').
\end{aligned}$$

We assume that the eligibility function $f_c(d_r) \approx f_c(d_r + r)$ if $|r|$ is on the time scale of a PSP, the learning window, or the reward kernel, and that d_r is large compared to these time scales. Then, we have

$$\int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r)\int_{-\infty}^{\infty} dr'\;\kappa(r') f_c(d_r)\epsilon(r-r') = f_c(d_r)\int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r)\epsilon_\kappa(r)$$

where $\epsilon_\kappa(r) = \int_{-\infty}^{\infty} dr'\;\kappa(r')\epsilon(r-r')$ is the convolution of the reward kernel with the PSP. Furthermore, we find

$$\begin{aligned}
&\int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r)\int_{-\infty}^{\infty} dr'\;\kappa(r')\int_0^{\infty} ds\; f_c(s)\epsilon(s+r-d_r)\epsilon(s+r-d_r-r') \\
&\quad \approx f_c(d_r)\int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r)\int_{-\infty}^{\infty} dr'\;\kappa(r')\int_0^{\infty} ds\;\epsilon(s+r-d_r)\epsilon(s+r-d_r-r') \\
&\quad = f_c(d_r)\int_{-\infty}^{\infty} dr\; \nu_{ji}^W(r)\int_0^{\infty} ds\;\epsilon(s)\epsilon_\kappa(s).
\end{aligned}$$

With these simplifications, and the abbreviation $\bar{\nu}_{ji}^W = \int_{-\infty}^{\infty} dr\nu_{ji}^W(r)$ we obtain the weight change at synapse ji

$$\begin{aligned}
\frac{d}{dt}w_{ji}(t) &\approx \bar{\kappa}\nu^*\nu_j^{post}\bar{\nu}_{ji}^W\bar{f}_c + f_c(d_r)\bar{\kappa}\bar{\nu}_{ji}^W\left[\nu^* + \nu^* w_{ji} + w_i^*\nu_j^{post}\right] \\
&\quad + f_c(d_r)w_i^*\int_{-\infty}^{\infty} drW(r)\nu_{ji}(r)\epsilon_\kappa(r) + f_c(d_r)w_{ji}w_i^*\bar{\nu}_{ji}^W\int_{-\infty}^{\infty} dr\;\epsilon(r)\epsilon_\kappa(r),
\end{aligned}$$

where $\bar{\nu}_{ji}^W = \int_{-\infty}^{\infty} drW(r)\nu_{ji}(r)$.

For uncorrelated Poisson input spike trains of rate ν_i^{pre} and the linear Poisson neuron model, the input-output correlations are $\nu_{ji}(r) = \nu_i^{pre}\nu_j^{post} + w_{ji}\nu_i^{pre}\epsilon(r)$. With these correlations, we obtain $\bar{\nu}_{ji}^W = \nu_i^{pre}\nu_j^{post}\bar{W} + w_{ji}\nu_i^{pre}\bar{W}_\epsilon$ where $\bar{W} = \int_{-\infty}^{\infty} dr\; W(r)$, and

$\bar{W}_\epsilon = \int_{-\infty}^{\infty} dr\ \epsilon(r) W(r)$. The weight change at synapse ji is then

$$\begin{aligned}
\frac{d}{dt} w_{ji}(t) \approx\ & \bar{\kappa} \bar{f}_c \nu^* \nu_i^{pre} \nu_j^{post} \left[\nu_j^{post} \bar{W} + w_{ji} \bar{W}_\epsilon \right] \\
& + \bar{\kappa} f_c(d_r) \nu_i^{pre} \left[\nu_j^{post} \bar{W} + w_{ji} \bar{W}_\epsilon \right] \left[\nu^* + \nu^* w_{ji} + w_i^* \nu_j^{post} \right] \\
& + f_c(d_r) w_i^* \nu_i^{pre} \left[\nu_j^{post} \int_{-\infty}^{\infty} dr\ W(r) \epsilon_\kappa(r) + w_{ji} \int_{-\infty}^{\infty} dr\ W(r) \epsilon(r) \epsilon_\kappa(r) \right] \\
& + f_c(d_r) w_i^* w_{ji} \nu_i^{pre} \left[\nu_j^{post} \bar{W} + w_{ji} \bar{W}_\epsilon \right] \int_0^{\infty} dr\ \epsilon(r) \epsilon_\kappa(r), \qquad (22)
\end{aligned}$$

We will now bound the expected weight change for synapses ji with $w_i^* = w_{max}$ and for synapses jk with $w_k^* = 0$. In this way we can derive conditions for which the expected weight change for the former synapses is positive, and that for the latter type is negative. First, we assume that the integral over the reward kernel is positive. In this case, the weight change given by (22) is negative for synapses i with $w_i^* = 0$ if and only if $\nu_i^{pre} > 0$, and $-\nu_j^{post} \bar{W} > w_{ji} \bar{W}_\epsilon$. In the worst case, w_{ji} is w_{max} and ν_j^{post} is small. We have to guarantee some minimal output rate ν_{min}^{post} such that even if $w_{ji} = w_{max}$, this inequality is fulfilled. This could be guaranteed by some noise current. Given such minimal output rate, we can state the first inequality which guarantees convergence of weights w_{ji} with $w_i^* = 0$

$$-\nu_{min}^{post} \bar{W} > w_{max} \bar{W}_\epsilon.$$

For synapses ji with $w_i^* = w_{max}$, we obtain two more conditions. The approximate weight change is given by

$$\begin{aligned}
\frac{d}{dt} w_{ji}(t) \frac{1}{\nu_i^{pre}} \approx\ & \bar{\kappa} \left[\nu_j^{post} \bar{W} + w_{ji} \bar{W}_\epsilon \right] \\
& \left[\nu^* \nu_j^{post} \bar{f}_c + f_c(d_r) \nu^* + f_c(d_r) \nu^* w_{ji} + f_c(d_r) \nu_j^{post} w_{max} \right] \\
& + f_c(d_r) w_{max} \nu_j^{post} \int_{-\infty}^{\infty} dr\ W(r) \epsilon_\kappa(r) \\
& + f_c(d_r) w_{max} w_{ji} \int_{-\infty}^{\infty} dr\ W(r) \epsilon(r) \epsilon_\kappa(r) \\
& + f_c(d_r) w_{max} w_{ji} \nu_j^{post} \bar{W} \int_0^{\infty} dr\ \epsilon(r) \epsilon_\kappa(r) \\
& + f_c(d_r) w_{max} w_{ji}^2 \bar{W}_\epsilon \int_0^{\infty} dr\ \epsilon(r) \epsilon_\kappa(r).
\end{aligned}$$

The last term in this equation is positive and small. We can ignore it in our sufficient condition. The second to last term is negative. We will include in our condition that the third to last term compensates for this negative term. Hence, the second condition is

$$\int_{-\infty}^{\infty} dr\ W(r) \epsilon(r) \epsilon_\kappa(r) \geq -\nu_j^{post} \bar{W} \int_0^{\infty} dr\ \epsilon(r) \epsilon_\kappa(r),$$

which should be satisfied in most setups. If we assume that this holds, we obtain

$$\begin{aligned}
\frac{d}{dt} w_{ji}(t) \geq\ & \bar{\kappa} \left[\nu_j^{post} \bar{W} + w_{ji} \bar{W}_\epsilon \right] \left[\nu^* \nu_j^{post} \bar{f}_c + f_c(d_r) \nu^* + f_c(d_r) \nu^* w_{ji} + f_c(d_r) \nu_j^{post} w_{max} \right] \\
& + f_c(d_r) w_{max} \nu_j^{post} \int_{-\infty}^{\infty} dr\ W(r) \epsilon_\kappa(r).
\end{aligned}$$

which should be positive. We obtain the following inequality

$$\int_{-\infty}^{\infty} dr\, W(r)\epsilon_\kappa(r) > -\bar{W}\bar{\kappa}\left[\frac{\nu^*\nu_j^{post}}{w_{max}}\frac{\bar{f}_c}{f_c(d_r)} + \frac{\nu^*}{w_{max}} + \nu^* + \nu^{post}\right].$$

All three inequalities are summarized in the following:

$$\begin{aligned}
-\nu_{min}^{post}\bar{W} &> w_{max}\bar{W}_\epsilon \\
\int_{-\infty}^{\infty} dr\, W(r)\epsilon(r)\epsilon_\kappa(r) &\geq -\nu_{max}^{post}\bar{W}\int_0^{\infty} dr\, \epsilon(r)\epsilon_\kappa(r) \\
\int_{-\infty}^{\infty} dr\, W(r)\epsilon_\kappa(r) &> -\bar{W}\bar{\kappa}\left[\frac{\nu^*\nu_{max}^{post}}{w_{max}}\frac{\bar{f}_c}{f_c(d_r)} + \frac{\nu^*}{w_{max}} + \nu^* + \nu_{max}^{post}\right],
\end{aligned}$$

where ν_{max}^{post} is the maximal output rate. If these inequalities are fulfilled and input rates are positive, then the weight vector converges on average from any initial weight vector to $\mathbf{w}^*$. The second condition is less severe, and should be easily fulfilled in most setups. If this is the case, the first condition (13) ensures that weights with $w^* = 0$ are depressed while the third condition (15) ensures that weights with $w^* = w_{max}$ are potentiated.

Analysis of the pattern discrimination task (derivation of equation (17)).

We assume that a trial consists of the presentation of a single pattern starting at time $t = 0$. We compute the weight change for a single trial given that pattern $X \in \{P, N\}$ was presented with the help of equations (1), (3), and (4) as

$$\begin{aligned}
\left.\frac{d}{dt}w_i(t)\right|_X &= \int_0^{\infty} ds f_c(s)\left[\int_0^{\infty} drW(r)S^{post}(t-s)\delta(t-s-r-t_i^X)\right. \\
&+ \left.\int_0^{\infty} drW(-r)S^{post}(t-s-r)\delta(t-s-t_i^X)\right] d(t) \\
&= \alpha^X\int_0^{\infty} ds f_c(s)\left[\int_0^{\infty} drW(r)S^{post}(t-s)\delta(t-s-r-t_i^X)\right. \\
&+ \left.\int_0^{\infty} drW(-r)S^{post}(t-s-r)\delta(t-s-t_i^X)\right]\int_0^{\infty} dr'\epsilon_r(r')S^{post}(t-d_r-r') \\
&= \alpha^X\int_0^{\infty} dr f_c(t-r-t_i^X)W(r)\int_0^{\infty} dr'\epsilon_r(r')S^{post}(r+t_i^X)S^{post}(t-d_r-r') \\
&+ \alpha^X\int_0^{\infty} dr f_c(t-t_i^X)W(-r)\int_0^{\infty} dr'\epsilon_r(r')S^{post}(t_i^X-r)S^{post}(t-d_r-r').
\end{aligned}$$

We can compute the average weight change given that pattern X was presented:

$$\begin{aligned}\left\langle \frac{d}{dt}w_i(t)\right\rangle_{E|X} &= \alpha^X \int_0^\infty dr\, f_c(t-r-t_i^X) \\ &\quad W(r)\int_0^\infty dr' \epsilon_r(r')\langle S^{post}(t_i^X+r)S^{post}(t-d_r-r')\rangle_{E|X} \\ &+ \alpha^X \int_0^\infty dr\, f_c(t-t_i^X) \\ &\quad W(-r)\int_0^\infty dr' \epsilon_r(r')\langle S^{post}(t_i^X-r)S^{post}(t-d_r-r')\rangle_{E|X}.\end{aligned}$$

If we assume that f_c is approximately constant on the time scale of the learning window W, we can simplify this to

$$\left\langle \frac{d}{dt}w_i(t)\right\rangle_{E|X} = \int_{-\infty}^\infty dr\, f_c(t-t_i^X)W(r)\int_0^\infty dr' \epsilon_r(r')\langle S^{post}(t_i^X+r)S^{post}(t-d_r-r')\rangle_{E|X}\alpha^X.$$

For the linear Poisson neuron, we can write the auto-correlation function as

$$\begin{aligned}\langle S^{post}(t_i^X+r)S^{post}(t-d_r-r')\rangle_{E|X} &= [\nu^X(t_i^X+r)\nu^X(t-d_r-r') \\ &\quad +\nu^X(t_i^X+r)\delta(t_i^X+r-t+d_r+r')] \\ &= \nu^X(t_i^X+r)[\nu^X(t-d_r-r') + \\ &\quad \delta(t_i^X+r-t+d_r+r')],\end{aligned}$$

where $\nu^X(t) = \langle S^{post}(t)\rangle_{E|X}$ is the ensemble average rate at time t given that pattern X was presented. If an experiment for a single pattern runs over the time interval $[0, T']$, we can compute the total average weight change Δw_i^X of a trial given that pattern X was presented as

$$\begin{aligned}\Delta w_i^X &= \int_0^{T'} dt \left\langle \frac{d}{dt}w_i(t)\right\rangle_{E|X} \\ &= \alpha^X \int_{-\infty}^\infty dr W(r)\nu^X(t_i^X+r)\int_0^{T'} dt f_c(t-t_i^X)\int_0^\infty dr'\epsilon_r(r') \\ &\quad [\nu^X(t-d_r-r') + \delta(t_i^X+r-t+d_r+r')] \\ &= \alpha^X \int_{-\infty}^\infty dr W(r)\nu^X(t_i^X+r)\int_0^\infty dr'\epsilon_r(r') \\ &\quad \left[f_c(r+d_r+r') + \int_{d_r}^{T'} dt f_c(t-t_i^X)\nu^X(t-d_r-r')\right] \\ &\approx \alpha^X \int_{-\infty}^\infty dr W(r)\nu^X(t_i^X+r)\int_0^\infty dr'\epsilon_r(r') \\ &\quad \left[f_c(d_r+r') + \int_0^{T'} dt f_c(t-t_i^X)\nu^X(t-d_r-r')\right] \qquad (23)\end{aligned}$$

By defining

$$A_i^X = \alpha^X \int_0^\infty dr' \epsilon_r(r') \left[f_c(d_r + r') + \int_0^{T'} dt f_c(t - t_i^X) \nu^X (t - d_r - r') \right],$$

we can write equation (23) as

$$\Delta w_i^X = \int_{-\infty}^{\infty} dr W(r) \nu^X (t_i^X + r) A_i^X.$$

We assume that eligibility traces and reward signals have settled to zero before a new pattern is presented. The expected weight change for the successive presentation of both patterns is therefore

$$\Delta w_i = \int_{-\infty}^{\infty} dr W(r) \left[\nu^P (t_i^P + r) A_i^P + \nu^N (t_i^N + r) A_i^N \right].$$

The equations can easily be generalized to the case where multiple input spikes per synapse are allowed and where jitter on the templates is allowed. However, the main effect of the rule can be read off the equations given here.

Common models and parameters of the computer simulations

We describe here the models and parameter values that were used in all our computer simulations. We will specify in a subsequent section the values of other parameters that had to be chosen differently in individual computer simulations, in dependence of their different setups and requirements.

LIF neuron model

For the computer simulations LIF neurons with conductance-based synapses were used. The membrane potential $V_m(t)$ of this neuron model is given by:

$$C_m \frac{dV_m(t)}{dt} = -\frac{V_m(t) - V_{resting}}{R_m} - \sum_{j=1}^{K_e} g_{e,j}(t)(V_m(t) - E_e) - \sum_{j=1}^{K_i} g_{i,j}(V_m(t) - E_i) - I_{noise}(t), \quad (24)$$

where C_m is the membrane capacitance, R_m is the membrane resistance, $V_{resting}$ is the resting potential, and $g_{e,j}(t)$ and $g_{i,j}(t)$ are the K_e and K_i synaptic conductances from the excitatory and inhibitory synapses respectively. The constants E_e and E_i are the reversal potentials of excitatory and inhibitory synapses. I_{noise} represents the synaptic background current which the neuron receives (see below for details).

Whenever the membrane potential reaches a threshold value V_{thresh}, the neuron produces a spike, and its membrane potential is reset to the value of the reset potential V_{reset}. After a spike, there is a refractory period of length $T_{refract}$, during which the membrane potential of the neuron remains equal to the value $V_m(t) = V_{reset}$. After the refractory period $V_m(t)$ continues to change according to equation (24).

For a given synapse, the dynamics of the synaptic conductance $g(t)$ is defined by

$$\frac{dg(t)}{dt} = -\frac{g(t)}{\tau_{syn}} + \sum_k A(t^{(k)} + t_{delay})\delta(t - t^{(k)} - t_{delay}) \quad , \tag{25}$$

where $A(t)$ is the amplitude of the postsynaptic response (PSR) to a single presynaptic spike, which varies over time due to the inherent short-term dynamics of the synapse, and $\{t^{(k)}\}$ are the spike times of the presynaptic neuron. The conductance of the synapse decreases exponentially with time constant τ_{syn}, and increases instantaneously by amount of $A(t)$ whenever the presynaptic neuron spikes.

In all computer simulations we used the following values for the neuron and synapse parameters. The membrane resistance of the neurons was $R_m = 100M\Omega$, the membrane capacitance $C_m = 0.3nF$, the resting potential, reset potential and the initial value of the membrane potential had the same value of $V_{resting} = V_{reset} = V_m(0) = -70$mV, the threshold potential was set to $V_{thresh} = -59mV$ and the refractory period $T_{refract} = 5$ms. For the synapses we used a time constant set to $\tau_{syn} = 5$ms, reversal potential $E_e = 0$ mV for the excitatory synapses and $E_e = -75$ mV for the inhibitory synapses. All synapses had a synaptic delay of $t_{delay} = 1$ms.

Short-term dynamics of synapses

We modeled the short-term dynamics of synapses according to the phenomenological model proposed in [38], where the amplitude $A_k = A(t_k + t_{delay})$ of the postsynaptic response for the k^{th} spike in a spike train with inter-spike intervals $\Delta_1, \Delta_2, \ldots, \Delta_{k-1}$ is calculated with the following equations

$$\begin{aligned} A_k &= w \cdot u_k \cdot R_k \\ u_k &= U + u_{k-1}(1-U)e^{-\Delta_{k-1}/F} \\ R_k &= 1 + (R_{k-1} - u_{k-1}R_{k-1} - 1)e^{-\Delta_{k-1}/D}, \end{aligned} \tag{26}$$

with hidden dynamic variables $u \in [0,1]$ and $R \in [0,1]$ whose initial values for the 1^{st} spike are $u_1 = U$ and $R = 1$ (see [39] for a justification of this version of the equations, which corrects a small error in [38]). The variable w is the synaptic weight which scales the amplitudes of postsynaptic responses. If long-term plasticity is introduced, this variable is a function of time. In the simulations, for the neurons in the circuits the values for the U, D and F parameters were drawn from Gaussian distributions with mean values which depended on whether the type of presynaptic and postsynaptic neuron of the synapse is excitatory or inhibitory, and were chosen according to the data reported in [38] and [40]. The mean values of the Gaussian distributions are given in Table 2, and the standard deviation was chosen to be 50% of its mean. Negative values were replaced with values drawn from uniform distribution with a range between 0 and twice the mean value. For the simulations involving individual trained neurons, the U, D and F parameters of these neurons were set to the values from Table 2.

We have carried out control experiments with current-based synapses that were not subject to short-term plasticity (see Suppl. Fig. 5,8,9; successful control experiments with static current-based synapses were also carried out for computer simulation 1, results not shown). We found that the results of all our computer simulations also hold for static current-based synapses.

Model of background synaptic activity

To reproduce the background synaptic input cortical neurons receive in vivo, the neurons in our models received an additional noise process as conductance input. The noise process we used is a point-conductance approximation model, described in [27]. According to [27], this noise process models the effect of a bombardment by a large number of synaptic inputs in vivo, which causes membrane potential depolarization, referred to as "high conductance" state. Furthermore, it was shown that it captures the spectral and amplitude characteristics of the input conductances of a detailed biophysical model of a neocortical pyramidal cell that was matched to intracellular recordings in cat parietal cortex in vivo. The ratio of average contributions of excitatory and inhibitory background conductances was chosen to be 5 in accordance to experimental studies during sensory responses (see [41],[42], and [43]). In this model, the noisy synaptic current I_{noise} in equation (24) is a sum of two currents:

$$I_{noise}(t) = g_e(t)(V_m(t) - E_e) + g_i(t)(V_m(t) - E_i), \tag{27}$$

where $g_e(t)$ and $g_i(t)$ are time-dependent excitatory and inhibitory conductances. The values of the respective reversal potentials were $E_e = 0$ mV and $E_i = -75$ mV. The conductances $g_e(t)$ and $g_i(t)$ were modeled according to [27] as a one-variable stochastic process similar to an Ornstein-Uhlenbeck process:

$$\frac{dg_e(t)}{dt} = -\frac{1}{\tau_e}[g_e(t) - g_{e0}] + \sqrt{D_e}\chi_1(t)$$
$$\frac{dg_i(t)}{dt} = -\frac{1}{\tau_i}[g_i(t) - g_{i0}] + \sqrt{D_i}\chi_2(t),$$

with mean $g_{e0} = 0.012\mu$S, noise-diffusion constant $D_e = 0.003\mu$S and time constant $\tau_e = 2.7$ms for the excitatory conductance, and mean $g_{i0} = 0.057\mu$S, noise-diffusion constant $D_i = 0.0066\mu$S, and time constant $\tau_i = 10.5$ms for the inhibitory conductance. $\chi_1(t)$ and $\chi_2(t)$ are Gaussian white noise of zero mean and unit standard deviation.

Since these processes are Gaussian stochastic processes, they can be numerically integrated by an exact update rule:

$$g_e(t + \Delta) = g_{e0} + [g_e(t) - g_{e0}]e^{-\frac{\Delta}{\tau_e}} + A_e N_1(0, 1)$$
$$g_i(t + \Delta) = g_{i0} + [g_i(t) - g_{i0}]e^{-\frac{\Delta}{\tau_i}} + A_i N_2(0, 1),$$

where $N_1(0, 1)$ and $N_2(0, 1)$ are normal random numbers (zero mean, unit standard deviation) and A_e, A_i are amplitude coefficients given by:

$$A_e = \sqrt{\frac{D_e\tau_e}{2}[1 - e^{\frac{-2\Delta}{\tau_e}}]}$$
$$A_i = \sqrt{\frac{D_i\tau_i}{2}[1 - e^{\frac{-2\Delta}{\tau_i}}]}.$$

Reward-modulated STDP

For the computer simulations we used the following parameters for the STDP window function $W(r)$: $A_+ = 0.01w_{max}$, $A_-/A_+ = 1.05$, $\tau_+ = \tau_- = 30$ms. w_{max} denotes

the hard bound of the synaptic weight of the particular plastic synapse. Note that the parameter A_+ can be given arbitrary value in this plasticity rule, since it can be scaled together with the reward signal, i.e. multiplying the reward signal by some constant and dividing A_+ by the same constant results in identical time evolution of the weight changes. We have set A_+ to be 1% of the maximum synaptic weight.

We used the α-function to model the eligibility trace kernel $f_c(t)$

$$f_c(t) = \begin{cases} \frac{t}{\tau_e} e^{-\frac{t}{\tau_e}} & , \quad \text{if } t > 0 \\ 0 & , \quad \text{otherwise} \end{cases} , \tag{28}$$

where the time constant τ_e was set to $\tau_e = 0.4$s in all computer simulations.

For computer simulations 1 and 4 we performed control experiments (see Suppl. Fig. 3,4 and 7) with the weight-dependent synaptic update rule proposed in [23], instead of the purely additive rule (3). We used the parameters proposed in [23], i.e. $\mu = 0.4$, $\alpha = 0.11$, $\tau_+ = \tau_- = 20$ms. The w_0 parameter was calculated according to the formula: $w_0 = \frac{1}{2} w_{max} \alpha^{1/1-\mu}$ where w_{max} is the maximum synaptic weight of the synapse. $\frac{1}{2} w_{max}$ is equal to the initial synaptic weight for the circuit neurons, or to the mean of the distribution of the initial weights for the trained neurons.

Initial weights of trained neurons

The synaptic weights of excitatory synapses to the trained neurons in experiments 2-5 were initialized from a Gaussian distribution with mean $w_{max}/2$. The standard deviation was set to $w_{max}/10$ bounded within the range $[3w_{max}/10,\ 7w_{max}/10]$.

Software

All computer simulations were carried out with the PCSIM software package (http://www.lsm.tugraz.at/pcsim). PCSIM is a parallel simulator for biologically realistic neural networks with a fast c++ simulation core and a Python interface. It has been developed by Thomas Natschläger and Dejan Pecevski. The time step of simulation was set to 0.1ms.

Details to individual computer simulations

For all computer simulations, both for the cortical microcircuits and readout neurons, the same parameters values for the neuron and synapse models and the reward-modulated STDP rule were used, as specified in the previous section (except in computer simulation 3, where the goal was to test the theoretical predictions for different values of the parameters). Each of the computer simulations in this article modeled a specific task or experimental finding. Consequently, the dependence of the reward signal on the behavior of the system had to be modeled in a specific way for each simulation (a more detailed discussion of the reward signal can be found in the Discussion section). The parameters for that are given below in separate subsections which address the individual simulations. Furthermore, some of the remaining parameters in the experiments, i.e. the values of the synaptic weights, the number of synapses of a neuron, number of neurons in the circuit

and the Ornstein-Uhlenbeck (OU) noise levels were chosen to achieve different goals depending on the particular experiment. Briefly stated, these values were tuned to achieve a certain level of firing activity in the neurons, a suitable dynamical regime of the activity in the circuits, and a specific ratio between amount of input the neurons receive from the input synapses and the input generated by the noise process.

We carried out two types of simulations: simulations of cortical microcircuits in computer simulations 1 and 5, and training of readout neurons in computer simulations 2, 3, 4 and 5. In the following we discuss these two types of simulations in more detail.

Cortical Microcircuits

The values of the initial weights of the excitatory and inhibitory synapses for the cortical microcircuits are given in Table 3. All synaptic weights were bounded in the range between 0 and twice the initial synaptic weight of the synapse.

The cortical microcircuit was composed of 4000 neurons connected randomly with connection probabilities described in Details to computer simulation 1. The initial synaptic weights of the synapses and the levels of OU noise were tuned to achieve a spontaneous firing rate of about 4.6 Hz, while maintaining an asynchronous irregular firing activity in the circuit.

50% of all neurons (randomly chosen, 50% excitatory and 50% inhibitory) received downscaled OU noise (by a factor 0.2 from the model reported in [27]), with the subtracted part substituted by additional synaptic input from the circuit. The input connection probabilities of these neurons were scaled up, so that the firing rates remain in the same range as for the other neurons. The reinforced neurons were randomly chosen from this group of neurons. This was done in order to observe how the learning mechanisms work when most of the input conductance in the neuron comes from a larger number of input synapses which are plastic, rather than from a static noise process.

We chose a smaller microcircuit, composed of 540 neurons, for the computer simulation 5 in order to be able to perform a large number of training trials. The synaptic weights in this smaller circuit were chosen (see Table 3) to achieve an appropriate level of firing activity in the circuit that is modulated by the external input. The circuit neurons had injected an Ornstein-Uhlenbeck (OU) noise multiplied by 0.4 in order to emulate the background synaptic activity in neocortical neurons in vivo, and test the learning in a more biologically realistic settings. This produced significant trial-to-trial variability in the circuit response (see Fig. 10D). A lower value of the noise level could also be used without affecting the learning, whereas increasing the amount of injected noise would slowly deteriorate the information that the circuit activity maintains about the injected inputs, resulting in a decline of the learning performance.

Readout neurons

The maximum values of the synaptic weights of readout neurons for computer simulations 2, 4 and 5, together with the number of synapses of the neurons, are given in Table 4.

The neuron in computer simulation 2 had 100 synapses. We chose 200 synapses for the neuron in computer simulation 4, in order to improve the learning performance. Such improvement of the learning performance for larger numbers of synapses is in accordance

with our theoretical analysis (see equation (17), since for learning the classification of temporal patterns the temporal variation of the voltage of the postsynaptic membrane turns out to be of critical importance (see the discussion after equation (17)). This temporal variation depends less on the shape of a single EPSP and more on the temporal pattern of presynaptic firing when the number of synapses is increased. In computer simulation 5 the readout neuron received inputs from all 432 excitatory neurons in the circuit. The synaptic weights were chosen in accordance with the number of synapses in order to achieve a firing rate suitable for the particular task, and to balance the synaptic input and the noise injections in the neurons.

For the pattern discrimination task (computer simulation 4) and the speech recognition task (computer simulation 5), the amount of noise had to be chosen to be high enough to achieve sufficient variation of the membrane potential from trial to trial near the firing threshold, and low enough so that it would not dominate the fluctuations of the membrane potential. In the experiment where the exact spike times were rewarded (computer simulation 2), the noise had a different role. As described in the Results section, there the noise effectively controls the amount of depression. If the noise (and therefore the depression) is too weak, $w^* = 0$ synapses do not converge to 0. If the noise is too strong, $w^* = w_{max}$ synapses do not converge to w_{max}. To achieve the desired learning result, the noise level should be in a range where it reduces the correlations of the synapses with $w^* = 0$ so that the depression of STDP will prevail, but at the same time is not strong enough to do the same for the other group of synapses with $w^* = w_{max}$, since they have stronger pre-before-post correlations. For our simulations, we have set the noise level to the full amount of OU noise.

Details to computer simulation 1: Model for biofeedback experiment

The cortical microcircuit model consisted of 4000 neurons with twenty percent of the neurons randomly chosen to be inhibitory, and the others excitatory. The connections between the neurons were created randomly, with different connectivity probabilities depending on whether the postsynaptic neuron received the full amount of OU noise, or downscaled OU noise with an additional compensatory synaptic input from the circuit. For neurons in the latter sub-population, the connection probabilities were $p_{ee} = 0.02$, $p_{ei} = 0.02$, $p_{ie} = 0.024$ and $p_{ii} = 0.016$ where the ee, ei, ie, ii indices designate the type of the presynaptic and postsynaptic neurons (e=excitatory or i=inhibitory). For the other neurons the corresponding connection probabilities were downscaled by 0.4. The resulting firing rates and correlations for both types of excitatory neurons are plotted in Suppl. Fig. 1 and 2.

The shape of the reward kernel $\epsilon_r(t)$ was chosen as a difference of two α-functions

$$\epsilon_r(t) = A_r^+ \frac{t}{\tau_r^+} e^{1-\frac{t}{\tau_r^+}} - A_r^- \frac{t}{\tau_r^-} e^{1-\frac{t}{\tau_r^-}}, \tag{29}$$

one positive α-pulse with a peak at 0.4 sec after the corresponding spike, and one long-tailed negative α-pulse which makes sure that the integral over the reward kernel is zero. The parameters for the reward kernel were $A_r^+ = 1.379$, $A_r^- = 0.27$, $\tau_r^+ = 0.2$s, $\tau_r^- = 1$s, and $d_r = 0.2$s, which produced a peak value of the reward pulse 0.4s after the spike that caused it.

Details to computer simulation 2: Learning spike times

We used the following function for the reward kernel $\kappa(r)$

$$\kappa(r) = \begin{cases} A_+^\kappa (e^{-\frac{t-t_\kappa}{\tau_1^\kappa}} - e^{-\frac{t-t_\kappa}{\tau_2^\kappa}}) & , \quad \text{if } t - t_\kappa \geq 0 \\ -A_-^\kappa (e^{\frac{t-t_\kappa}{\tau_1^\kappa}} - e^{\frac{t-t_\kappa}{\tau_2^\kappa}}) & , \quad \text{otherwise} \end{cases} \tag{30}$$

where A_+^κ and A_-^κ are positive scaling constants, τ_1^κ and τ_2^κ define the shape of the two double-exponential functions the kernel is composed of, and t_κ defines the offset of the zero-crossing from the origin. The parameter values used in our simulations were $A_+^\kappa = 0.1457$, $A_-^\kappa = -0.1442$, $\tau_1^\kappa = 30$ms, $\tau_2^\kappa = 4$ms and $t_\kappa = -1$ms. The reward delay was equal to $d_r = 0.4$s.

Details to computer simulation 3: Testing the analytically derived conditions

We used a linear Poisson neuron model as in the theoretical analysis with static synapses and exponentially decaying postsynaptic responses $\epsilon(s) = e^{(-s/\tau_\epsilon)}/\tau_\epsilon$. The neuron had 100 excitatory synapses, except in experiment #6, where we used 200 synapses. In all experiments the target neuron received additional 10 excitatory synapses with weights set to w_{max}. The input spike trains were Poisson processes with a constant rate of $r_{pre} = 6$Hz, except in experiment # 6 where the rate was $r_{pre} = 3$Hz. The weights of the target neuron were set to $w_i^* = w_{max}$ for $0 \leq i < 50$ and $w_i^* = 0$ for $50 \leq i < 100$.

The time constants of the reward kernel were $\tau_2^\kappa = 4$ms, whereas τ_1^κ had different values in different experiments (reported in table 1). The value of t_κ was always set to an optimal value such that the $\epsilon_\kappa(0) = \int_0^\infty \kappa(-s)\epsilon(s) = 0$. The time constant τ_- of the negative part of the STDP window function $W(r)$ was set to τ_+. The reward signal was delayed by $\tau_d = 0.4$s. The simulations were performed for varying durations of simulated biological time (see the t_{sim}-column in Table 1).

Details to computer simulation 4: Learning pattern classification

We used the reward signal from equation (16), with an α-function for the reward kernel $\epsilon_r(r) = \frac{e}{\tau} t e^{-t/\tau}$, and the reward delay d_r set to 300 ms. The amplitudes of the positive and negative pulses were $\alpha_P = -\alpha_N = 1.435$. and the time constant of the reward kernel was $\tau = 100$ms.

Details to computer simulation 5: Training a readout neuron with reward-modulated STDP to recognize isolated spoken digits

Spike representations of speech utterances. The speech utterances were preprocessed by the cochlea model described in [44], which captures the filtering properties of the cochlea and hair cells in the human inner ear. The resulting analog signals were encoded by spikes with the BSA spike encoding algorithm described in [45]. We used the same preprocessing to generate the spikes as in [46]. The spike representations had a duration of about 400 ms and 20 input channels. The input channels were connected topographically to the cortical microcircuit model. The neurons in the circuit were split into 20 disjunct subsets of 27 neurons, and each input channel was connected to the 27

neurons in its corresponding subsets. The readout neuron was trained with 20 different spike inputs to the circuit, where 10 of them resulted from utterances of digit "one", and the other 10 resulted from utterances of digit "two" by the same speaker.

Training procedure. We performed 2000 training trials, where for each trial a spike representation of a randomly chosen utterance out of 10 utterances for one digit was injected into the circuit. The digit changed from trial to trial. Whenever the readout neuron spiked during the presentation of an utterance of digit "two", a positive pulse was generated in the reward signal, and accordingly, for utterances of digit "one", a negative pulse in the reward was generated. We used the reward signal from equation (16). The amplitudes of the positive and negative pulses were $\alpha_P = -\alpha_N = 0.883$. The time constant of the reward kernel $\epsilon_r(r)$ was $\tau = 100$ms. The pulses in the reward were delayed $d_r = 300$ ms from the spikes that caused them.

Cortical microcircuit details. The cortical microcircuit model consisted of 540 neurons with twenty percent of the neurons randomly chosen to be inhibitory, and the others excitatory. The recurrent connections in the circuit were created randomly with a connection probability of 0.1. Long-term plasticity was not modeled in the circuit synapses.

The synapses for the connections from the input neurons to the circuit neurons were static, current based with axon conduction delay of 1ms, and exponentially decaying PSR with time constant $\tau_e = 3$ ms and amplitude $w_{input} = 0.715$ nA.

4 Discussion

We have presented in this article analytical tools which make it possible to predict under which conditions reward-modulated STDP will achieve a given learning goal in a network of neurons. These conditions specify relationships between parameters and auxiliary functions (learning curves for STDP, eligibility traces, reward signals etc.) that are involved in the specification of the reward-modulated STDP learning rule. Although our analytical results are based on some simplifying assumptions, we have shown that they predict quite well the outcomes of computer simulations of quite complex models for cortical networks of neurons.

We have applied this learning theory for reward-modulated STDP to a number of biologically relevant learning tasks. We have shown that the biofeedback result of Fetz and Baker [17] can in principle be explained on the basis of reward-modulated STDP. The underlying credit assignment problem was extremely difficult, since the monkey brain had no direct information about the identity of the neuron whose firing rate was relevant for receiving rewards. This credit assignment problem is even more difficult from the perspective of a single synapse, and hence for the application of a local synaptic plasticity rule such as reward-modulated STDP. However our theoretical analysis (see equation (10), (11)) has shown that the longterm evolution of synaptic weights depended only on the correlation of pairs of pre- and postsynaptic spikes with the reward signal. Therefore the firing rate of the rewarded neuron increased (for a computer simulation of a recurrent network consisting of 4000 conductance based LIF neurons with realistic background noise typical for in-vivo conditions, and 228954 synapses that exhibited data-based short

term synaptic plasticity) within a few minutes of simulated biological time, like in the experimental data of [17], whereas the firing rates of the other neurons remained invariant (see Fig. 3B). We were also able to model differential reinforcement of two neurons in this way (Fig. 4). These computer simulations demonstrated a remarkable stability of the network dynamics (see Fig. 3A, 4A, 5) in spite of the fact that all excitatory synapses were continuously subjected to reward-modulated STDP. In particular, the circuit remained in the asynchronous irregular firing regime, that resembles spontaneous firing activity in the cortex [21]. Other STDP-rules (without reward modulation) that maintain this firing regime have previously been exhibited in [23].

Whereas this learning task focused on firing rates, we have also shown (see Fig. 7) that neurons can learn via reward-modulated STDP to respond to inputs with particular spike trains, i.e., particular temporal output patterns. It has been pointed out in [28] that this is a particularly difficult learning task for reward-modulated STDP, and it was shown there that it can be accomplished with a modified STDP rule and more complex reward prediction signals without delays. We have complemented the results of [28] by deriving specific conditions (equation (13)-(15)) under which this learning task can be solved by the standard version of reward-modulated STDP. Extensive computer simulations have shown that these analytically derived conditions for a simpler neuron model predict also for a LIF neuron with conductance based synapses whether it is able to solve this learning task. Fig. 8 shows that this learning theory for reward-modulated STDP is also able to predict quite well *how fast* a neuron can learn to produce a desired temporal output pattern. An interesting aspect of [28] is that there also the utility of third signals that provide information about changes in the expectation of reward was explored. We have considered in this article only learning scenarios where reward prediction is not possible. A logical next step will be to extend our learning theory for reward-modulated STDP to scenarios from classical reinforcement learning theory that include reward prediction.

We have also addressed the question to what extent neurons can learn via reward-modulated STDP to respond with different firing rates to different spatio-temporal presynaptic firing patterns. It had already been shown in [12] that this learning rule enables neurons to classify spatial firing patterns. We have complemented this work by deriving an analytic expression for the expected weight change in this learning scenario (see equation (17)), which clarifies to what extent a neuron can learn by reward-modulated STDP to distinguish differences in the temporal structure of presynaptic firing patterns. This theoretical analysis showed that in the extreme case, where all incoming information is encoded in the relative timing of presynaptic spikes, reward-modulated STDP is not able to produce a higher average membrane potential for selected presynaptic firing patterns, even if that would be rewarded. But it is able to increase the variance of the membrane potential, and thereby also the number of spikes of any neuron model that has (unlike the simple linear Poisson neuron) a firing threshold. The simulation results in Fig. 9 confirm that in this way a LIF neuron can learn with the standard version of reward-modulated STDP to discriminate even purely temporal presynaptic firing patterns, by producing more spikes in response to one of these patterns.

A surprising feature is, that although the neuron was rewarded here only for responding with a higher firing rate to one presynaptic firing pattern P, it automatically started to respond to this pattern P with a specific temporal spike pattern, that advanced in time during training (see Fig. 9A).

Finally, we have shown that a spiking neuron can be trained by reward-modulated STDP to read out information from a simulated cortical microcircuit (see Fig. 10). This is insofar of interest, as previous work [32, 47, 35] had shown that models of generic cortical microcircuits have inherent capabilities to serve as preprocessors for such readout neurons, by combining in diverse linear and nonlinear ways information that was contained in different time segments of spike inputs to the circuit ("liquid computing model"). The classification of spoken words (that were first transformed into spike trains) had been introduced as a common benchmark task for the evaluation of different approaches towards computing with spiking neurons [48, 32, 33, 34, 46]. But so far all approaches that were based on learning (rather than on clever constructions) had to rely on supervised training of a simple linear readout. This gave rise to the question whether also biologically more realistic models for readout neurons can be trained through a biologically more plausible learning scenario to classify spoken words. The results of Fig. 10 may be interpreted as a tentative positive answer to this question. We have demonstrated that LIF neurons with conductance based synapses (that are subject to biologically realistic short term plasticity) can learn without a supervisor through reward-modulated STDP to classify spoken digits. In contrast to the result of Fig. 9, the output code that emerged here was a rate code. This can be explained through the significant in-class variance of circuit responses to different utterances of the same word (see Fig. 10C, D). Although the LIF neuron learnt here without a supervisor to respond with different firing rates to utterances of different words by the same speaker (whereas the rate output was very similar for both words at the beginning of learning, see Fig. 10E), the classification capability of these neurons has not yet reached the level of linear readouts that are trained by a supervisor (for example, speaker independent word classification could not yet be achieved in this way). Further work is needed to test whether the classification capability of LIF readout neurons can be improved through additional preprocessing in the cortical microcircuit model, through a suitable variation of the reward-modulated STDP rule, or through a different learning scenario (mimicking for example preceding developmental learning that also modifies the presynaptic circuit).

The new learning theory for reward-modulated STDP will also be useful for biological experiments that aim at the clarification of details of the biological implementation of synaptic plasticity in different parts of the brain, since it allows to make predictions which types and time courses of signals would be optimal for a particular range of learning tasks. For each of the previously discussed learning tasks, the theoretical analysis provided conditions on the structure of the reward signal $d(t)$ which guaranteed successful learning. For example, in the biofeedback learning scenario (Fig. 3), every action potential of the reinforced neuron led – after some delay – to a change of the reward signal $d(t)$. The shape of this change was defined by the reward kernel $\epsilon(r)$. Our analysis revealed that this reward kernel can be chosen rather arbitrarily as long as the integral over the kernel is zero, and the integral over the product of the kernel and the eligibility function is positive. For another learning scenario, where the goal was that the output spike train S_j^{post} of some neuron j approximates the spike timings of some target spike train S^* (Fig. 7), the reward signal has to depend on both, S_j^{post} and S^*. The dependence of the reward signal on these spike timings was defined by a reward kernel $\kappa(r)$. Our analysis showed that the reward kernel has to be chosen for this task so that the synapses receive positive rewards if the postsynaptic neuron fires close to the time of a spike in the target spike train S^* or

somewhat later, and negative rewards when an output spike occurs in the order of ten milliseconds too early. In the pattern discrimination task of Fig. 9 each postsynaptic action potential was followed – after some delay – by a change of the reward signal which depended on the pattern presented. Our theoretical analysis predicted that this learning task can be solved if the integrals A_i^P and A_i^N defined by equation (18) are such that $A_i^P > 0$ and $A_i^N \approx -A_i^P$. Again, this constraints are fulfilled for a large class of reward kernels, and a natural choice is to use a non-negative reward kernel ϵ_r. There are currently no data available on the shape of reward kernels in biological neural systems. The previous sketched theoretical analysis makes specific prediction for the shape of reward kernels (depending on the type of learning task in which a biological neural system is involved) which can potentially be tested through biological experiments.

An interesting general aspect of the learning theory that we have presented in this article is that it requires substantial trial-to-trial variability in the neural circuit, which is often viewed as "noise" of imperfect biological implementations of theoretically ideal circuits of neurons. This learning theory for reward-modulated STDP suggests that the main functional role of noise is to maintain a suitable level of spontaneous firing (since if a neuron does not fire, it cannot find out whether this will be rewarded), which should vary from trial to trial in order to explore which firing patterns are rewarded.[5] On the other hand if a neuron fires primarily on the basis of a noise current that is directly injected into that neuron, and not on the basis of presynaptic activity, then STDP does not have the required effect on the synaptic connections to this neuron (see Suppl. Fig 6). This perspective opens the door for subsequent studies that compare for concrete biological learning tasks the theoretically derived optimal amount and distribution of trial-to-trial variability with corresponding experimental data.

Related Work

The theoretical analysis of this model is directly applicable to the learning rule considered in [12]. There, the network behavior of reward-modulated STDP was also studied some situations different from the ones in this article. The computer simulations of [12] operate apparently in a different dynamic regime, where LTD dominates LTP in the STDP-rule, and most weights (except those that are actively increased through reward-modulated STDP) have values close to 0 (see Fig. 1b and d in [12], and compare with Fig. 5 in this article). This setup is likely to require for successful learning a larger dominance of pre-before-post over post-before-pre pairs than the one shown in Fig. 3E. Furthermore, whereas a very low spontaneous firing rate of 1 Hz was required in [12], computer simulation 1 shows that reinforcement learning is also feasible at spontaneous firing rates which correspond to those reported in [17] (the preceding theoretical analysis had already suggested that the success of the model does not depend on particularly low firing rates). The articles [15] and [13] investigate variations of reward-modulated STDP rules that do not employ learning curves for STDP that are based on experimental data, but modified curves that arise in the context of a very interesting top-down theoretical approach (distributed reinforcement learning [14]). The authors of [16] arrive at similar learning rules

[5] It had been shown in [32, 47, 35] that such highly variable circuit activity is compatible with a stable performance of linear readouts.

in a supervised scenario which can be reinterpreted in the context of reinforcement learning. We expect that a similar theory as we have presented in this article for the more commonly discussed version of STDP can also be applied to their modified STDP rules, thereby making it possible to predict under which conditions their learning rules will succeed. Another reward based learning rule for spiking neurons was recently presented in [49]. This rule exploits correlations of a reward signal with noisy perturbations of the neuronal membrane conductance in order to optimize some objective function. One crucial assumption of this approach is that the synaptic plasticity mechanism "knows" which contributions to the membrane potential arise from synaptic inputs, and which contributions are due to internal noise. Such explicit knowledge of the noise signal is not needed in the reward-modulated STDP rule of [12], which we have considered in this article. The price one has to pay for this potential gain in biological realism is a reduced generality of the learning capabilities. While the learning rule in [49] approximates gradient ascent on the objective function, this cannot be stated for reward-modulated STDP at present. Timing-based pattern discrimination with a spiking neuron, as discussed in the section "Pattern discrimination with reward-modulated STDP" of this article, was recently tackled in [50]. The authors proposed the tempotron learning rule, which increases the peak membrane voltage for one class of input patterns (if no spike occurred in response to the input pattern) while decreasing the peak membrane voltage for another class of input patterns (if a spike occurred in response to the pattern). The main difference between this learning rule and reward-modulated STDP is that the tempotron learning rule is sensitive to the peak membrane voltage, whereas reward-modulated STDP is sensitive to local fluctuations of the membrane voltage. Since the time of the maximal membrane voltage has to be determined for each pattern by the synaptic plasticity mechanism, the basic tempotron rule is perhaps not biologically realistic. Therefore, an approximate and potentially biologically more realistic learning rule was proposed in [50], where plasticity following error trials is induced at synapse i only if the voltage within the postsynaptic integration time after their activation exceeds a plasticity threshold κ. One potential problem of this rule is the plasticity threshold κ, since a good choice of this parameter strongly depends on the mean membrane voltage after input spikes. This problem is circumvented by reward-modulated STDP, which considers instead the local change in the membrane voltage. Further work is needed to compare the advantages and disadvantages of these different approaches.

Conclusion

Reward-modulated STDP is a very promising candidate for a synaptic plasticity rule that is able to orchestrate local synaptic modifications in such a way that particular functional properties of larger networks of neurons can be achieved and maintained (we refer to [12] and [28] for discussion of potential biological implementations of this plasticity rule). We have provided in this article analytical tools which make it possible to evaluate this rule and variations of this rule not just through computer simulations, but through theoretical analysis. In particular we have shown that successful learning is only possible if certain relationships hold between the parameters that are involved. Some of these predicted relationships can be tested through biological experiments.

Provided that these relationships are satisfied, reward-modulated STDP turns out to be a powerful rule that can achieve self-organization of synaptic weights in large recurrent

networks of neurons. In particular, it enables us to explain seemingly inexplicable experimental data on biofeedback in monkeys. In addition reward-modulated STDP enables neurons to distinguish complex firing patterns of presynaptic neurons, even for data-based standard forms of STDP, and without the need for a supervisor that tells the neuron when it should spike. Furthermore reward-modulated STDP requires substantial spontaneous activity and trial-to-trial variability in order to support successful learning, thereby providing a functional explanation for these ubiquitous features of cortical networks of neurons. In fact, not only spontaneous activity but also STDP itself may be seen in this context as a mechanism that supports the exploration of different firing chains within a recurrent network, until a solution is found that is rewarded because it supports a successful computational function of the network.

Acknowledgment: We would like to thank Markus Diesmann, Eberhard Fetz, Razvan Florian, Yves Fregnac, Wulfram Gerstner, Nikos Logothetis, Abigail Morrison, Matthias Munk, Gordon Pipa and Dan Shulz for helpful discussions. In addition we would like to thank Malcolm Slaney for providing a MATLAB implementation of the cochlea model from [44], as well as Benjamin Schrauwen, David Verstraeten, Michiel D'Haene and Stefan Klampfl for additional code that we used in our speech classification task (computer simulation 5). Written under partial support by the Austrian Science Fund FWF, project # P17229-N04, project # S9102-N04, as well as project # FP6-015879 (FACETS) and project # FP7-216886 (PASCAL2) of the European Union.

References

[1] L. F. Abbott and S. B. Nelson. Synaptic plasticity: taming the beast. *Nature Neuroscience*, 3:1178–1183, 2000.

[2] V. Jacob, D.J. Brasier, I. Erchova, D. Feldman, and D. E. Shulz. Spike timing-dependent synaptic depression in the in vivo barrel cortex of the rat. *J Neuroscience*, 27(6):1271–84, 2007.

[3] C. H. Bailey, M. Giustetto, Y.-Y. Huang, R. D. Hawkins, and E. R. Kandel. Is heterosynaptic modulation essential for stabilizing Hebbian plasticity and memory? *Nature Reviews Neuroscience*, 1:11–20, 2000.

[4] Q. Gu. Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. *Neuroscience*, 111(4):815–835, 2002.

[5] W. Schultz. Behavioral dopamine signals. *Trends in Neuroscience*, 30:203–210, 2007.

[6] J. N. Reynolds, B. I. Hyland, and J. R. Wickens. A cellular mechanism of reward-related learning. *Nature*, 413:67–70, 2001.

[7] John N. Reynolds and Jeffery R. Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15(4-6):507–521, 2002.

[8] S. Bao, V. T. Chan, and M. M. Merzenich. Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature*, 412(6842):79–83, 2001.

[9] D. E. Shulz, R. Sosnik, V. Ego, S. Haidarliu, and E. Ahissar. A neuronal analogue of state-dependent learning. *Nature*, 403(6769):549–553, 2000.

[10] C. M. Thiel, K. J. Friston, and R. J. Dolan. Cholinergic modulation of experience-dependent plasticity in human auditory cortex. *Neuron*, 35(3):567–574, 2002.

[11] D. E. Shulz, V. Ego-Stengel, and E. Ahissar. Acetylcholine-dependent potentiation of temporal frequency representation in the barrel cortex does not depend on response magnitude during conditioning. *J Physiol Paris*, 97(4–6):431–439, 2003.

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

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

[14] J. Baxter and P. L. Bartlett. Direct gradient-based reinforcement learning: I. gradient estimation algorithms. Technical report, Research School of Information Sciences and Engineering, Australian National University, 1999.

[15] D. Baras and R. Meir. Reinforcement learning, spike-time-dependent plasticity, and the bcm rule. *Neural Computation*, 19(8):2245–2279, 2007.

[16] J.-P. Pfister, T. Toyoizumi, D. Barber, and W. Gerstner. Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. *Neural Computation*, 18(6):1318–1348, 2006.

[17] E. E. Fetz and M. A. Baker. Operantly conditioned patterns of precentral unit activity and correlated responses in adjacent cells and contralateral muscles. *J Neurophysiol*, 36(2):179–204, Mar 1973.

[18] E. E. Fetz. Operant conditioning of cortical unit activity. *Science*, 163(870):955–958, 1969.

[19] E. E. Fetz. Volitional control of neural activity: implications for brain-computer interfaces. *J Physiol*, 579(3):571–579, 2007.

[20] E. E. Fetz and D. V. Finocchio. Correlations between activity of motor cortex cells and arm muscels during operantly conditioned response patterns. *Exp. Brain Research*, 23(3):217–240, 1975.

[21] N. Brunel. Dynamics of networks of randomly connected excitatory and inhibitory spiking neurons. *Journal of Physiology-Paris*, 94:445–463, 2000.

[22] W. Gerstner and W. M. Kistler. *Spiking Neuron Models*. Cambridge University Press, Cambridge, 2002.

[23] A. Morrison, A. Aertsen, and M. Diesmann. Spike-timing-dependent plasticity in balanced random networks. *Neural Computation*, 19:1437–1467, 2007.

[24] G.Q. Bi and M.M. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neuroscience*, 18(24):10464–10472, 1998.

[25] S. Song, K. D. Miller, and L. F. Abbott. Competitive hebbian learning through spike-timing dependent synaptic plasticity. *Nature Neuroscience*, 3:919–926, 2000.

[26] R. Kempter, W. Gerstner, and J. L. van Hemmen. Intrinsic stabilization of output rates by spike-based hebbian learning. *Neural Computation*, 13:2709–2741, 2001.

[27] A. Destexhe, M. Rudolph, J. M. Fellous, and T. J. Sejnowski. Fluctuating synaptic conductances recreate in vivo-like activity in neocortical neurons. *Neuroscience*, 107(1):13–24, 2001.

[28] M. A. Farries and A. L. Fairhall. Reinforcement learning with modulated spike timing-dependent synaptic plasticity. *Journal of Neurophysiology*, 98:3648–3665, 2007.

[29] C. F. Stevens and A. M. Zador. Input synchrony and the irregular firing of cortical neurons. *Nature Neuroscience*, 1:210–217, 1998.

[30] Z.F. Mainen and T.J. Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268:1503–1505, 1995.

[31] G. Silberberg, M. Bethge, H. Markram, K. Pawelzik, and M. Tsodyks. Dynamics of population rate codes in ensembles of neocortical neurons. *J Neurophysiology*, 91(2):704–709, 2004.

[32] W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

[33] W. Maass, T. Natschläger, and H. Markram. Fading memory and kernel properties of generic cortical microcircuit models. *Journal of Physiology – Paris*, 98(4–6):315–330, 2004.

[34] A. Destexhe and E. Marder. Plasticity in single neuron and circuit computations. *Nature*, 431:789–795, 2004.

[35] W. Maass, P. Joshi, and E. D. Sontag. Computational aspects of feedback in neural circuits. *PLOS Computational Biology*, 3(1):e165, 1–20, 2007.

[36] D. Nikolić, S. Haeusler, W. Singer, and W. Maass. Temporal dynamics of information content carried by neurons in the primary visual cortex. In *Proc. of NIPS 2006, Advances in Neural Information Processing Systems*, volume 19, pages 1041–1048. MIT Press, 2007.

[37] R. Kempter, W. Gerstner, and J. L. van Hemmen. Hebbian learning and spiking neurons. *Phys. Rev. E*, 59(4):4498–4514, 1999.

[38] H. Markram, Y. Wang, and M. Tsodyks. Differential signaling via the same axon of neocortical pyramidal neurons. *PNAS*, 95:5323–5328, 1998.

[39] W. Maass and H. Markram. Synapses as dynamic memory buffers. *Neural Networks*, 15:155–161, 2002.

[40] A. Gupta, Y. Wang, and H. Markram. Organizing principles for a diversity of GABAergic interneurons and synapses in the neocortex. *Science*, 287:273–278, 2000.

[41] L. J. Borg-Graham, C. Monier, and Y. Frégnac. Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature*, 393:369–373, 1998.

[42] J. A. Hirsch, J. M. Alonso, R. C. Reid, and L. M. Martinez. Synaptic integration in striate cortical simple cells. *J. Neurosci.*, 18(22):9517–9528, 1998.

[43] J. Anderson, I. Lampl, I. Reichova, M. Carandini, and D. Ferster. Stimulus dependence of two-state fluctuations of membrane potential in cat visual cortex. *Nature Neuroscience*, 3(6):617–621, 2000.

[44] R. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *Proceedings of IEEE International Conference on ICASSP*, pages 1282–1285, 1982.

[45] B. Schrauwen and J. Van Campenhout. BSA, a fast and accurate spike train encoding scheme. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 2825–2830, 2003.

[46] D. Verstraeten, B. Schrauwen, D. Stroobandt, and J. Van Campenhout. Isolated word recognition with the liquid state machine: a case study. *Information Processing Letters*, 95(6):521–528, 2005.

[47] S. Häusler and W. Maass. A statistical analysis of information processing properties of lamina-specific cortical microcircuit models. *Cerebral Cortex*, 17(1):149–162, 2007.

[48] J. J. Hopfield and C. D. Brody. What is a moment? Transient synchrony as a collective mechanism for spatio-temporal integration. *Proc. Nat. Acad. Sci. USA*, 98(3):1282–1287, 2001.

[49] I. R. Fiete and H. S. Seung. Gradient learning in spiking neural networks by dynamic perturbation of conductances. *Physical Review Letters*, 97(4):048104–1 to 048104–4, 2006.

[50] R. Gütig and H. Sompolinsky. The tempotron: a neuron that learns spike timing-based decisions. *Nature Neuroscience*, 9(3):420–428, 2006.

Ex.	τ_ϵ[ms]	w_{max}	ν_{min}^{post} [Hz]	$A_+ 10^6$	$\frac{A_-}{A_+}$	τ_+ [ms]	A_+^κ, A_-^κ	τ_2^κ [ms]	t_{sim} [h]
1	10	0.012	10	16.62	1.05	20	3.34, -3.12	20	5
2	7	0.020	5	11.08	1.02	15	4.58, -4.17	16	10
3	20	0.010	6	5.54	1.10	25	1.50, -1.39	40	19
4	7	0.020	5	11.08	1.07	25	4.67, -4.17	16	13
5	10	0.015	6	20.77	1.10	25	3.75, -3.12	20	2
6	25	0.005	3	13.85	1.01	25	3.34, -3.12	20	18

Table 1: Parameter values used for computer simulation 3 (see Fig. 8).

source/dest.	exc.(U,D,F)	inh. (U,D,F)
exc.	0.5, 1.1, 0.02	0.25, 0.7, 0.02
inh.	0.05, 0.125, 1.2	0.32, 0.144, 0.06

Table 2: Mean values of the U, D and F parameters in the model from [38] for the short-term dynamics of synapses, depending on the type of the presynaptic and postsynaptic neuron (excitatory or inhibitory). These mean values, based on experimental data from [38, 40], were used in all computer simulations.

Cortical microcircuits					
simulation No.	neurons	$p_{ee}, p_{ei}, p_{ei}, p_{ii}$	$w_{exc}(0)$ $[nS]$	$w_{inh}(0)$ $[nS]$	C_{OU}
1	4000	0.02,0.02,0.024,0.016	10.7	211.6	1.0, 0.2
5	540	0.1	0.784	5.1	0.4

Table 3: Specific parameter values for the cortical microcircuits in computer simulation 1 and 5. p_{conn} is the connection probability, $w_{exc}(0)$ and $w_{inh}(0)$ are the initial synaptic weights for the excitatory and inhibitory synapses respectively, and C_{OU} is the scaling factor for the Ornstein-Uhlenbeck noise injected in the neurons.

Trained (readout) neurons			
simulation No.	num. synapses	w_{max} $[nS]$	C_{OU}
2	100	11.9	1.0
4	200	5.73	0.2
5	432	2.02	0.2

Table 4: Specific parameter values for the trained neurons in computer simulation 2, 4 and 5. w_{max} is the upper hard bound of the synaptic weights of the synapses. C_{OU} is the scaling factor for the Ornstein-Uhlenbeck noise injected in the neurons.

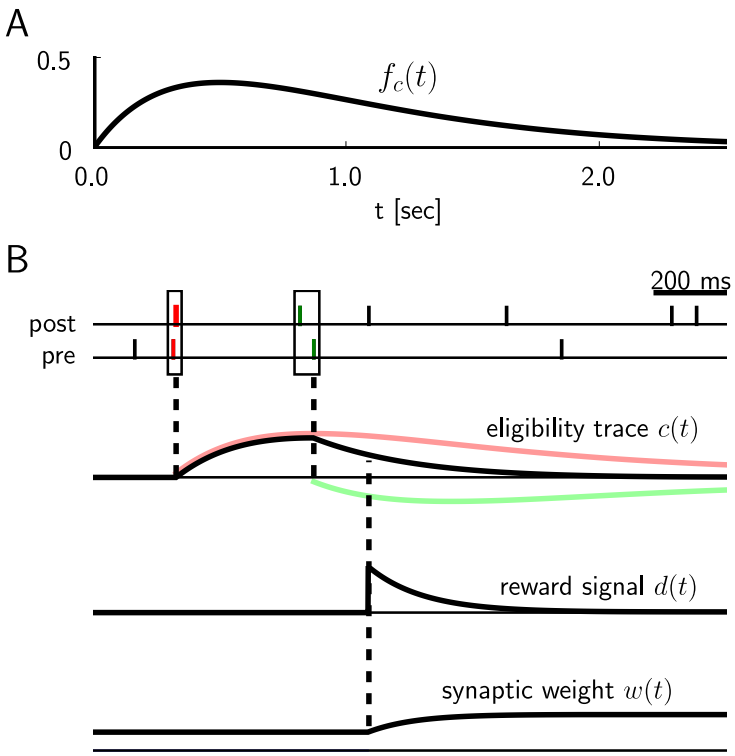


Figure 1: Scheme of reward-modulated STDP according to equations (1) - (4). **A)** Eligibility function $f_c(t)$, which scales the contribution of a pre/post spike pair (with the second spike at time 0) to the eligibility trace $c(t)$ at time t. **B)** Contribution of a pre-before-post spike pair (in red) and a post-before-pre spike pair (in green) to the eligibility trace $c(t)$ (in black), which is the sum of the red and green curves. According to equation (1) the change of the synaptic weight w is proportional to the product of $c(t)$ with a reward signal $d(t)$.

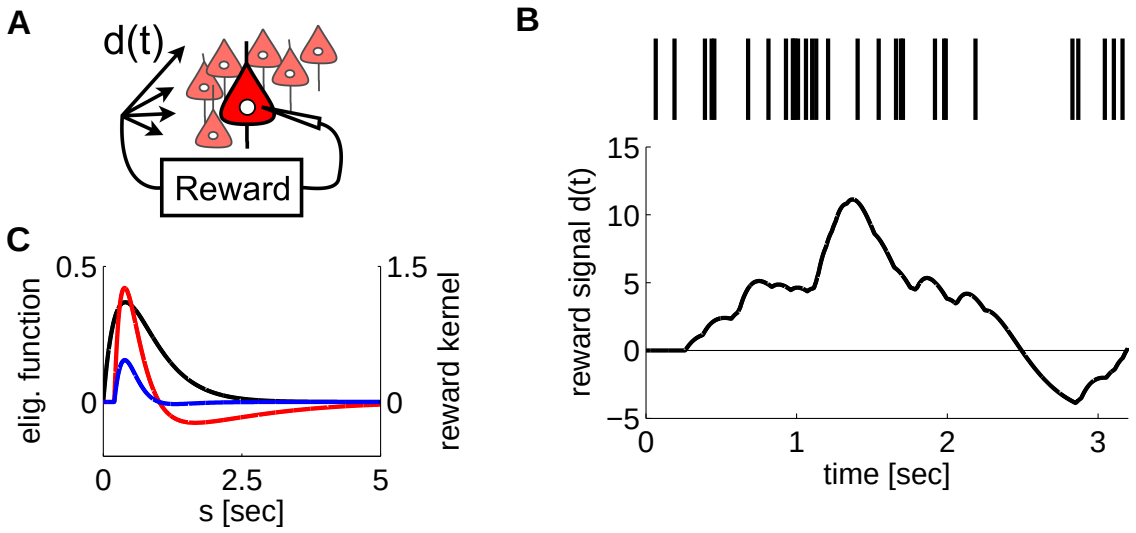


Figure 2: Setup of the model for the experiment by Fetz and Baker [17]. **A)** Schema of the model: The activity of a single neuron in the circuit determines the amount of reward delivered to all synapses between excitatory neurons in the circuit. **B)** The reward signal $d(t)$ in response to a spike train (shown at the top) of the arbitrarily selected neuron (which was selected from a recurrently connected circuit consisting of 4000 neurons). The level of the reward signal $d(t)$ follows the firing rate of the spike train. **C)** The eligibility function $f_c(s)$ (black curve, left axis), the reward kernel $\epsilon_r(s)$ delayed by 200 ms (red curve, right axis), and the product of these two functions (blue curve, right axis) as used in our computer experiment. The integral of $f_c(s+d_r)\epsilon_r(s)$ is positive, as required according to equation (10) in order to achieve a positive learning rate for the synapses to the selected neuron.

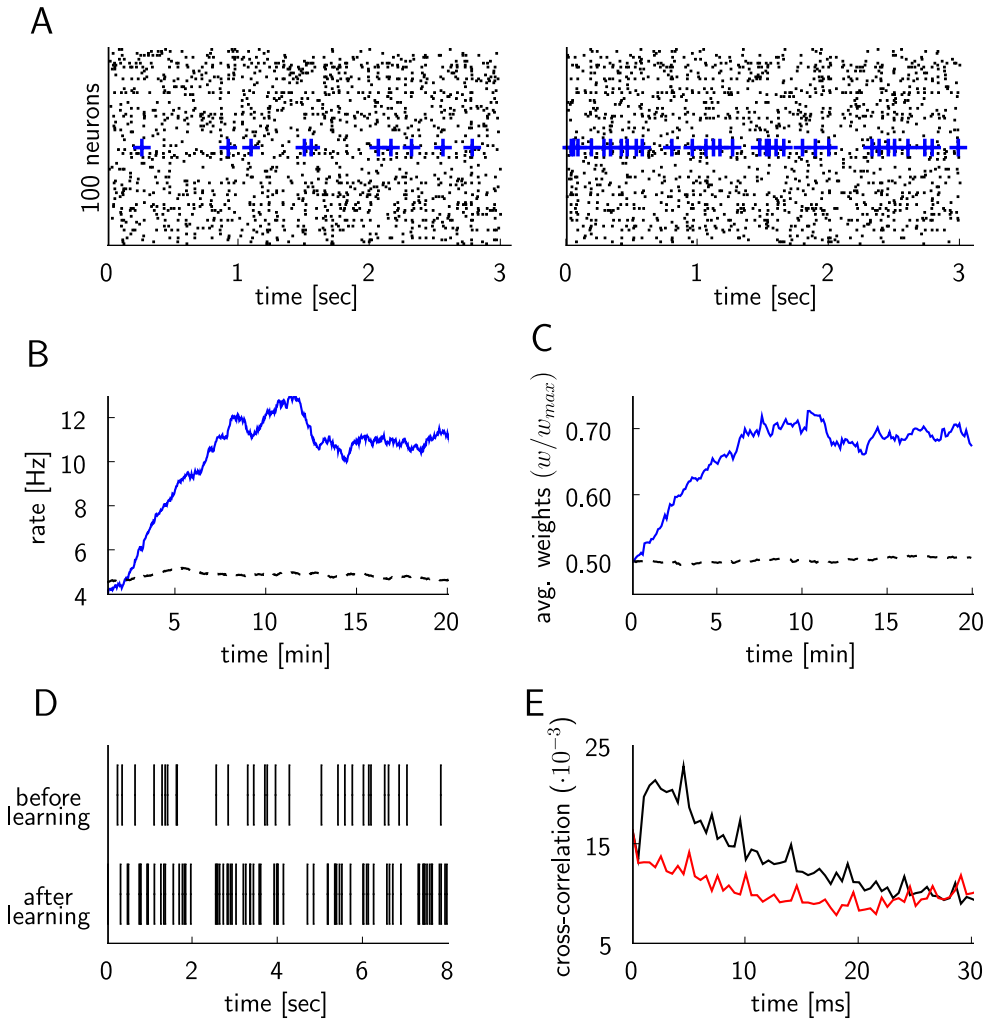


Figure 3: Simulation of the experiment by Fetz and Baker [17] for the case where an arbitrarily selected neuron triggers global rewards when it increases its firing rate. **A)** Spike response of 100 randomly chosen neurons within the recurrent network of 4000 neurons at the beginning of the simulation (20sec - 23sec, left plot), and at the end of the simulation (the last 3 seconds, right plot). The firing times of the reinforced neuron are marked by blue crosses. **B)** The firing rate of the positively rewarded neuron (blue line) increases, while the average firing rate of 20 other randomly chosen neurons (dashed line) remains unchanged. **C)** Evolution of the average weight of excitatory synapses to the reinforced neuron (blue line), and of the average weight of 1663 randomly chosen excitatory synapses to other neurons in the circuit (dashed line). **D)** Spike trains of the reinforced neuron before and after learning. **E)** Histogram of the time-differences between presynaptic and postsynaptic spikes (bin size 0.5ms), averaged over all excitatory synapses to the reinforced neuron. The black curve represents the histogram values for positive time differences (when the presynaptic spike precedes the postsynaptic spike), and the red curve represents the histogram for negative time differences.

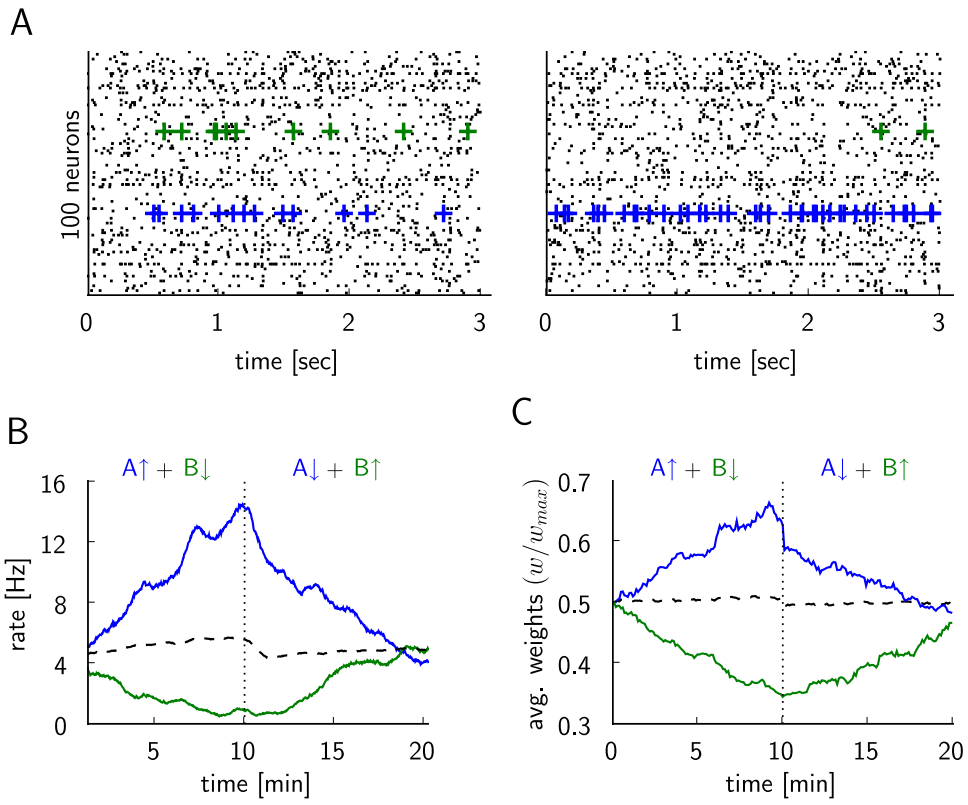


Figure 4: Differential reinforcement of two neurons (within a simulated network of 4000 neurons, the two rewarded neurons are denoted as A and B), corresponding to the experimental results shown in Fig. 9 of [17] and Fig. 1 of [19]. **A)** The spike response of 100 randomly chosen neurons at the beginning of the simulation (20sec - 23sec, left plot), and at the middle of simulation just before the switching of the reward policy (597sec-600sec, right plot). The firing times of the first reinforced neuron A are marked by blue crosses and those of the second reinforced neuron B are marked by green crosses. **B)** The dashed vertical line marks the switch of the reinforcements at $t = 10$min. The firing rate of neuron A (blue line) increases while it is positively reinforced in the first half of the simulation and decreases in the second half when its spiking is negatively reinforced. The firing rate of the neuron B (green line) decreases during the negative reinforcement in the first half and increases during the positive reinforcement in the second half of the simulation. The average firing rate of 20 other randomly chosen neurons (dashed line) remains unchanged. **C)** Evolution of the average weight of excitatory synapses to the rewarded neurons A and B (blue and green lines respectively), and of the average weight of 1744 randomly chosen excitatory synapses to other neurons in the circuit (dashed line).

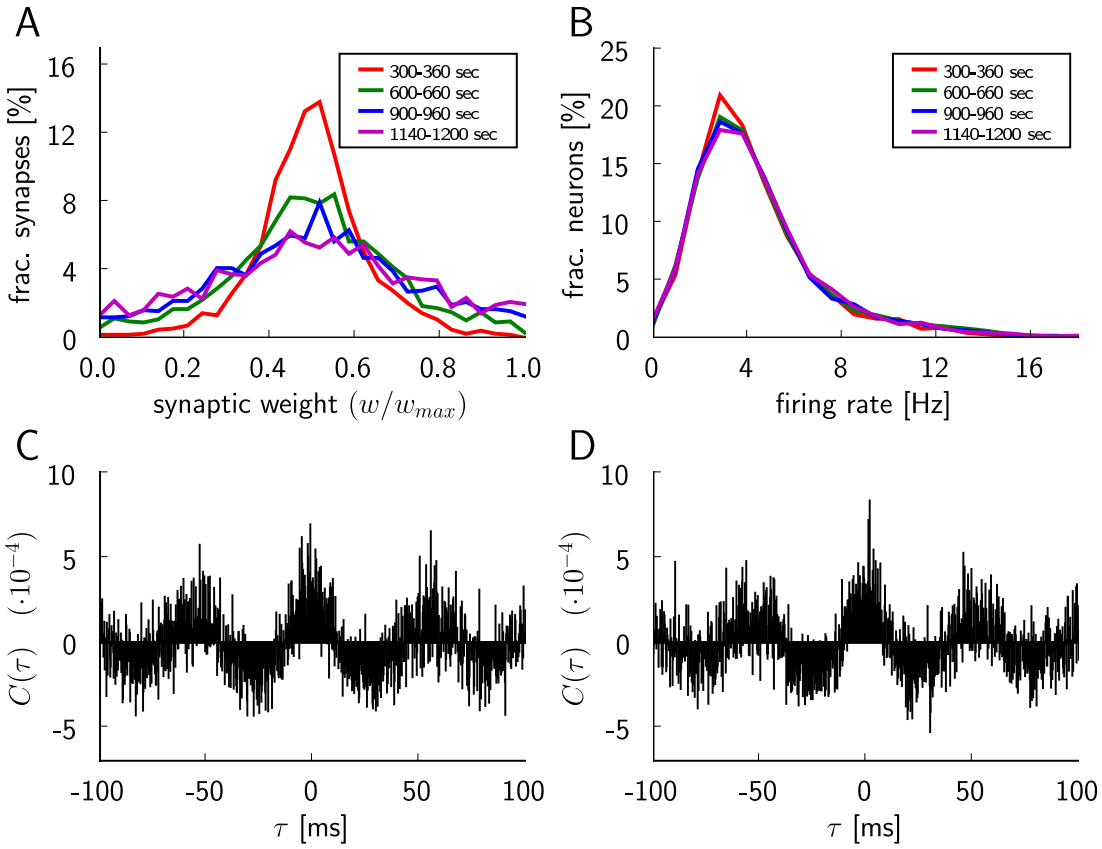


Figure 5: Evolution of the dynamics of a recurrent network of 4000 LIF neurons during application of reward-modulated STDP. **A)** Distribution of the synaptic weights of excitatory synapses to 50 randomly chosen non-reinforced neurons, plotted for 4 different periods of simulated biological time during the simulation. The weights are averaged over 10 samples within these periods. The colors of the curves and the corresponding intervals are as follows: red ($300 - 360$ sec), green ($600 - 660$ sec), blue ($900 - 960$ sec), magenta ($1140 - 1200$ sec). **B)** The distribution of average firing rates of the non-reinforced excitatory neurons in the circuit, plotted for the same time periods as in A). The colors of the curves are the same as in A). The distribution of the firing rates of the neurons in the circuit remains unchanged during the simulation, which covers 20 minutes of biological time. **C)** Cross-correlogram of the spiking activity in the circuit, averaged over 200 pairs of non-reinforced neurons and over 60 s, with a bin size of 0.2 ms, for the period between 300 and 360 seconds of simulated biological time. It is calculated as the cross-covariance divided by the square root of the product of variances. **D)** As in C), but between seconds 1140 and 1200.(Separate plots of panel B, C, D for two types of excitatory neurons that received different amounts of noise currents are given in Suppl. Fig. 1, 2.)

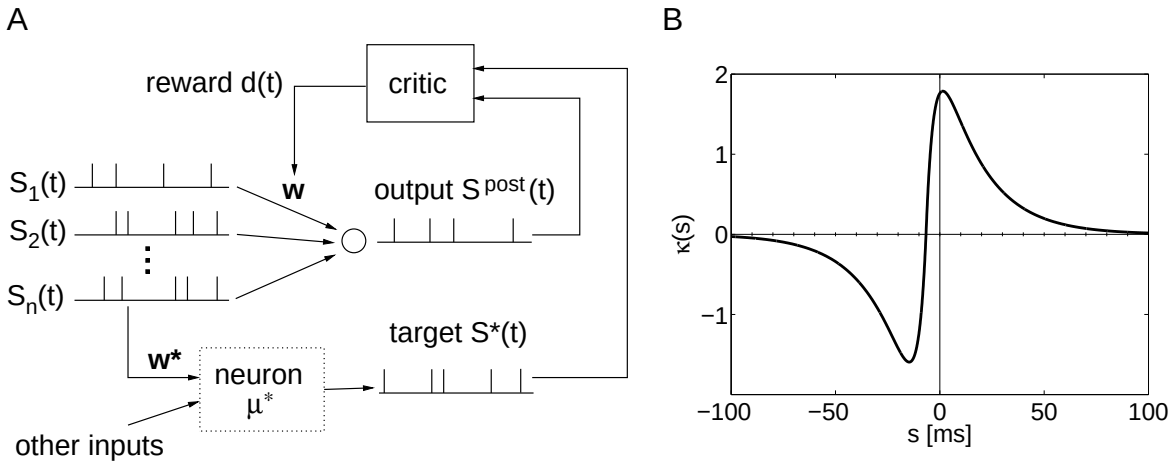


Figure 6: Setup for reinforcement learning of spike times. **A)** Architecture. The trained neuron receives n input spike trains. The neuron μ^* receives the same inputs plus additional inputs not accessible to the trained neuron. The reward is determined by the timing differences between the action potentials of the trained neuron and the neuron μ^*. **B)** A reward kernel with optimal offset from the origin of $t_\kappa = -6.6$ms. The optimal offset for this kernel was calculated with respect to the parameters from computer simulation 1 in Table 1. Reward is positive if the neuron spikes around the target spike or somewhat later, and negative if the neuron spikes much too early.

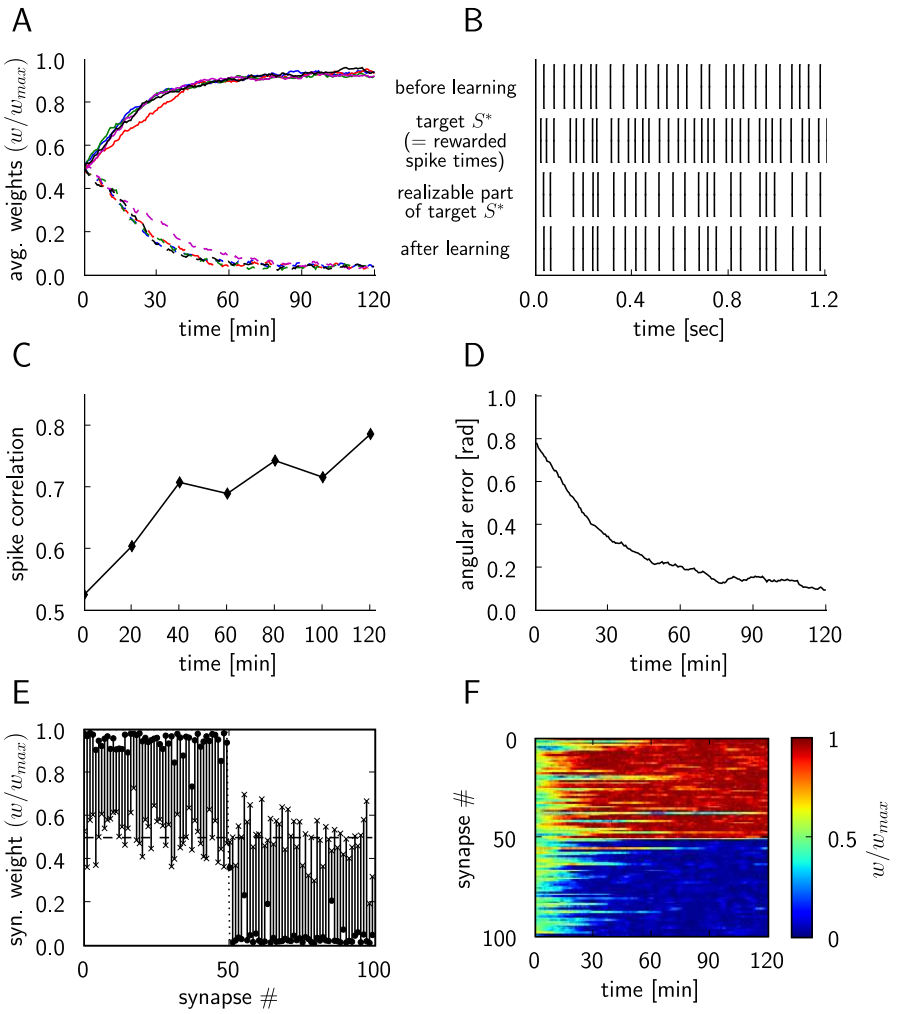


Figure 7: Results for reinforcement learning of exact spike times through reward-modulated STDP. **A)** Synaptic weight changes of the trained LIF neuron, for 5 different runs of the experiment. The curves show the average of the synaptic weights that should converge to $w_i^* = 0$ (dashed lines), and the average of the synaptic weights that should converge to $w_i^* = w_{max}$ (solid lines) with different colors for each simulation run. **B)** Comparison of the output of the trained neuron before (top trace) and after learning (bottom trace). The same input spike trains and the same noise inputs were used before and after training for 2 hours. The second trace from above shows those spike times S^* which are rewarded, the third trace shows the realizable part of S^* (i.e. those spikes which the trained neuron could potentially learn to reproduce, since the neuron μ^* produces them without its 10 extra spike inputs). The close match between the third and fourth trace shows that the trained neuron performs very well. **C)** Evolution of the spike correlation between the spike train of the trained neuron and the realizable part of the target spike train S^*. **D)** The angle between the weight vector $\mathbf{w}$ of the trained neuron and the weight vector $\mathbf{w}^*$ of the neuron μ^* during the simulation, in radians. **E)** Synaptic weights at the beginning of the simulation are marked with $\times$, and at the end of the simulation with $\bullet$, for each plastic synapse of the trained neuron. **F)** Evolution of the synaptic weights w/w_{max} during the simulation (we had chosen $w_i^* = w_{max}$ for $i < 50$, $w_i^* = 0$ for $i \geq 50$).

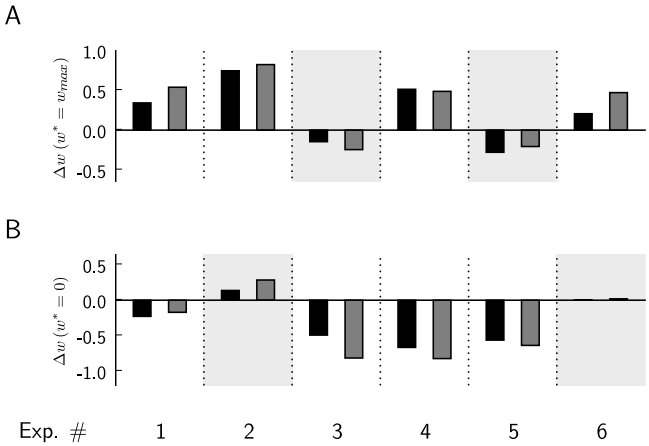


Figure 8: Test of the validity of the analytically derived conditions (13)-(15) on the relationship between parameters for successful learning with reward-modulated STDP. Predicted average weight changes (black bars) calculated from equation (22) match in sign and magnitude the actual average weight changes (gray bars) in computer simulations, for 6 different experiments with different parameter settings (see Table 1) . **A)** Weight changes for synapses with $w_i^* = w_{max}$. **B)** Weight changes for synapses with $w_i^* = 0$. Four cases where the constraints (13) - (15) are not fulfilled are shaded in light gray. In all of these four cases the weights move into the opposite direction, i.e., a direction that decreases rewards.

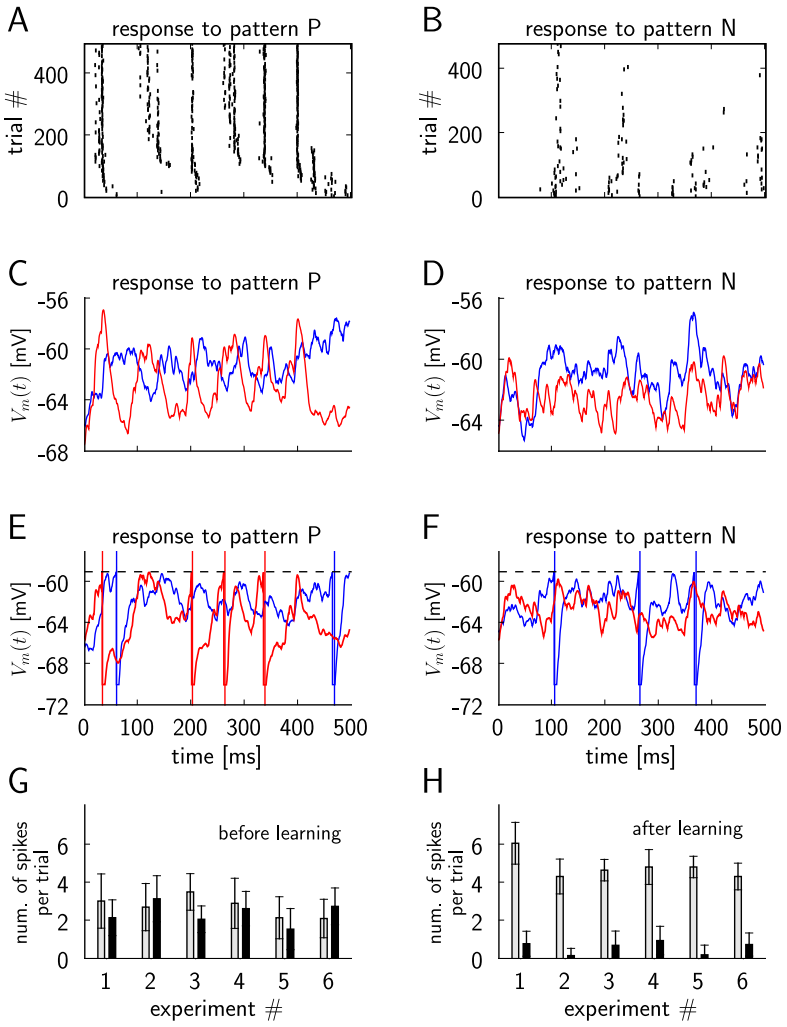


Figure 9: Training a LIF neuron to classify purely temporal presynaptic firing patterns: a positive reward is given for firing of the neuron in response to a temporal presynaptic firing pattern P, and a negative reward for firing in response to another temporal pattern N. **A)** The spike response of the neuron for individual trials, during 500 training trials when pattern P is presented. Only the spikes from every 4-th trial are plotted. **B)** As in A), but in response to pattern N. **C)** The membrane potential $V_m(t)$ of the neuron during a trial where pattern P is presented, before (blue curve) and after training (red curve), with the firing threshold removed. The variance of the membrane potential increases during learning, as predicted by the theory. **D)** As in C), but for pattern N. The variance of the membrane potential for pattern N decreases during learning, as predicted by the theory. **E)** The membrane potential $V_m(t)$ of the neuron (including action potentials) during a trial where pattern P is presented before (blue curve) and after training (red curve). The number of spikes increases. **F)** As in E), but for trials where pattern N is given as input. The number of spikes decreases. **G)** Average number of output spikes per trial before learning, in response to pattern P (gray bars) and pattern N (black bars), for 6 experiments with different randomly generated patterns P and N, and different random initial synaptic weights of the neuron. **H)** As in G), for the same experiments, but after learning. The average number of spikes per trial increases after training for pattern P, and decreases for pattern N.

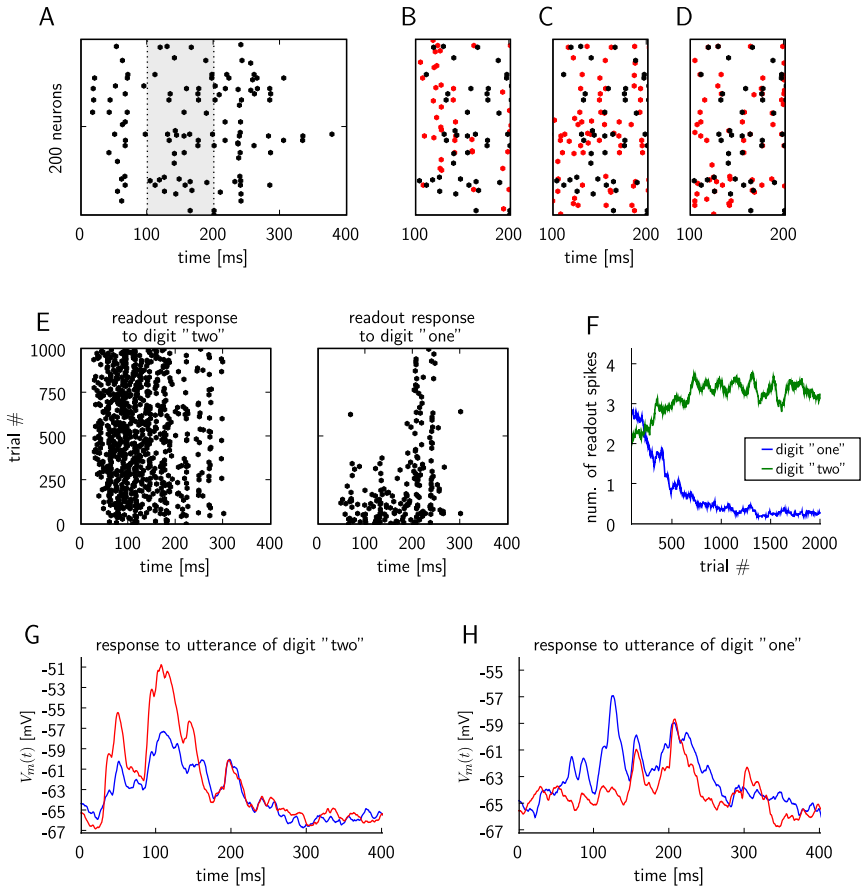


Figure 10: A LIF neuron is trained through reward-modulated STDP to discriminate as a "readout neuron" responses of generic cortical microcircuits to utterances of different spoken digits. **A)** Circuit response to an utterance of digit "one" (spike trains of 200 out of 540 neurons in the circuit are shown). The response within the time period from 100 to 200 ms (marked in gray) is used as a reference in the subsequent 3 panels. **B)** The circuit response from A) (black) for the period between 100 and 200 ms, and the circuit response to an utterance of digit "two" (red). **C)** The circuit spike response from A) (black) and a circuit response for another utterance of digit "two" (red), also shown for the period between 100 and 200 ms. **D)** The circuit spike response from A) (black), and another circuit response to the same utterance in another trial (red). The responses differ due to the presence of noise in the circuit. **E)** Spike response of the LIF readout neuron for different trials during learning, for trials where utterances of digit "two" (left plot) and digit "one" (right plot) are presented as circuit inputs. The spikes from each 4th trial are plotted. **F)** Average number of spikes in the response of the readout during training, in response to digit "one" (blue) and digit "two" (green). The number of spikes were averaged over 40 trials. **G)** The membrane potential $V_m(t)$ of the neuron during a trial where an input pattern corresponding to an utterance of digit "two" is presented, before (blue curve) and after training (red curve), with the firing threshold removed. **H)** As in G), but for an input pattern corresponding to an utterance of digit "one". The variance of the membrane potential increases during learning for utterances of the rewarded digit, and decreases for the non-rewarded digit.

Supplementary Information to "A Learning Theory for Reward-Modulated Spike-timing-dependent Plasticity with an Application to Biofeedback"

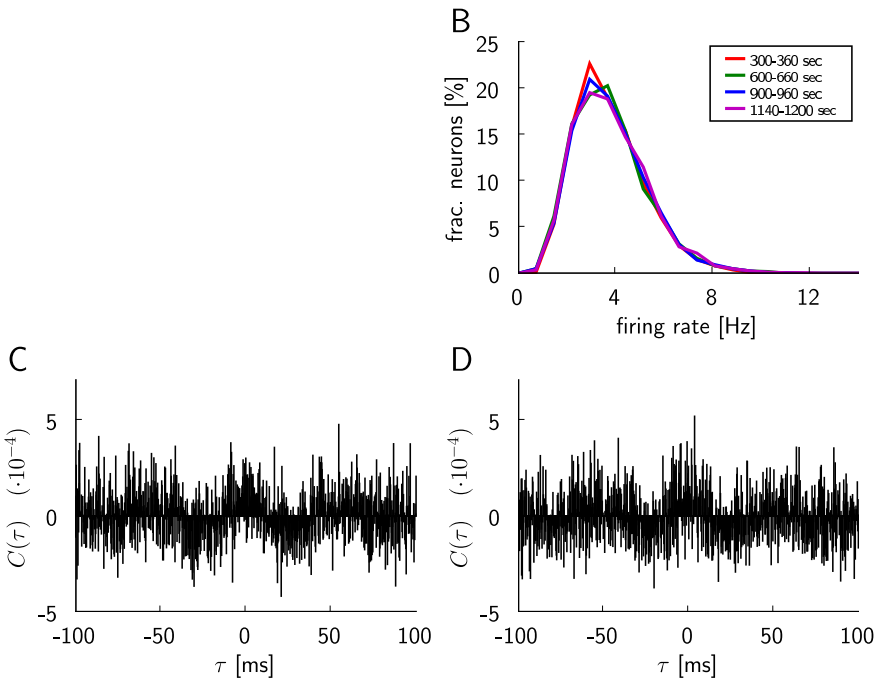
Robert Legenstein, Dejan Pecevski, Wolfgang Maass
Institute for Theoretical Computer Science
Graz University of Technology
A-8010 Graz, Austria
{legi,dejan,maass}@igi.tugraz.at

June 23, 2008

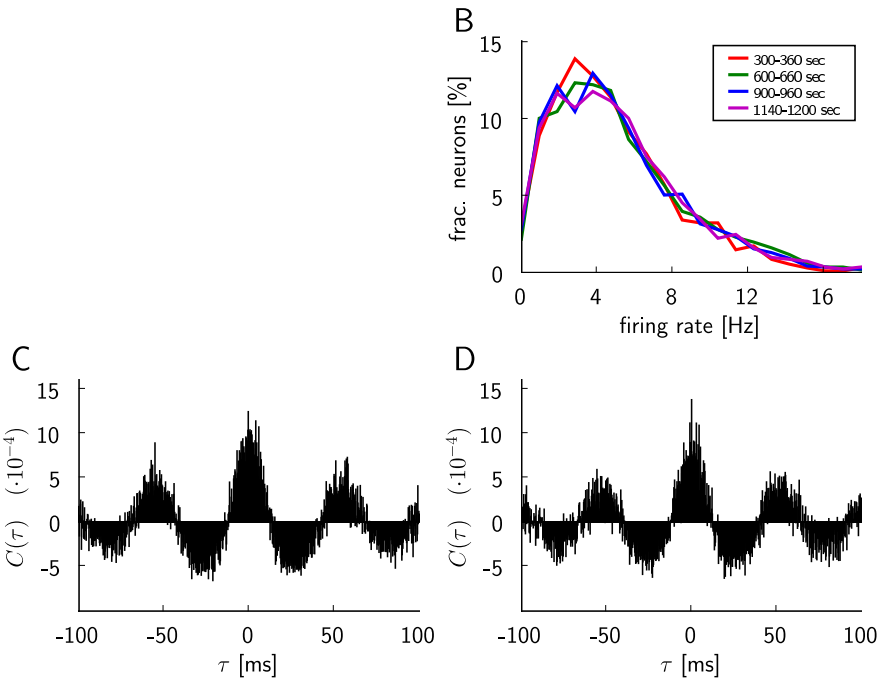
References

[1] A. Morrison, A. Aertsen, and M. Diesmann. Spike-timing-dependent plasticity in balanced random networks. *Neural Computation*, 19:1437–1467, 2007.

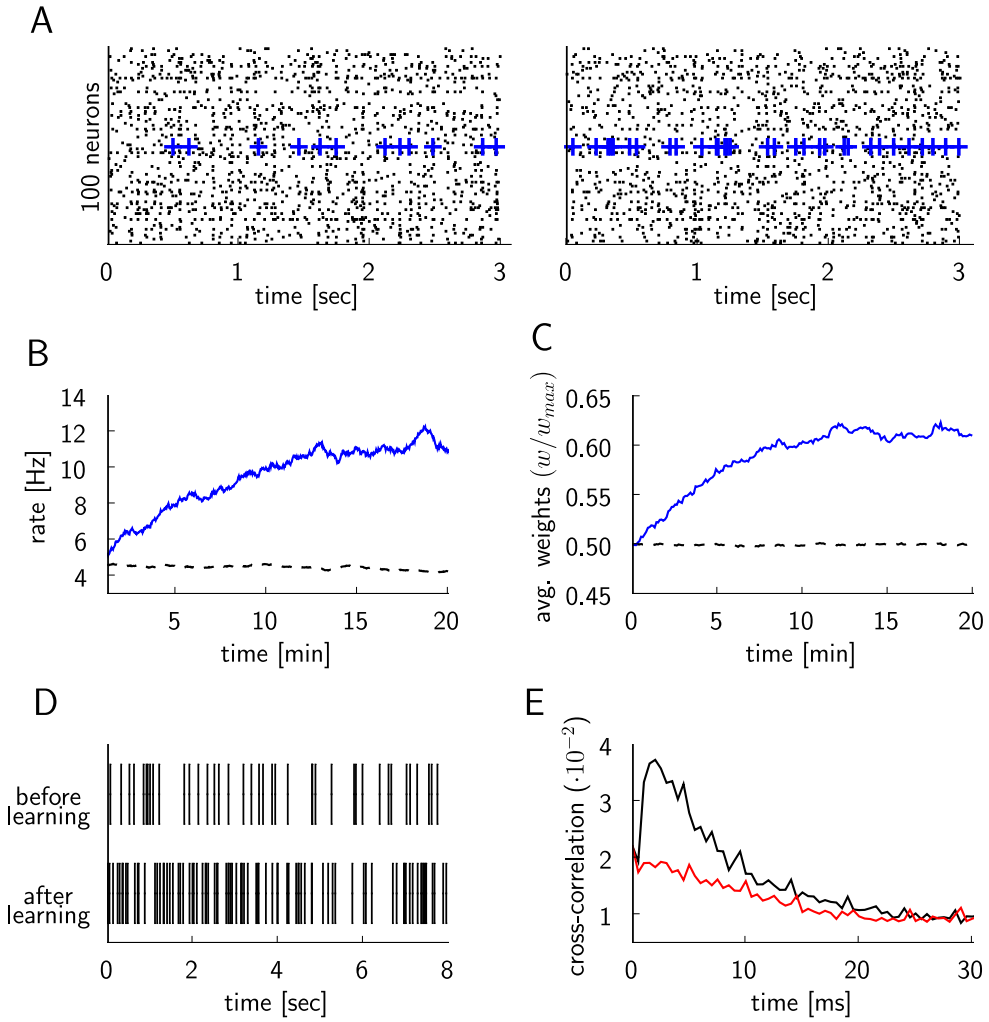
[2] R. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *Proceedings of IEEE International Conference on ICASSP*, pages 1282–1285, 1982.



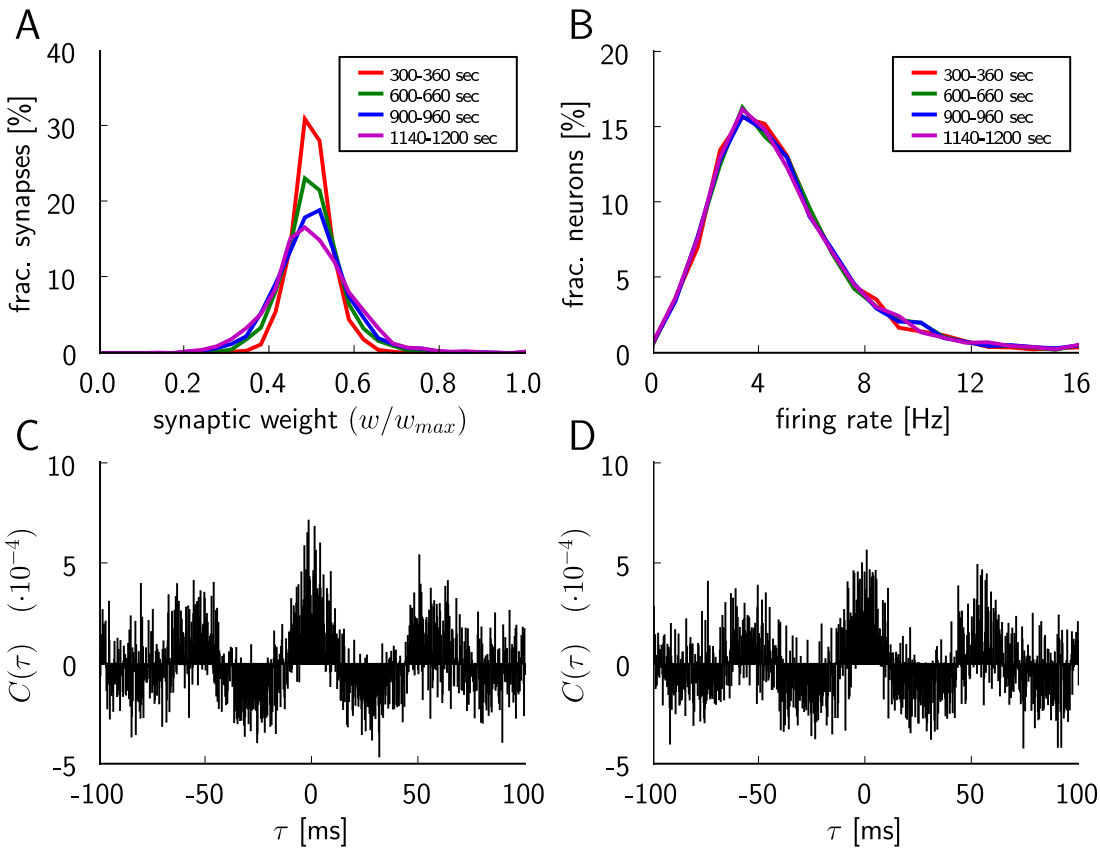
Supplementary Figure 1: Variations of panels **B) - D)** in Fig. 5 for those excitatory neurons which receive the full amount of Ornstein-Uhlenbeck noise. **B)** The distribution of the firing rates of these neurons remains unchanged during the simulation. The colors of the curves and the corresponding intervals are as follows: red ($300 - 360$ sec), green ($600 - 660$ sec), blue ($900 - 960$ sec), magenta ($1140 - 1200$ sec). **C)** Cross-correlogram of the spiking activity of these neurons, averaged over 200 pairs of neurons and over 60 s, with a bin size of 0.2 ms, for the period between 300 and 360 seconds of simulation time. It is calculated as the cross-covariance divided by the square root of the product of variances. **D)** As in **C)**, but for the last 60 seconds of the simulation. The correlation statistics in the circuit is stable during learning.



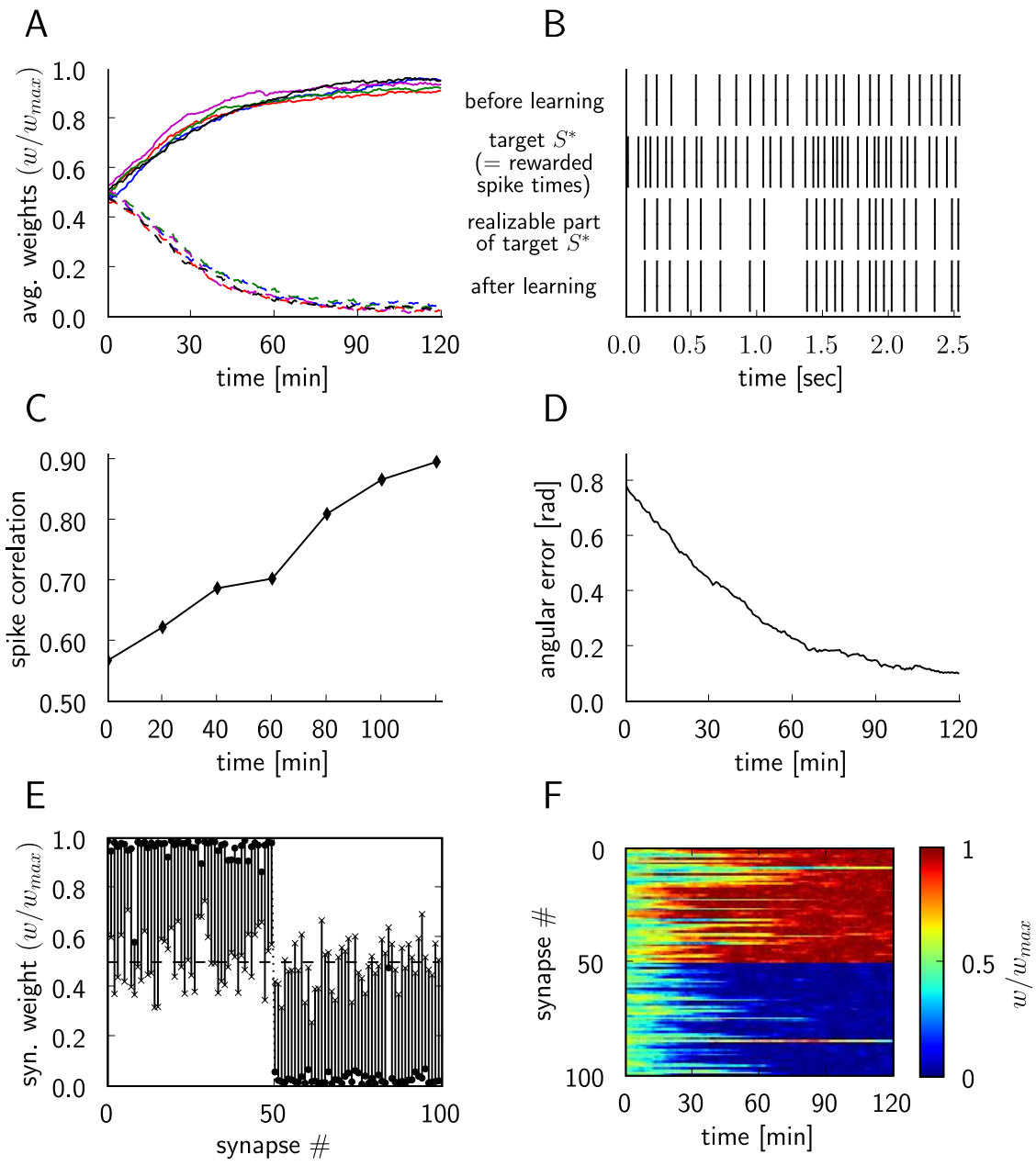
Supplementary Figure 2: Variations of panels **B) - D)** in Fig. 5 for those excitatory neurons which receive a reduced amount of Ornstein-Uhlenbeck noise, but receive more synaptic inputs from other neurons. **B)** The distribution of the firing rates of these neurons remains unchanged during the simulation. The colors of the curves and the corresponding intervals are as follows: red ($300 - 360$ sec), green ($600 - 660$ sec), blue ($900 - 960$ sec), magenta ($1140 - 1200$ sec). **C)** Cross-correlogram of the spiking activity in the circuit, averaged over 200 pairs of these neurons and over 60 s, with a bin size of 0.2 ms, for the period between 300 and 360 seconds of simulation time. It is calculated as the cross-covariance divided by the square root of the product of variances. **D)** As in **C)**, but for the last 60 seconds of the simulation. The correlation statistics in the circuit is stable during learning.



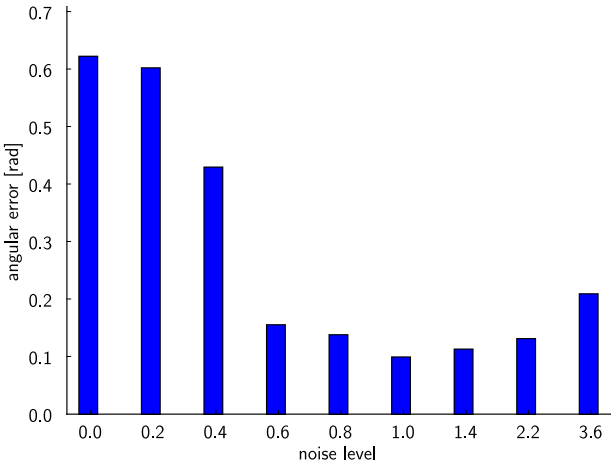
Supplementary Figure 3: Variation of Fig. 3 from computer simulation 1 with results from a simulation where the weight-dependent version of STDP proposed in [1] was used. This STDP rule is defined by the following equations: $\Delta w_+ = \lambda w_0^{1-\mu} w^\mu e^{-|\Delta t|/\tau_+}$ and $\Delta w_- = \lambda \alpha w e^{-|\Delta t|/\tau_-}$. We used the parameters proposed in [1], i.e. $\mu = 0.4$, $\alpha = 0.11$, $\tau_+ = \tau_- = 20$ms, $\lambda = 0.1$ and $w_0 = 272.6$ pS. The w_0 parameter was calculated according to the formula: $w_0 = \frac{1}{2} w_{max} \alpha^{\frac{1}{1-\mu}}$ where w_{max} is the maximum synaptic weight of the synapse. The amplitude parameters A_r^+, A_r^- for the reward kernel were set to $A_r^+ = 1.104$ and $A_r^- = 0.221$. All other parameter values were the same as in computer simulation 1.



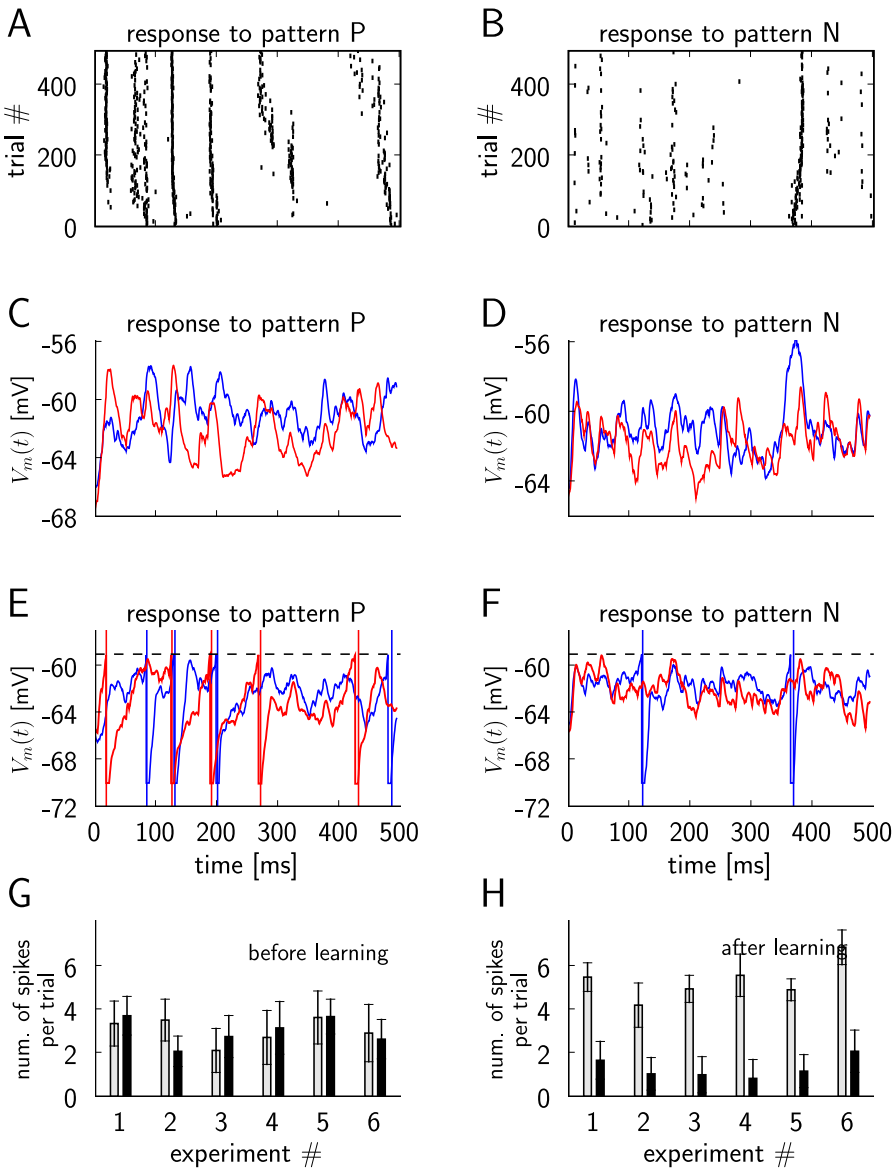
Supplementary Figure 4: Variation of Fig. 5 for the weight-dependent STDP rule from [1] (as in Suppl. Fig. 3).



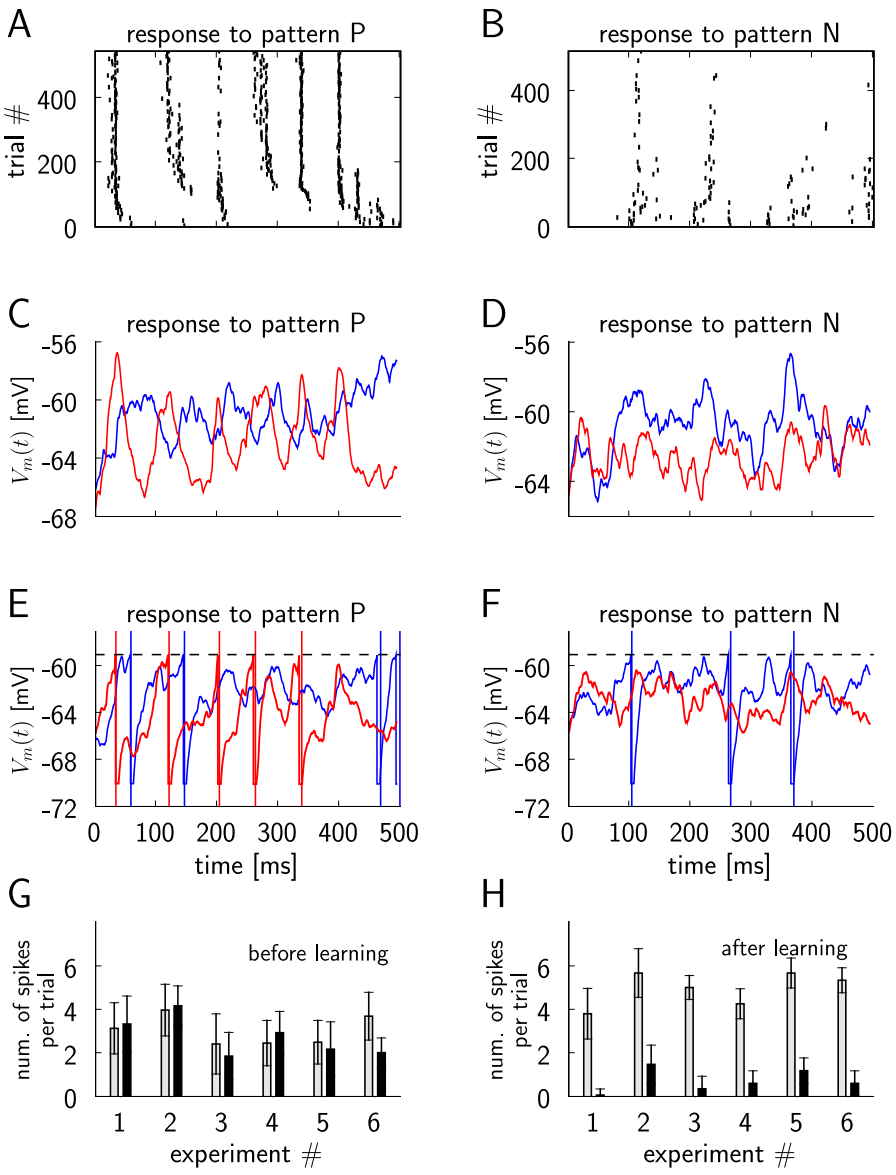
Supplementary Figure 5: Variation of Fig. 7 (i.e., of computer simulation 2) for a simulation where we used current-based synapses without short-term plasticity. The postsynaptic response had an exponentially decaying form $\epsilon(s) = e^{-s/\tau_\epsilon}/\tau_\epsilon$, with $\tau_\epsilon = 5ms$. The value of the maximum synaptic weight was $w_{max} = 32.9$ pA. All other parameter values were the same as in computer simulation 2.



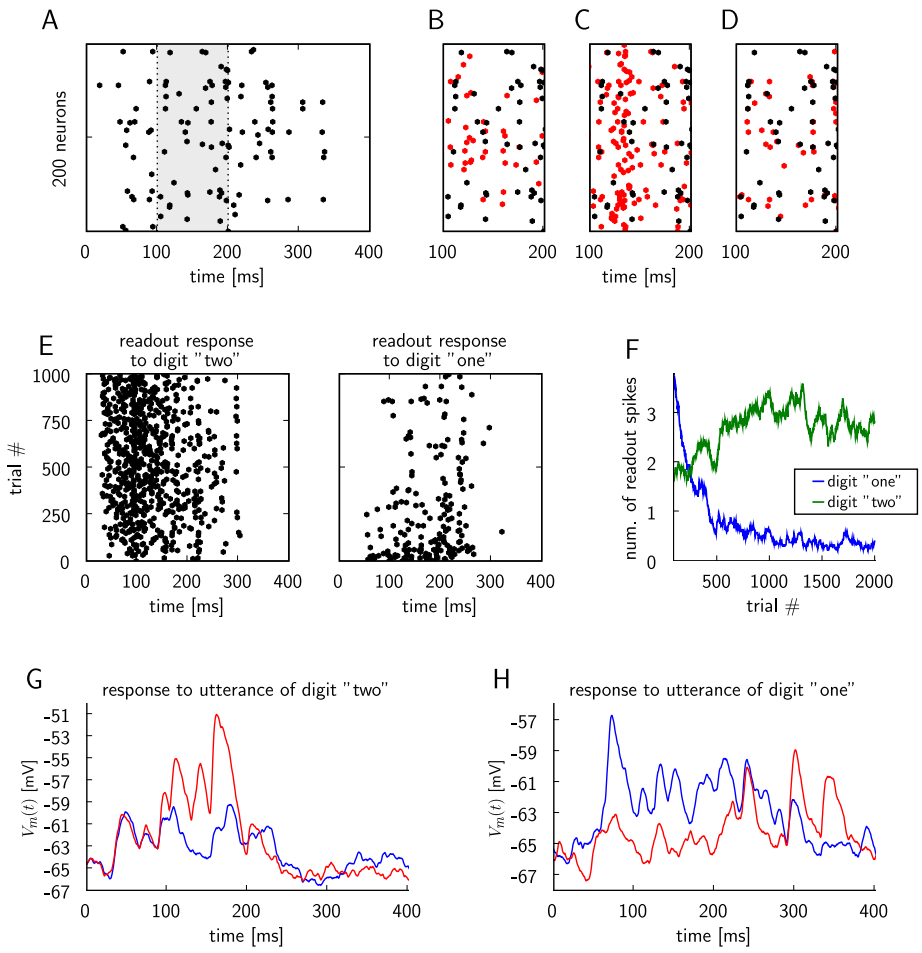
Supplementary Figure 6: Dependence of the learning performance on the noise level in computer simulation 2. The angular error (defined as the angle between the weight vector $\mathbf{w}$ of the trained neuron at the end of the simulation and the weight vector $\mathbf{w}^*$ of the neuron μ^*) is taken as measure for the learning performance, and plotted for 9 simulations with different noise levels that are given on the x-axis (in terms of multiples of the noise level chosen for Fig. 7). All other parameters values were the same as in computer simulation 2. The figure shows that the learning performance declines both for too little and for too much noise.



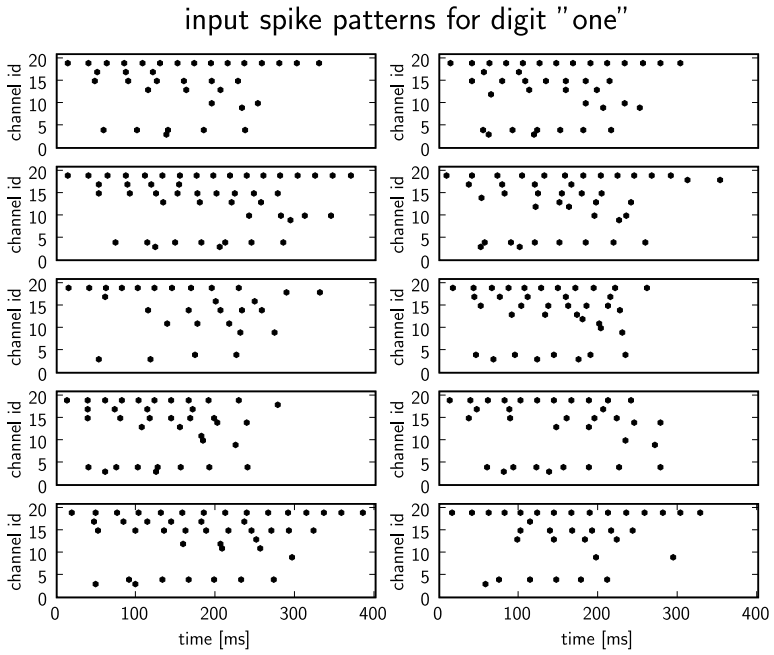
Supplementary Figure 7: Variation of Fig. 9 (i.e., of computer simulation 4) with the weight-dependent STDP rule proposed in [1]. This rule is defined by the following equations: $\Delta w_+ = \lambda w_0^{1-\mu} w^\mu e^{-|\Delta t|/\tau_+}$ and $\Delta w_- = \lambda \alpha w e^{-|\Delta t|/\tau_-}$. We used the parameters proposed in [1], i.e. $\mu = 0.4$, $\alpha = 0.11$, $\tau_+ = \tau_- = 20$ms, $\lambda = 0.1$ and $w_0 = 72.4$pS. The w_0 parameter was calculated according to the formula: $w_0 = \frac{1}{2} w_{max} \alpha^{\frac{1}{1-\mu}}$ where w_{max} is the maximum synaptic weight of the synapse. The amplitude parameters of the reward kernel were set to $\alpha_P = -\alpha_N = 1.401$. All other parameter values were the same as in computer simulation 4. The variance of the membrane potential increased for pattern P from $2.35(mV)^2$ to $3.66(mV)^2$ (panel C), and decreased for pattern N (panel D), from $2.27(mV)^2$ to $1.54(mV)^2$.



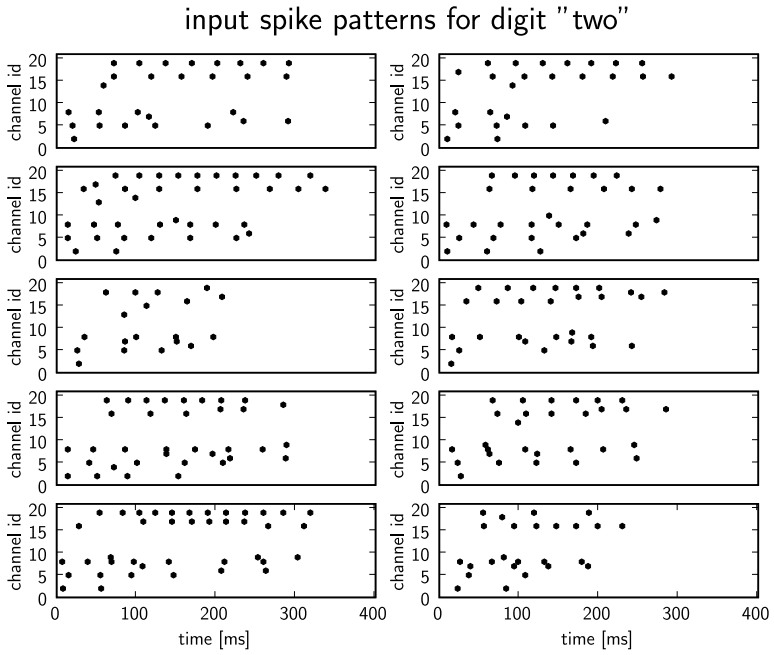
Supplementary Figure 8: Variation of Fig. 9 for a simulation where we used current-based synapses without short-term plasticity. The post-synaptic response had an exponentially decaying form $\epsilon(s) = e^{-s/\tau_\epsilon}/\tau_\epsilon$, with $\tau_\epsilon = 5ms$. The value of the maximum synaptic weight was $w_{max} = 106.2$ pA All other parameter values were the same as in computer simulation 4. The variance of the membrane potential increased for pattern P from $2.84(mV)^2$ to $5.89(mV)^2$ (panel C), and decreased for pattern N (panel D), from $2.57(mV)^2$ to $1.22(mV)^2$.



Supplementary Figure 9: Variation of Fig. 10 (i.e., of computer simulation 5) for a simulation where we used current-based synapses without short-term plasticity. The postsynaptic response had an exponentially decaying form $\epsilon(s) = e^{-s/\tau_\epsilon}/\tau_\epsilon$, with $\tau_\epsilon = 5ms$. The synaptic weights of the excitatory and inhibitory synapses in the cortical microcircuit were set to $w_{exc} = 65.4$ pA and $w_{inh} = 238$ pA respectively. The maximum synaptic weight of the synapses to the readout neuron was $w_{max} = 54.3$ pA. All other parameter values were the same as in computer simulation 5.



Supplementary Figure 10: Spike encodings of 10 utterances of digit ”one” by one speaker with the Lyon cochlea model [2], which were used as circuit inputs for computer simulation 5.



Supplementary Figure 11: Spike encodings of 10 utterances of digit ”two” by one speaker with the Lyon cochlea model [2], which were used as circuit inputs for computer simulation 5.