Learning with bounded synapses generates synaptic democracy and balanced neurons

Walter Senn and Stefano Fusi*

December 23, 2003

Submitted to Neural Computation.

Abstract. Learning in a neuronal network is often thought of as a linear superposition of synaptic modifications induced by individual stimuli. However, since biological synapses are naturally bounded, a linear superposition would cause fast forgetting of previously acquired memory. Here we show that this forgetting can be avoided by additional simple constraints. We consider Hebbian plasticity of excitatory synapses which modifies a synapse only if the postsynaptic response does not match the desired output. With this learning rule the original memory capacity with unbounded weights is regained, provided there is (1) some global inhibition, (2) a small learning rate, and (3) a small neuronal threshold. We prove, in the form of a generalized perceptron convergence theorem, that under these constraints a neuron learns to classify any linearly separable set of patterns. The maximal storage capacity is also reestablished if the synapses are distributed over a spatially extended dendritic tree, provided that distal synapses are allowed to attain stronger weights. After successful learning, excitation will roughly balance inhibition. Moreover, learning a large number of patterns urges the synapses to acquire similar strengths when measured in the soma. The fact that synapses saturate has the additional benefit that non-separable patterns, e.g. similar patterns with contradicting outputs, eventually generate a subthreshold response, and therefore silence neurons which can not provide any information.

1 Introduction

Realistic synaptic efficacies vary within a limited range of values. Synaptic saturation induced by new stimuli to be learned can provoke a rapid deterioration of the memories acquired in the past. In general, neural networks with bounded synapses are forgetful (Parisi, 1986) and the memory traces of past experiences are destroyed at a rate which is dramatically high: if one assumes that the long term changes cannot be arbitrarily small, the memory trace decays exponentially with the number

*Physiological Institute, University of Bern, Bühlplatz 5, CH-3014 Bern. Email: {wsenn,fusi}@cns.unibe.ch. Phone: +41 31 631 87 21 (ws), ...87 78 (sf).

of stored patterns. The neural network remembers only the most recent stimuli and the memory span cannot surpass a number of patterns which is proportional to the logarithm of the number of neurons (Amit and Fusi, 1994; Fusi, 2002). Slowing down the learning process by changing a small fraction of synapses solves the forgetting problem and it allows in principle to store an extensive number of random uncorrelated patterns, as is the case for unbounded synaptic strengths (Amit and Fusi, 1994). These studies were restricted to patterns with uniform statistics and fixed coding level. Moreover they focused on the maintenance of the memory trace, and not on the dynamic mechanisms to store and retrieve the information.

Here we focus on the dynamics of a biologically realistic network which distinguishes between excitation and inhibition, and which is able to store more complex patterns (i.e. linearly separable patterns, but otherwise arbitrary correlations among the components). To assure the robustness of the long-term memory against noise and smooth degradation, the synapse must be able to sustain discrete synaptic states. As can be formally proven (Senn and Fusi, 2003a), the weight assignment problem for discrete synapses can be solved by a stochastic learning rule, to the expense of an increased number of neurons. To study the meanfield dynamics of a stochastic model with discrete (binary) synapses, we focus on the case of continuous synaptic states with multiplicative saturation. Simulations with discrete synaptic states show that this meanfield description is accurate.

We consider a learning scenario in which each stimulus imposes a pattern of activities to the individual neurons of the network. The learning rule is designed to imprint the imposed activity patterns into the synaptic matrix, such that after learning a pattern can be retrieved by presenting a distorted version. Each neuron within such a network is taught in a 'supervised learning' mode since the pre- and postsynaptic activities are 'clamped' by the stimulus. The goal of the learning process is to reproduce the clamped output of a postsynaptic neuron in response to the presynaptic input pattern, or a distortion of it. When the specific value of the output activity is discarded, the postsynaptic neuron has learned to dichotomize the input patterns. It separates the inputs into two classes of patterns which either should or should not activate the postsynaptic neuron: patterns of the first class will generate a supra-threshold input, while patterns of the second class generate a sub-threshold input.

We show that a Hebbian learning rule with an additional stop-learning condition will find appropriate synaptic weights projecting onto an individual postsynaptic neuron, provided that the two classes of input patterns are linearly separable. In case of unbounded synapses a faithful learning is assured by the classical perceptron convergence theorem (see e.g. Hertz et al. 1991). The perceptron learning rule imprints the patterns into the weight vector by adding or subtracting a fraction of the input pattern, provided that the postsynaptic neuron does not yet give the required response (of 1 or 0, coding for active or inactive, respectively). With increasing number of patterns to be learned, however, the weight vector during the learning procedure becomes longer and longer. It is not clear a priori, how a local algorithm could find an appropriate weight vector if the individual components have a fixed upper and lower bound. In fact, additional requirements are necessary. Only in the presence of global inhibition, with a learning rate which is small enough, and with a neuronal threshold which is small compared to the total amount of excitation, will a faithful learning become possible. In turn, these constraints assure that any sets of linearly separable patterns can be learned by a Hebbian rule with a stop-learning condition and bounded synapses. This statement represents a generalization of the classical perceptron convergence theorem (Rosenblatt, 1962; Block, 1962; Minsky and Papert, 1969; Diederich and Opper, 1987; Arbib, 1987; Hertz et al., 1991) to the case of separate excitatory and inhibitory synapses with bounded strengths.

By imposing global inhibition and small neuronal thresholds we recover further properties which seem to be shared by the biology. A recent experimental finding shows that the effective strength of synapses projecting onto hippocampal pyramidal neurons is almost the same when measured in the soma, independently of their dendritic position (Magee and Cook, 2000). This equalization of the synaptic response in the soma is only possible if the synapses correct for the dendritic attenuation, and increase their local strength with distance from the soma. Another finding is that the somatic strength of excitatory synapses is relatively high compared to the their number and the neuronal threshold. In fact, $10'000$ afferents with a somatic amplitude of 0.2 mV and a spontaneous firing rate of 1 Hz, say, would give a depolarization of 2 mV per millisecond (see e.g. Abeles, 1991). With a voltage threshold of 20 mV this would lead to a spontaneous firing rate of roughly 100 Hz instead of 1 Hz. Only a strong balancing of excitation by inhibition can resolve this puzzle and prevent the neurons from constantly being active at a high rate.

The properties of the synaptic equalization and the neuronal balancing, as described above, emerge as a byproduct of successful learning with bounded synapses and the stop-learning condition. Such successful learning requires a small neuronal threshold to prevent the individual synapses from running into saturation. As a consequence, the total excitation will be roughly cancelled by inhibition. Moreover, overlaps in the patterns to be separated urge the synaptic weights to be roughly equal (although complete equality would fully destroy the memory). Interestingly, the constraint of bounded synaptic strengths turns out to be advantageous when dealing with non-separable sets of patterns. Due to synaptic saturation, learning similar patterns with contradicting outputs tends to erase any synaptic structure, and eventually the postsynaptic response is suppressed by the global inhibition. Such a suppression mechanism appears as a self-correcting feedback of a network in response to an overflow of unstructured inputs: instead of responding with random activity onto contradictory information, the neurons prefer to silence themselves.

2 The model

Neuron model. We consider a single postsynaptic neuron which receives excitatory inputs from N presynaptic neurons, and an inhibitory input which is proportional to the total activity of the N excitatory neurons (Fig. 1a). The postsynaptic neuron is either active or inactive, depending on whether the total postsynaptic current h is above or below the neuronal threshold $\theta_\circ$. The total postsynaptic current is calculated by the weighted sum of the excitatory synaptic input ξ_j, minus global inhibition. Global inhibition is represented by an inhibitory neuron (or a population thereof) with a linear transfer function, which uniformly sums up the excitatory input. More precisely, the total postsynaptic current is $h = \frac{1}{N}\sum_{j=1}^{N}(G_j - g_I)\xi_j$, where ξ_j can take on any value from (and including) 0 to R. The excitatory weights G_j and the global inhibitory weight g_I take on real values in the interval $[0, 1]$. In the simulations with binary synapses the excitatory weights take on values $J_j = 0$ or 1.

Training protocol. During training the network is repeatedly presented with all the p patterns ξ of two classes C^+ and C^-. At each presentation, the activities ξ_j are clamped to the N presynaptic neurons, and the output of the postsynaptic neuron is clamped to the desired response ($\xi_{post} = 0$ or 1, depending on whether ξ belongs to class C^+ or C^-, respectively). The synaptic learning rule is designed such that, after successful training, the total synaptic current h generated by a pattern ξ falls either above or below the threshold $\theta_\circ$, depending on whether ξ is in class C^+ or C^-.

Synaptic dynamics. Upon presentation of a pattern ξ the excitatory weights are modified in a Hebbian way, depending on the pre- and postsynaptic activities and the total current h. When the pre and postsynaptic cells are both active (clamped to $\xi_{post} > 0$, $\xi_j = 1$) and the total synaptic current is not too large ($h \leq \theta_\circ + \delta_\circ$, with a learning margin $\delta_\circ \geq 0$), the weight G_j is increased by $q^+\xi_j(1 - G_j)$. The weight increase is proportional to the learning rate q^+, the presynaptic activity ξ_j, and the saturation factor $(1 - G_j)$. When the presynaptic neuron is active ($\xi_j > 0$), the postsynaptic cell inactive ($\xi_{post} = 0$), and the total synaptic input not too low ($h \geq \theta_\circ - \delta_\circ$), the weight G_j is decreased by $q^-\xi_j G_j$. The weight decrease is proportional to the learning rate q^-, the presynaptic activity ξ_j, and the saturation factor G_j. Summarized, the weight change at time t writes

$$G_j^{t+1} = \begin{cases} G_j^t + q^+\,\xi_j^t(1 - G_j^t)\,, & \text{if } \xi_{post}^t = 1,\ \xi_j^t > 0, \text{ and } h^t \leq \theta_\circ + \delta_\circ\,, \\ G_j^t - q^-\,\xi_j^t G_j^t\,, & \text{if } \xi_{post}^t = 0,\ \xi_j^t > 0, \text{ and } h^t \geq \theta_\circ - \delta_\circ\,. \end{cases} \tag{1}$$

The condition onto the total synaptic current h^t represents a 'stop-learning' condition: learning stops as soon as the total synaptic current would be able to reproduce the desired postsynaptic activity (with some margin $\delta_\circ$ for overlearning).

The motivation to study the learning rule (1) comes from a probabilistic synaptic model with binary states. In this model the synapse stochastically flips its state upon presentation of a pattern ξ, depending on the conditions on the pre- and postsynaptic activities and the total current h imposed in (1). Downregulated synapses ($J_j = 0$) are potentiated with probability $q^+\xi_j$ if $\xi_{post} = 1$, $\xi_j > 0$, and $h \leq \theta_\circ + \delta_\circ$. Potentiated synapses ($J_j = 1$) downregulate with probability $q^-\xi_j$ if $\xi_j > 0$, $\xi_{post} = 0$, and $h \geq \theta_\circ - \delta_\circ$. The dynamics of the expected synaptic strengths, $G_j^t = \langle J_j^t \rangle$, can be well approximated by the dynamics (1). Note that the stochastic update can formally be described by $J_j^{t+1} = J_j^t + \zeta_j^+\,(1 - J_j^t)$ and $J_j^{t+1} = J_j^t - \zeta_j^-\,J_j^t$, respectively, where $\zeta^\pm$ are random variables which are 1 with probability $q^\pm\xi_j^t$ and 0 otherwise. Since the fluctuations of the total postsynaptic current h^t for different realizations of the stochastic process ζ typically shrink to zero with growing N, the expected total current $\langle h^t \rangle$ (which is again denoted by h^t in Eq. 1) does well approximate the actual total current h^t. A formal treatment of the stochastic model with a convergence proof for linearly separable patterns is found in (Senn and Fusi, 2003a).

3 Results

Linearly separable patterns can be learned. Given any two sets $C^{\pm}$ of linearly separable patterns, a neuron endowed with global inhibition and the stochastic learning rule described above will always learn to correctly classify the patterns in a finite number of presentations. The tighter the separation between the two classes $C^{\pm}$, the smaller the neuronal threshold $\theta_\circ$, learning margin $\delta_\circ$, and learning rate q must be (for simplicity we assume $q^+ = q^- = q$). More precisely, we assume that there is a separation vector S of length $\|S\| = N$ (not necessarily binary and positive), and a separation threshold θ, such that the classes are separated by S and θ with a positive margin (Fig. 1b). Writing this separation margin as $\delta + \epsilon$, the linear separability states that $\xi S > (\theta + \delta + \epsilon)N$ for $\xi \in C^+$, and $\xi S < (\theta - \delta - \epsilon)N$ for $\xi \in C^-$. Classification is then also possible by a separation vector which is scaled by a factor ϱ, provided that also the threshold and the margins are scaled by the same factor. These different solutions correspond to output neurons which would separate the patterns around different thresholds at the end of the training session (i.e. $h > \varrho\theta + \varrho\delta$ for $\xi \in C^+$ and $h < \varrho\theta - \varrho\delta$ for $\xi \in C^-$). However, as we show, the synaptic dynamics can only converge to a scaled separation vector if the scaling factor is small enough, $\varrho \leq \epsilon \overline{g}_I/(2R)$, where ϵ is the partial separation margin of the sets $C^{\pm}$, $\overline{g}_I = \min\{g_I, 1 - g_I\}$ the 'distance' of the inhibitory weight g_I from the boundaries 0 and 1, and R is the maximal activity of an input ξ_j (Fig. 1b). Given such a scaling factor and any global inhibition g_I between 0 and 1, the synaptic dynamics (1) converges (i.e. all the patterns will be classified correctly) in at most $n_\circ = 6/(q\varrho\epsilon\overline{g}_I)$ synaptic updates, provided that the learning rate is small enough, $q \leq \varrho\epsilon\overline{g}_I/(2R^2)$. This is valid for any presentation order of the patterns to be learned and for any initial conditions for the synaptic states. The detailed proof of the theorem is found in the Appendix.

Sketch of the proof. The idea behind the threshold scaling and the global inhibition is to keep the synaptic strength G^t far away from the lower and upper boundaries. This prevents the weight vector G^t from being distorted by synaptic saturation. Let us write the synaptic update in the form $G^{t+1} = G^t + q\Delta G^t$, where we assume equal learning rates for LTP and LTD, $q^+ = q^- = q$. The normalized change ΔG can be decomposed into a 'linear' and a 'forgetting' (saturation) part ΔL and ΔF. If the updating conditions are met we can write (1) in the form

$$\Delta G = \Delta L + \Delta F = \left\{ \begin{array}{rll} \xi * (\mathbf{1} - G) & = (1 - g_I)\xi - \xi * G_I \ , & \text{if } \xi \in C_+ \ , \\ -\xi * G & = -g_I\xi - \xi * G_I \ , & \text{if } \xi \in C_- \ , \end{array} \right. \qquad (2)$$

where $G_I = G - g_I\mathbf{1}$ and '$*$' is the componentwise product of vectors. The linear term $\Delta L = (1-g_I)\xi$ in case of $\xi \in C^+$ and $\Delta L = -g_I\xi$ in case of $\xi \in C^-$, respectively, is the learning component which is parallel to the pattern to be learned (Fig. 1c). This linear term is also present in the case of the classical perceptron learning with analog unbounded synapses, and would always bring G^t toward a solution vector. Selecting a pattern $\xi \in C^+$, for instance, we have $\xi\varrho S > \varrho(\theta+\delta+\epsilon)N$ by assumption that the solution vector S (and therefore ϱS) separates the classes. In the case that this pattern is not yet correctly implemented by the neuron, i.e. if $hN = \xi G_I < \varrho(\theta+\delta)N$, the synaptic weight vector is updated by $q\Delta G$ (as the inequality is equivalent to the update condition $h^t \leq \theta_\circ + \delta_\circ$ in Eq. 1). By subtracting this inequality from the previous one we get $(\varrho S - G_I)\xi \geq \varrho\epsilon N$. Multiplying with the factor $(1-g_I)$ and using the definition of ΔL and $\overline{g}_I = \min\{g_I, 1-g_I\}$ given above, we obtain,

$$(\varrho S - G_I)\Delta L \geq \varrho\epsilon\overline{g}_I N \ . \tag{3}$$

The same estimate (3) is obtained when $\xi \in C^-$ and ΔL has the form $-g_I\xi$. Were the forgetting part negligible, we would have $\Delta G \approx \Delta L$, and (3) would ensure that total weight vector G_I^t moves towards the solution vector ϱS, provided that the learning rate q is small. In fact, if the angle between $(\rho S - G_I)$ and ΔG is smaller than 90°, the weight vector at the next time step, $G_I + q\Delta G$, is always closer to the target vector ρS than G_I was, assuring that q is small enough (Fig. 1d).

The forgetting part ΔF in the decomposition $\Delta G = \Delta L + \Delta F$ (Eq. 2) takes the form $\Delta F = -\xi G_I$ for both of up- and downregulation. It arises from the synaptic saturation and tends to bring $G_I = G - g_I$ towards 0, where $G_j = g_I$ for all j. In this asymptotic limit no structure would be present in the synaptic weight vector, showing that synaptic saturation might neutralize previous learning steps (see Fig. 1c). However, synaptic saturation is strongly reduced and can become negligible if all the weights are far from the boundary. This is the case if the weight vector is close to the main diagonal where all the synaptic strengths are roughly equal. If the uniform component is subtracted by the global inhibition, and if the neuronal threshold is small, the remaining structure in the weight vector is enough to separate the patterns.

The supervised learning, together with the small neuronal threshold $\theta_\circ$, enforces the dynamics to reach the region of the diagonal where synaptic saturation is negligible. Given the separation threshold θ and the separation parameter ϵ of the two classes, the neuronal threshold leading to a correct separation must be in the range of $\epsilon\theta$. More precisely, the convergence of the weight vector is guaranteed with a threshold $\theta_\circ = \varrho\theta$, provided that $\varrho \leq \epsilon\overline{g}_I/(2R)$. In fact, it is possible to show that $(\varrho S - G_I)\Delta F \geq -\varrho^2 RN$, and that for small ϱ the distortion by the synaptic saturation therefore vanishes. Together with (3) we obtain $(\varrho S - G_I)(\Delta L + \Delta F) > 0$, asserting that the effective synaptic change $\Delta G = \Delta L + \Delta F$, including the forgetting term, points towards the target vector ϱS. Hence, provided that ϱ is small, convergence of the learning procedure is guaranteed as outlined above.

Global inhibition and a small threshold are necessary. To test the statement of the theorem and to show the necessity of the different requirements we consider a simple numerical example. We randomly chose a set of $p = 10$ patterns ξ with activities ξ_j uniformly distributed between 0 and 40 (=R, e.g. in units of spikes/s) and components $j = 1, \ldots, N = 20$. The excitatory synaptic weights G_j of the 20 synapses were randomly initialized between 0 and 1. The threshold and the margin ($\theta_\circ = 5.2$ and $\delta_\circ = 0.08$) were set such they separate the 10 patterns ξ into two classes of 5, after projection to a random separation vector S. As predicted, the separation of the postsynaptic current, $h^t > \theta_\circ + \delta_\circ$ and $h^t < \theta_\circ - \delta_\circ$ for patterns in C^+ and C^-, respectively, is reached after a few synaptic updates (cf. Fig. 2a). The simulation confirms that learning makes always some progress due to its linear part, in the sense that in case of a synaptic update we have $(\varrho S - G_I)\Delta L > 0$, Eq. 3, while the forgetting (saturation) part may work against this progress as $(\varrho S - G_I)\Delta F$ can become negative (Fig. 2b).

The value of the global inhibition plays a crucial role. As predicted by the theorem, many more learning steps are necessary if g_I is close to the boundary 0 or 1 (Fig. 3a). In fact, the theorem predicts that the number of synaptic updates required to learn the patterns is roughly $n_\circ \propto \frac{1}{\overline{g}_I} \approx \frac{1}{g_I(1-g_I)}$. The chance of finding a configuration of excitatory synapses which balance inhibition shrinks when g_I tends to a boundary value. Similarly, only when the neuronal threshold is small, expressed by a small threshold scaling factor, will it be possible to converge to a solution (Fig. 3b). The simulation result is expressed by the requirement $\varrho \leq \epsilon \overline{g}_I/(2R)$ appearing in the theorem (see also Fig. 1b). If global inhibition is kept away from 0 and 1, the drawback of synaptic saturation is fully compensated, provided that the learning rate and the threshold are sufficiently small.

Intuitively, global inhibition is necessary since the separation of the patterns into two classes may require, for instance, assigning an output $\xi_{post} = 0$ to a pattern with high coding (activity) level $f = \frac{1}{N}\sum_{j=1}^{N} \xi_j$ (many presynaptic neurons strongly active). This would not be possible with excitatory synapses alone because a pattern with a high coding level would always lead to a suprathreshold response. However, if the global activity level is subtracted by the global inhibition, $h = \sum G_j \xi_j - g_I f = \sum (G_j - g_I)\xi_j$, then the assignment of the output 0 becomes possible, even if the activity level of the pattern is high (choose $G_j < g_I$ for components j with strong input ξ_j). Intuitively, a small threshold is necessary because tightly separated classes (ϵ small) require that small differences in the inputs ξ_j, independently of the size of ξ_j, may turn a subthreshold response into a suprathreshold response.

A small learning rate and the stop-learning condition are necessary. To prevent overshooting of the target vector ϱS, the learning rates $q^\pm$ $(= q)$ must be small enough. A monotonic convergence towards the target vector is expected if the learning rate is small compared to the neuronal threshold. Since the threshold itself scales with the separation parameter ϵ, the learning rate must scale, for instance, with ϵ^2. In fact, the convergence is guaranteed if $q \leq \varrho \epsilon \overline{g}_I/(2R^2)$ (cf. Fig. 4a). The requirement of a small threshold is also confirmed by the simulations, see (Senn and Fusi, 2003a; Senn and Fusi, 2003b).

Learning is also severely impaired if the stop-learning condition on the total postsynaptic current h^t in (1) is not imposed. Only if the learning process stops when the desired output is reached is it possible to learn any set of separable patterns. Otherwise, the dynamics may learn a dominant cluster of patterns while other patterns far from such a cluster may fall off from the correct classification (Fig. 4a). In fact, dropping the stop-learning condition leads to sustained oscillations in the total postsynaptic currents and no further learning progress is achieved (Fig. 4b). Although the ongoing learning prevents the weights from being settled in an appropriate state, the synaptic weights tend to be equalized by the forgetting part, $(\varrho S - G_I)\Delta F \to 0$ (decaying curve in Fig. 4b). Any learning rule which is able to learn tightly separated classes must incorporate some form of stopping condition.

Learning equalizes synaptic strengths and balances inputs. Another physiological prediction is that the synaptic strengths, due to their boundedness, become more similar to each other, the more difficult the separation task. This is because during learning the algorithm must find a synaptic configuration for which the detrimental effect of synaptic saturation is weak compared to the difficulty of the separation task. The tighter the two classes C^+ and C^- are separated, the less distortion by synaptic saturation can be afforded, and the more uniform the distribution becomes. A relatively uniform distribution of the excitatory synaptic weights G_j around the value of the global inhibition g_I is enforced by *a priori* choosing a small threshold (small scaling factor ϱ) depending on the separation margin of the classes to be learnt (cf. Fig. 1b,c). The balancing of excitation and inhibition and the equalization of the synaptic weights appears as a byproduct of efficient learning.

The balancing and the equalization of the synaptic weights are confirmed by our simulations. Due to the random initialization, the weights originally span the whole possible range of values (Fig. 5a). After a few synaptic updates evoked by the initially incorrectly classified patterns, the weights all adopted roughly the same value (Fig. 5b, solid line). Note that the weaker excitatory weights are cancelled by inhibition, in the sense that $G_j - g_I < \theta_\circ$ (dotted line). If the learning algorithm is not able to separate the patterns, for instance because the stop-learning condition is discarded, the weight equalization does not fully develop (dashed-dotted line).

Equalized synaptic strength and dendritic democracy. The equalization of the synaptic strength also works when the synapses are distributed across a dendritic tree. In this case, it is the amplitude of an excitatory postsynaptic potential (EPSP) measured in the soma which is equalized by the local learning rules. This is because it is at the level of the soma where the total synaptic input is compared with the neuronal threshold, and where the postsynaptic signal for the modification of the synaptic strengths is triggered. However, to locally control the degree of synaptic saturation, distal synapses must be allowed to attain stronger weights. In fact, if the upper bound imposed on the synaptic strength linearly increases with the distance from the soma, our theorem can be reformulated in the context of learning on a dendritic tree (by scaling the synaptic strength G_j by an attenuation factor a_j, and its upper bound by $1/a_j$).

To test the undiminished learning capabilities in the presence of the dendritic tree we distributed $N = 350$ synapses along a cable of $350 \mu m$ length, showing attenuation factors a_j from 1 down to 0.2 (Magee and Cook, 2000), and upper bounds $1/a_j$ for the maximally achievable synaptic strength. The initial synaptic weights, measured at the corresponding dendritic site, were uniformly distributed between 0.2 and 0.6 (Fig. 5c). The neuron was trained with $p = 350$ random activity patterns with individual presynaptic activities ξ_j ranging uniformly from 0 to 40 (e.g. in Hz), and a random splitting of these patterns into two classes C^+ and C^-, requiring a supra- and subthreshold postsynaptic response, respectively. The neuron learned to correctly separate the patterns within a total of 3023 presentations. After faithful learning the somatically recorded synaptic weights $(G_j a_j)$ were equalized, while the local synaptic weights (G_j) increased with distance from the soma (Fig. 5d). We again assumed that global inhibition is fixed $(g_I = 0.1)$, and that it directly projects onto the soma, where the rough balance between excitation and inhibition is established.

Conflicting patterns shut down neuronal activity. An interesting property of (multiplicative) synaptic saturation is that it tends to stabilize the synaptic weights. This property can be advantageous when dealing with unstructured patterns, or with similar patterns requiring different outputs, since in these cases it leads to a uniform excitatory weight distribution around some common equilibrium weight. If this excitatory equilibrium weight is dominated by the global inhibition, the neuron will no longer respond to these patterns, and therefore not try to make an impossible classification of unstructured or conflicting stimulus.

To be more concrete, we stimulate our neuron with a set of input patterns which repeatedly lead to sequential potentiations and depressions of the same synapses. According to the update rule (1) the equilibrium weight of synapse j is then determined by the equation

$$\Delta G_j = \tilde{q}^+(1 - G_j) - \tilde{q}^- G_j = 0 \; , \qquad (4)$$

where $\tilde{q}^\pm$ represent the effective rates of up- and down-regulations. These rates are the product of the learning rates $q^\pm$, the expected presynaptic activity $\langle \xi_j \rangle$, and the relative frequency of requiring a postsynaptic response 1 or 0, respectively. Solving (4) for G_j gives the unique equilibrium weight $G^* = \tilde{q}^+ / (\tilde{q}^+ + \tilde{q}^-)$. This equilibrium is an attractor of (4), as shown by the negative derivative of ΔG_j with respect to G_j at the fixed point, $\frac{d\Delta G_j}{dG_j} = -\tilde{q}^+ - \tilde{q}^-$. Whatever the initial synaptic weight is, the saturation factors $(1 - G_j)$ and G_j in (4) always drive the synapse to the unique steady state. If the equilibrium weight is dominated by the global inhibition, $G^* < g_I$, the total postsynaptic current would become negative in response to an arbitrary stimulus ξ, $h = \frac{1}{N} \sum (G_j - g_I) \xi_j < 0$. Taking the stop-learning condition into account, however, the weights G_j are only depressed until the lower learning threshold $\theta_\circ - \delta_\circ$ is reached, $h = \frac{1}{N} \sum (G_j - g_I) \xi_j \approx \theta_\circ - \delta_\circ$. In any case, trying to learn different outputs to similar input patterns, will eventually lead to a subthreshold activation.

The neuronal suppressing mechanism is confirmed by the simulations. As an example we show the evolution of the total postsynaptic currents h^t for the case of 5 pairs of identical patterns $\xi^{\pm}$ (i.e. $p = 10$ and identical classes $C^+ = C^-$). As predicted, the total postsynaptic currents eventually become, or remain, subthreshold for all patterns (Fig. 6a). The downward drift of the total postsynaptic current h^t is caused by the synaptic saturation which strongly homogenizes the synaptic weights until excitation is dominated by the global inhibition (Fig. 6c). In fact, without synaptic saturation (mimicked by cancelling the forgetting part $\Delta F = -\xi G_I$ in the update rule (2)), the suppression effect vanishes and the total postsynaptic currents incoherently become either sub- or suprathreshold (Fig. 6b). This is also reflected in the uncontrolled growth of the synaptic weights beyond the upper boundary (Fig. 6d). Hence, teaching the neuron with different outputs for the same patterns will uniformly depress the synaptic weights and silence the neuron.

Convergence for binary synapses with stochastic modifications. We finally provide a partial account of the results that learning with discrete synapses converges in a finite number of steps, provided that 1) the number of neurons is large enough, 2) the low learning rate is replaced by low transition probabilities between stable discrete states. If these two conditions are satisfied, the average values of the discrete synapses are well described by the continuous synaptic variables introduced in the Model. As a consequence, the convergence of the learning process is well predicted by the theorem in the Appendix (see Senn and Fusi, 2003a, for more details and for extensive simulations with highly correlated patterns).

Simulations with binary synapses projecting to a single output cell confirm that the stochastic learning rule is successful (Fig. 7). The parameters of the learning dynamics are the same as in the simulations of the deterministic example (Fig. 2), and the activities ξ_j of the 10 patterns are either 0 or 40, with probability of 0.5. As expected, the convergence in the stochastic case is more noisy and it takes a larger number of presentations than in the case with continuous synapses (Fig. 7a). With increasing number of neurons, however, the prediction of the synaptic dynamics by the meanfield equation (1) becomes more reliable. The redundancy in the synaptic encoding speeds learning up until it approaches the convergence speed of continuous-valued synapses. In fact, the number of presentations per pattern, required to correctly classify the stimuli, shrinks with increasing number of presynaptic neurons towards an asymptotic value (Fig. 7b).

4 Discussion

We showed that despite the synaptic boundedness and restriction of plasticity to the excitatory synapses, any set of linearly separable patterns can be learned with Hebbian plasticity incorporating a stop-learning condition. These biologically plausible restrictions, however, require that (1) there is some global inhibition, (2) a small learning rate, and (3) a threshold which is small compared to the the overall excitatory synaptic strengths. The restrictions are shown to be necessary to prevent fast forgetting which may arise during the learning process by driving the synaptic strengths into saturation. As a byproduct of learning, the synaptic strengths roughly (but not fully) equalize, and a rough balancing between the total excitation and inhibition emerges. Synaptic saturation further causes a neuron to suppress its activity if it is learned with similar patterns, but opposing outputs.

The stop-learning condition protects from overlearning. The stop-learning condition is necessary to not lose previously acquired memory when repeatedly presenting the same or similar patterns. There are many ways of implementing such a stopping mechanism. It could be inherent to the individual synapse, governed by the postsynaptic neuron, or depend on an external feedback. For instance, the synapse may not undergo potentiation if the pre- and postsynaptic activities and the postsynaptic calcium concentration are each above some critical level (cf. also (Fusi, 2003; Amit and Mogillo, 2003)). The stop-learning condition might also be implemented by an anti-Hebbian term in the learning rule. Unfortunately, experimental data leave the question of such an intrinsic nonlinearity open (see e.g. (Cho et al., 2001; Rumsey and Abbott, 2003)). Another possibility would be that the stop-learning signal is carried by a third signal, for instance in the reduction of dopamine release, as observed after successful reinforcement learning (see e.g. Fiorillo et al., 2003). A similar stop-learning phenomenon is observed in V4 of a monkey performing a delayed match-to-sample task, where no learning effect is seen if the visual stimuli are not degraded by noise and easy to classify (Rainer et al., 2003).

Global inhibition sets the range of excitatory weights. Global inhibition is a general property often assumed in neural networks to normalize the total synaptic input. In fact, recent experimental findings show that inhibitory neurons in neocortex, but also in hippocampus, may form a large network, tightly coupled through gap junctions (see e.g. Amitai et al., 2002). In our framework, inhibition defines a range, far from saturation, into which the excitatory weights will tend during learning. Inhibition must be global to assert that any set of linearly separable patterns with any correlations (i.e. clustering of the patterns) can be learned. Non-global inhibition could make it difficult to learn a specific set of correlated patterns when plasticity is restricted to only excitatory synapses. In fact, non-global inhibition may lead to a strong and unequal forgetting across the synapses due to unequal synaptic saturation, unless also inhibition is plastic.

Slow learning prevents fast forgetting. Slow learning becomes important if the set of patterns to be learned is large. This is because a slow learning prevents the synaptic weights from overshooting, but also from heading off into the saturation regime. In the continuous-valued synaptic model, slow learning is implemented by a small learning rate (q). However, biological synapses do not admit arbitrarily small changes. Synapses must operate with discrete states, allowing them to benefit for the long-term maintenance of their strengths. In a discrete-valued synaptic model, slow learning is achieved by a stochastic selection of a small number of synapses to be modified. In general, networks with bounded synapses which do not allow arbitrarily small changes share the palimpsest property (Fusi, 2002): new patterns to be learnt overwrite the oldest ones, and only a limited number of patterns are remembered. When learning is slow, in the sense that only a few synapses are changed with each presentation of a stimulus, the speed of overwriting is small and more patterns are remembered. Slow learning also occurs in biology, for instance in infero-temporal and perirhinal cortex (Miyashita, 1993; Yakovlev et al., 1998; Erikson and Desimone, 1999). The slow learning observed in these experiments is consistent with the implicit task for the monkey to maximize the number of memorized patterns.

Small neuronal thresholds allow to separate similar patterns. The assumption of a small neuronal threshold relative to the total excitatory synaptic strength seems to be satisfied in biology by virtue of the huge number of excitatory synapses projecting onto a single neuron (Abeles, 1991). As we showed, the ratio between the neuronal threshold and the total excitatory synaptic strength must decrease with the difficulty of the learning task, i.e. with decreasing separation margin between the two classes to be learned. The correct tuning of this threshold-to-synaptic strength ratio could be performed by additional homeostatic processes (see e.g. Desai et al., 2002). Homeostatic plasticity may also tune the global inhibition (g_I) to dominate the excitatory equilibrium weight (G^*), such that neurons silence themselves in response to unstructured input.

Synaptic democracy increases memory capacity. The tendency to equalize the synaptic weights when learning tightly separated classes may also underly the effective weight equalization (measured as EPSP amplitudes in the soma) in hippocampal pyramidal cells (Magee and Cook, 2000; Häusser, 2001). These cells are thought to perform an associative memory task, and therefore need to classify presynaptic activity patterns as described here. However, the cells will only reach the maximal storage capacity if distal synapses have the same chance to evoke an action potential as proximal synapses have. To allow distal synapses to increase their strengths, without disproportional saturation deteriorating the storage capacity, the upper bound imposed on the synaptic strength must grow with distance from the soma. As we showed, synaptic democracy then automatically emerges while learning tightly separated classes of input patterns. The synaptic equalization is caused by the synaptic boundedness, which tends to uniform the somatically measured synaptic strength, and the stopping conditions, which prevents an overlearning of repeatedly presented stimuli. The neuron can therefore regain its maximal storage capacity, despite the strong distortions of the synaptic inputs by the dendritic attenuation. An alternative model explaining the emergence of synaptic equalization was recently suggested by combining Hebbian and anti-Hebbian spike-timing dependent plasticity (Rumsey and Abbott, 2003). It would be interesting to reconcile this bottom-up with the present top-down approach, and specify how Hebbian and anti-Hebbian plasticity may jointly improve learning (e.g. of output distributions as in the Boltzmann machine) or maximize the memory capacity (as achieved here by the stop-learning condition).

Weight variability and noise robustness Our assumption that the local inhibitory weights are uniform and fixed requires that they directly project onto the soma. The proximal location of inhibitory input is consistent with experimental findings in hippocampal pyramidal cells (see e.g. Pouille and Scanziani, 2001). However, the somatically measured excitatory synaptic strengths in biology seem to show a larger variance than predicted by the present theory (see e.g. Magee and Cook, 2000). A possible explanation for this larger variance could be that biological neurons avoid to exploit the maximal, theoretically available, memory capacity. This is reasonable in the presence of noise, since otherwise the effective synaptic strengths would become too equal and the response of the neurons would become noise sensitive. A strategy to achieve noise robustness is to increase the strength of the global inhibition. This is because strong global inhibition would start to suppress the neuronal responses already before the synaptic strengths became too uniform, and the Hebbian modifications would therefore stop while some variance in the synaptic weights remains.

Silenced neurons allow to deal with non-separable patterns. Suppressing the activity of a neuron which receives contradictory information is an important property when dealing with non-separable patterns. As the number of random input patterns increases ($p > 2N$), the chance that they are inseparable, and therefore not classifiable by a neuron, also increases (Cover, 1965). A mechanism is therefore required which protects the neuron from an overflow of data, and eventually avoids misclassifications. The studied suppression mechanism represents an intrinsic self-regulation of a neuron against such an overflow. Non-separable patterns homogenize the synaptic weights by means of the synaptic saturation, until their response becomes suppressed by the global inhibition. The same suppression mechanism can also be exploited to improve the classification of noisy data like handwritten digits (Senn and Fusi, 2003a). Since patterns which are incorrectly classified typically evoke a subthreshold response (false negative), the classification can be improved by adding several stochastic output neurons in parallel. This naturally leads to a sparse representation of the input patterns, as often seen in cortical recordings (see e.g. Vinje and Gallant, 2000). We conclude that synaptic saturation, and the way to prevent it or to make use of it, may have important implications onto the neuronal organization of the cortex.

Appendix

Perceptron convergence theorem for bounded synapses

The theorem asserts that with the classical Hebbian rule incorporating a stop-learning condition any set of linearly separable patterns can be learned with bounded synaptic strengths, provided that the learning rate is small, that there is some global inhibition, and that the neuronal threshold is small compared to the overall sum of the presynaptic excitatory weights. For notational convenience we consider equal learning rates for LTP and LTD, $q^- = q^+ = q$.

Theorem. *Let $C^\pm$ be any sets of linearly $(\delta + \epsilon)$-separable activity patterns $\xi \in [0, R]^N$ with separability threshold $\theta \in \mathbf{R}$ and separability parameters $\delta \geq 0$, $\epsilon > 0$. Let us choose any globally inhibitory weight $g_I \in (0, 1)$, any scaling factor $\varrho \leq \epsilon\overline{g}_I/(2R)$, and any learning rate $q \leq \varrho\epsilon\overline{g}_I/(2R^2)$, where $\overline{g}_I = \min\{g_I, 1-g_I\}$. Set the threshold of the postsynaptic neuron to $\theta_\circ = \varrho\theta$, and the learning margin to $\delta_\circ = \varrho\delta$. Then, for any repeated presentation of the patterns $\xi \in C^\pm$ and any initial condition $G_j^0 \in [0, 1]^N$, the synaptic dynamics* (1) *converges in at most $n_\circ = 6/(q\varrho\epsilon\overline{g}_I)$ synaptic updates.*

Note that the maximal number of stochastic updates, $n_\circ$, which is required to learn the patterns, is independent of the number of patterns p to be learned. This apparent paradox arises because $n_\circ$ only counts the number of presentations which trigger synaptic updates, i.e. for which the update conditions in (1) are satisfied. Since the patterns satisfying these conditions are not known a priori, however, an online algorithm needs to repeatedly cycle through all the p patterns. Hence, for a periodic cycling, an upper bound for the number of presentations, t, until learning stops is $t_\circ = pn_\circ = 6p/(q\varrho\epsilon\overline{g}_I)$.

Proof of the theorem. The condition on the linear separability of the sets $C^\pm$ states that there is an $S \in \mathbf{R}^N$ with $\|S\|^2 = N$ and a separation threshold $\theta \in \mathbf{R}$ such that $\xi S > (\theta + \delta + \epsilon)N$ for $\xi \in C^+$ (i.e. $\xi_{post} = 1$), and $\xi S < (\theta - \delta - \epsilon)N$ for $\xi \in C^-$ (i.e. $\xi_{post} = 0$). Writing the learning rule (1) in the form $G^{t+1} = G^t + q\Delta G^t$ and assuming that the conditions for a synaptic update are satisfied, we can decompose (1) into the linear and forgetting part according to (2). Recall that the condition for a synaptic update is satisfied if either $h = \frac{1}{N} G_I \xi \leq \varrho(\theta + \delta)$ or $h = \frac{1}{N} G_I \xi \geq \varrho(\theta - \delta)$ for $\xi \in C^+$ and $\xi \in C^-$, respectively.

Learning with the linear part. According to the update and separability condition for the case $\xi \in C^+$ we have $\xi G_I < \varrho(\theta + \delta)N$ and $\xi \varrho S > \varrho(\theta + \delta + \epsilon)N$, respectively. Subtracting the first from the second inequality we get $(\varrho S - G_I)\xi \geq \varrho\epsilon N$. Similarly, for the case $\xi \in C^-$ we have the two conditions $\xi G_I > \varrho(\theta - \delta)N$ and $\xi\varrho S < \varrho(\theta - \delta - \epsilon)N$, respectively, and by subtraction we get $-(\varrho S - G_I)\xi \geq \varrho\epsilon N$. Defining the linear part in the learning rule (2) by $\Delta L = \xi(1 - g_I)$ in case of $\xi \in C^+$ and $\Delta L = -g_I \xi$ in case of $\xi \in C^-$, respectively, we get the basic inequality (3) presented previously in the main text,

$$(\varrho S - G_I)\Delta L \geq \varrho\epsilon\overline{g}_I N \; . \tag{3}$$

Controlling the forgetting part. We next estimate the impact of the forgetting (saturation) term $\Delta F = -\xi * G_I$. We show that updating G with $q\Delta F$ either supports learning (in the sense of Eq. 3), or at least does not move G_I too far away from ϱS. Inserting the definition of ΔF, writing $\xi = \sqrt{\xi} * \sqrt{\xi}$ and applying twice the Cauchy-Schwartz inequality in the form $x\,y \leq \|x\|\,\|y\|$, with equality if $x = y$, we get sequentially

$$\begin{aligned}(\varrho S - G_I)\Delta F &= G_I(\xi * G_I) - \varrho S(\xi * G_I) = \left(\sqrt{\xi} * G_I\right)^2 - \varrho\left(\sqrt{\xi} * S\right)\left(\sqrt{\xi} * G_I\right) \\ &\geq \|\sqrt{\xi} * G_I\| \left(\|\sqrt{\xi} * G_I\| - \varrho\,\|\sqrt{\xi} * S\|\right) \; . \end{aligned} \tag{5}$$

When forgetting supports learning. In the case of $\|\sqrt{\xi} * G_I\| \geq \varrho\,\|\sqrt{\xi} * S\|$, the parenthesis on the right-hand side of (5) is non-negative, and one immediately concludes from (5) that $(\varrho S - G_I)\Delta F \geq 0$. Note that the condition on the norm of G_I roughly states that G_I lies ‘behind’ ϱS when looking from the origin in the direction of $\sqrt{\xi}$. In this case the forgetting term ΔF speeds up, or at least does not counteract, the convergence of G_I towards ϱS. In fact, since $\Delta G = \Delta L + \Delta F$ we obtain from $(\varrho S - G_I)\Delta F \geq 0$, together with (3), that for any ϱ,

$$(\varrho S - G_I)\Delta G \geq \varrho\epsilon\overline{g}_I N \ , \ \text{provided} \ \|\sqrt{\xi} * G_I\| \geq \varrho\,\|\sqrt{\xi} * S\| \ . \tag{6}$$

When forgetting counteracts learning. We next consider the case that $\|\sqrt{\xi} * G_I\| \leq \varrho\,\|\sqrt{\xi} * S\|$. Inserting this into (5), while neglecting the term $\|\sqrt{\xi} * G_I\|$ in the parenthesis on the right-hand side, we get the estimate

$$(\varrho S - G_I)\Delta F \geq -\|\sqrt{\xi} * G_I\|\,\varrho\,\|\sqrt{\xi} * S\| \geq -\varrho^2\,\|\sqrt{\xi} * S\|^2 \geq -\varrho^2 RN \ . \tag{7}$$

For the last inequality we used the definition of the norm square, the fact that $\xi_j \leq R$, and the assumption on the separation vector that $\|S\|^2 = N$ to obtain $\|\sqrt{\xi} * S\|^2 = \sum_{i=1}^N \xi_i S_i^2 \leq R \sum_i S_i^2 = RN$. Since the above estimate cannot exclude that $(\varrho S - G_I)\Delta F$ becomes negative, we cannot preclude that forgetting counteracts learning. However, since the scaling factor ϱ enters as the square, forgetting becomes disproportionally weak if ϱ gets small. Let us choose $\varrho \leq \epsilon\overline{g}_I/(2R)$. Using again $\Delta G = \Delta L + \Delta F$ we then get from estimate (7), together with (3), that

$$(\varrho S - G_I)\Delta G \geq \varrho N(\epsilon\overline{g}_I - \varrho R) \geq \varrho\epsilon\overline{g}_I N/2 \ , \ \text{provided} \ \|\sqrt{\xi} * G_I\| \leq \varrho\,\|\sqrt{\xi} * S\| \ . \tag{8}$$

Learning in the general case stops. We next show that with each synaptic update the distance from G_I to ϱS decreases at least by some fixed quantity. We conclude that the learning process must terminate, since otherwise the distance from G_I to ϱS would become negative. Let t_μ denote the time(s) when pattern ξ^μ is presented and the synapses are updated. At a subsequent time step $t_\mu + 1$ there is $G_I^{t_\mu+1} = G_I^{t_\mu} + q\Delta G^{t_\mu}$. Combining (6) and (8) we estimate $(\varrho S - G_I^{t_\mu})\Delta G^{t_\mu} \geq \varrho\epsilon\overline{g}_I N/2$, independently of the value of $\|\sqrt{\xi} * G_I\|$. Substituting $G_I^{t_\mu+1}$ in the following line, multiplying the norm squares out, inserting $(\varrho S - G_I^{t_\mu})\Delta G^{t_\mu} \geq \varrho\epsilon\overline{g}_I N/2$, and choosing a learning rate $q \leq \varrho\epsilon\overline{g}_I/(2R^2)$ yields

$$\begin{aligned} \|\varrho S - G_I^{t_\mu+1}\|^2 - \|\varrho S - G_I^{t_\mu}\|^2 &= -2q(\varrho S - G_I^{t_\mu})\Delta G^{t_\mu} + q^2\|\Delta G^{t_\mu}\|^2 \leq \ldots \\ \ldots &\leq qN(qR^2 - \varrho\epsilon\overline{g}_I) \leq -q\varrho\epsilon\overline{g}_I N/2 \ . \end{aligned} \tag{9}$$

Note that by definition of ΔG, see (2), we have $\|\Delta G^{t_\mu}\|^2 \leq R^2 N$. This is because the synaptic weights $G_j^{t_\mu}$ are between 0 and 1, and the stimuli ξ_j^μ are between 0 and R. Summing up the contributions of all the updates up to time t evoked by the different patterns, $G_I^t = G_I^0 + q\sum_{t'_\mu < t} \Delta G^{t'_\mu}$, while repeatedly using estimate (9), we get an estimate of the telescope sum

$$\begin{aligned} \|\varrho S - G_I^t\|^2 - \|\varrho S - G_I^0\|^2 &= \|\varrho S - G_I^t\|^2 - \|\varrho S - G_I^{t-1}\|^2 + \\ &\quad \|\varrho S - G_I^{t-1}\|^2 - \|\varrho S - G_I^{t-2}\|^2 + - \ldots \\ &\leq -n_t q\varrho\epsilon\overline{g}_I N/2 \,, \end{aligned} \tag{10}$$

where n_t is the number of synaptic updates up to the t-th presentation of a pattern. From (10) we immediately obtain

$$0 \le \|\varrho S - G_I^t\|^2 \le \|\varrho S - G_I^0\|^2 - n_t q \varrho \epsilon \overline{g}_I N/2\,. \tag{11}$$

Since $\|\varrho S - G_I^0\|^2 \le (\varrho^2 + g_I^2 + 1)N \le 3N$ we conclude from (11) that $\|\varrho S - G_I^t\|^2 \le 0$ after $n_t = 6/(q\varrho\epsilon\overline{g}_I)$ updates. Hence, the number of synaptic updates until learning stops must be smaller $n_\circ = 6/(q\varrho\epsilon\overline{g}_I)$. If we set $\varrho = \epsilon\overline{g}_I/(2R)$ and $q = \varrho\epsilon\overline{g}_I/(2R^2)$, consistent with the smallness requirements above, we obtain $n_\circ = 48(R/\epsilon\overline{g}_I)^4$. Note that this estimate is independent of the initial state of the synaptic weight vector $G^0 \in [0,1]^N$. q.e.d.

Acknowledgements: This work was supported by the SNF grant 3152-065234.01, the Silva Casa foundation, and the EU grant IST-2001-38099 ALAVLSI. We would like to thank Nicolas Brunel for helpful comments and for checking the theorem, Massimo Mascaro for motivating discussions, and Joe Brader for proof reading.

References

Abeles, M. (1991). *Corticonics.* Cambridge University Press.

Amit, D. & Fusi, S. (1994). Learning in neural networks with material synapses. *Neural Computation*, 6:957–982.

Amit, D. & Mogillo, G. (2003). Spike-driven synaptic dynamics generating working memory states. *Neural Computation*, 15:565–596.

Amitai, Y., Gibson, J., Beierlein, M., Patrick, S., Ho, A., Connors, B., & Golomb, D. (2002). The spatial dimensions of electrically coupled networks of interneurons in the neocortex. *Journal of Neuroscience*, 22:4142–4152.

Arbib, M. (1987). *Brains, Machines, and Mathematics.* Springer Verlag, Berlin.

Block, H. (1962). The perceptron: a model for brain functioning. *Reviews of Modern Physics*, 34:123–135. Reprinted in: Anderson and Rosenfeld (eds.), *Neurocomputing: Foundations of Research.*

Cho, K., Aggleton, J., Brown, M., & Bashir, Z. (2001). An experimental test of the role of postsynaptic calcium levels in determining synaptic strength using perirhinal cortex of rat. *J. Physiology*, 532.2:459–466.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with application in pattern recognition. *IEEE Transaction on Electronic Computers*, 14(3):326–334.

Desai, N., Cudmore, R., Nelson, S., & Turrigiano, G. (2002). Critical periods for experience-dependent synaptic scaling in visual cortex. *Nature Neuroscience*, 5(8):783–789.

Diederich, S. & Opper, M. (1987). Learning of correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett.*, 58:929–952.

Erikson, C. & Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. *J. Neurosci*, 19:10404–10416.

Fiorillo, C., Tobler, P., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299:1898–1902.

Fusi, S. (2002). Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. *Biol. Cybernetics*. "Hebb in Perspective".

Fusi, S. (2003). Spike-driven synaptic plasticity for learning correlated patterns of mean firing rates. *Reviews in the Neurosciences*, 14:73–84.

Häusser, M. (2001). Synaptic function: dendritic democracy. *Current Biology*, 11:R10–R12.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison Wesley.

Magee, J. & Cook, E. (2000). Somatic EPSP amplitude is independent of synapse location in hippocampal pyramidal neurons. *Nature Neuroscience*, 3(9):895–903.

Minsky, M. L. & Papert, S. A. (1969). *Perceptrons*. MIT Press. Expanded edition 1988.

Miyashita, Y. (1993). Inferior temporal cortex: where visual perception meets memory. *Ann. Rew. Neurosci.*, 16:245–263.

Parisi, G. (1986). A memory which forgets. *Journal of Physics A - Mathematical & General*, 19(10):L–617–620.

Pouille, F. & Scanziani, M. (2001). Enforcement of temporal fidelity in pyramidal cells by somatic feed-forward inhibition. *Science*, 293:1159–1163.

Rainer, G., Lee, H., & Logothetis, N. (2003). The effect of learning on the function of monkey extrastriate visual cortex. page in press.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan Books.

Rumsey, C. & Abbott, L. (2003). Equalization of synaptic efficacy by activity- and timing-dependent synaptic plasticity. *Journal of Neurophysiology*, page In press.

Senn, W. & Fusi, S. (2003a). Learning with discrete-valued bounded synapses without forgetting. In preparation.

Senn, W. & Fusi, S. (2003b). Slow stochastic learning with global inhibition: a biological solution to the binary perceptron problem. *Neurocomputing*. To appear.

Vinje, W. & Gallant, J. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287:1273–1267.

Yakovlev, V., Fusi, S., Berman, E., & Zohary, E. (1998). Inter-trial neuronal activity in infero-temporal cortex: a putative vehicle to generate long term associations. *Nature Neuroscience*, 1:310–317.

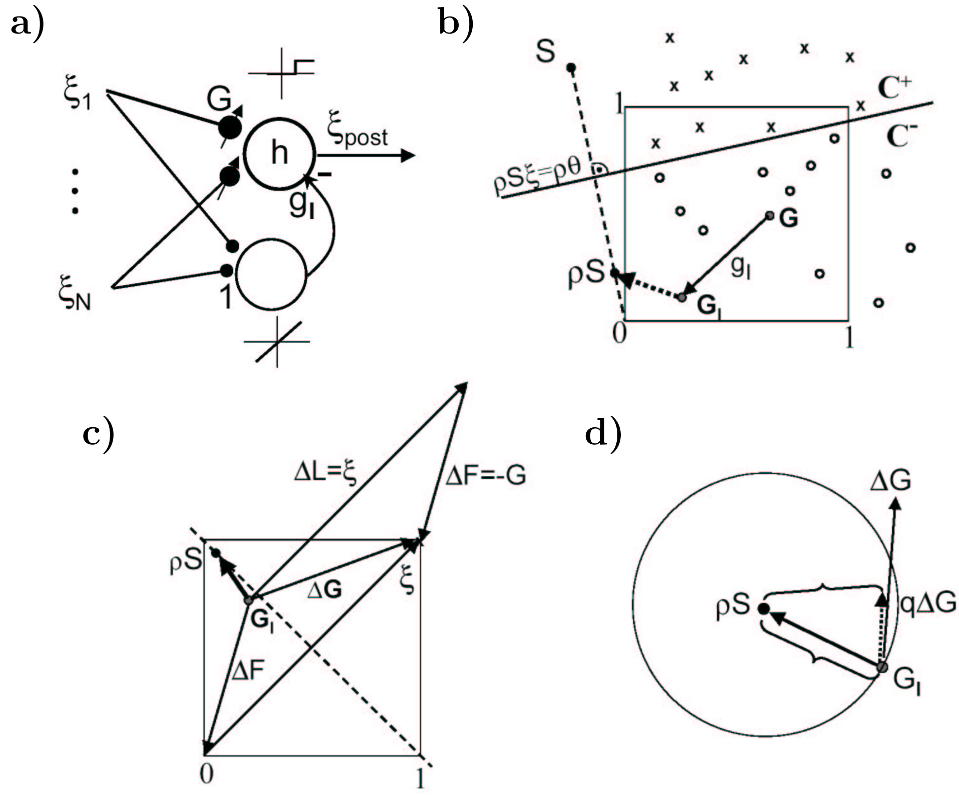

Figure 1: Neuronal architecture and sketch of the convergence proof. **a)** We consider a postsynaptic neuron receiving direct excitatory input from N presynaptic neurons (ξ_j), and indirect input through a inhibitory neuron with linear input-output relationship. The excitatory weights (G_j) are subject to Hebbian plasticity with weight saturation and a stop-learning condition. The globally inhibitory weight (g_I) is fixed. The postsynaptic response (ξ_{post}) is the thresholded total synaptic current h, but any other nonlinear input-output relationship which dichotomizes the input is also possible. **b)** The sets C^+ (crosses) and C^- (circles) of patterns ξ are assumed to be linearly separable, with a separation vector S and a threshold θ. Since S may contain negative components and components larger than 1, it cannot in general be approximated by the excitatory weight vector G. Only if the solution vector S (and with it the threshold θ) is scaled down by ϱ, and if some global inhibition g_I is present, is it possible to approximate the solution vector, $\varrho S \approx G_I = G - g_I\mathbf{1}$, with a G which is far from saturation at the boundaries 0 and 1 of the hypercube. **c)** The synaptic change ΔG triggered by pattern ξ is decomposed into a linear and forgetting (saturation) part, $\Delta G = \Delta L + \Delta F$. Without global inhibition ($g_I = 0$ and $G_I = G$), synaptic saturation (ΔF) may prevent the weight vector G to be updated in the 'correct' direction ΔL, in the sense that $(\varrho S - G_I)\Delta G > 0$. In the shown example we have $(\varrho S - G_I)\Delta G < 0$, i.e. the update moves G_I away from the solution vector ϱS. This is because an update of G_I in the desired direction ΔL is distorted by the nearby boundaries and, instead, G_I moves in the direction of $\Delta G = \Delta L + \Delta F$ towards the upper right corner. Such a distortion is not possible if G is close to the main diagonal and far from $\mathbf{0}$ and $\mathbf{1}$ (achieved by a small ϱ, and a g_I in between 0 and 1, see a). **d)** A positive scalar product $(\varrho S - G_I)\Delta G > 0$ ensures that the G_I moves towards ϱS, provided that the learning rate q is small (distance indicated by the upper brace is smaller than that indicated by the lower brace).

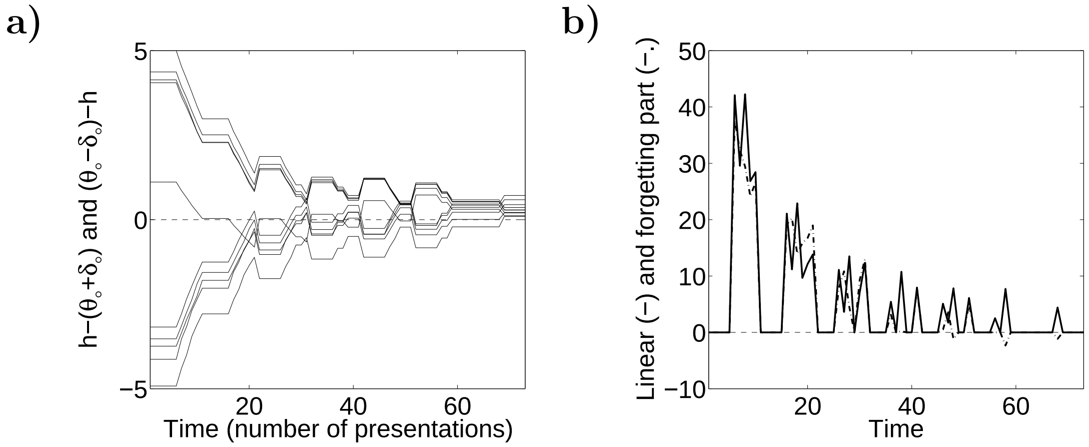

Figure 2: Any linearly separable set of patterns is learnable with limited synaptic strengths. **a)** Evolution of the signed distance between the total postsynaptic current and the learning threshold, $h^t(\xi) - (\theta_\circ + \delta_\circ)$, for patterns ξ of class C^+, and $(\theta_\circ - \delta_\circ) - h^t(\xi)$, for patterns of class C^-. According to the update condition (Eq. 1), learning stops as soon as these quantities become all positive, here after a total of 69 pattern presentations (out of which 27 satisfied the condition on h^t and let to synaptic updates). Note that the monotonic convergence of the total weight vector G_I^t towards the scaled solution vector ϱS does not imply that for all patterns the total input $h^t(\xi)$ monotonically converges. Further parameters: $q = q^\pm = 2 \cdot 10^{-3}$, $\rho = 0.3$, $g_I = 0.5$. The same set of patterns is used in the subsequent Figures 3-6. **b)** Evolution of the learning progress represented by the 'linear part', $(\varrho S - G_I^t)\Delta L^t$, (solid line) and the 'forgetting part', $(\varrho S - G_I^t)\Delta F^t$, (dashed-dotted line). The quantities represent the learning progress due to the non-saturating and saturating part: they indicate by how much the two learning components ΔF and ΔG move the weight vector G_I towards the target vector ϱS. The flat parts correspond to presentations which did not trigger synaptic updates because the patterns were already correctly implemented, and the condition on h^t in the update rule (1) therefore was not satisfied. As shown in the proof, the linear part always supports learning, $(\varrho S - G_I^t)\Delta L^t > 0$, while the forgetting part may counteract learning when G_I^t comes close to ϱS, as happens at the 48^{th}, 58^{th} and 68^{th} presentation, where $(\varrho S - G_I^t)\Delta F^t < 0$. Such forgetting could become dominant if the threshold (the scaling factor ϱ) were not small enough.

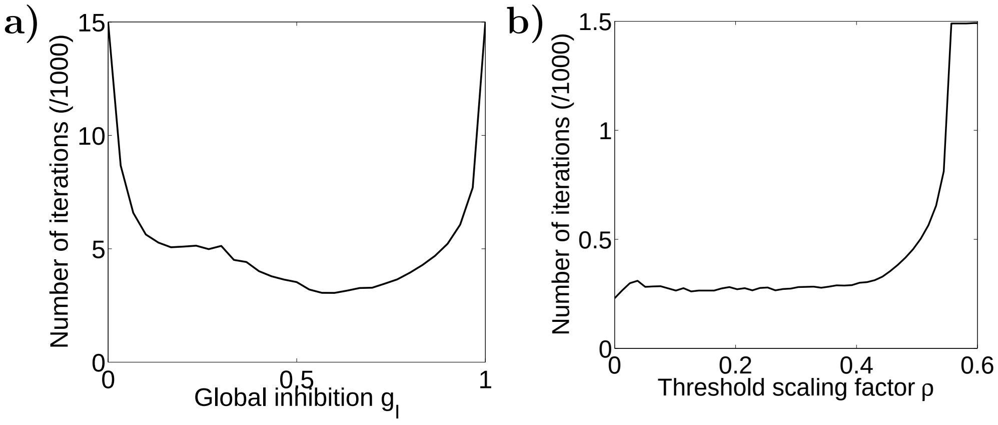

Figure 3: Learning requires global inhibition and a small scaling factor. **a)** The number of iterations (in thousands) required to learn the random set of patterns is minimal if the global inhibitory strength g_I is roughly 0.5, as predicted by the theory. A inhibitory weight close to 0 or 1 urges the excitatory weights to 'catch up' the inhibitory weight, and the emerging synaptic saturation ('forgetting') strongly impairs the learning (cf. Fig. 1c). The learning rate and the scaling factor were reduced by a factor of 100, yielding $q = 2 \cdot 10^{-5}$ and $\varrho = 0.003$, such that it is still possible to separate the patterns with values of the global inhibition near 0 and 1. **b)** Number of synaptic updates (in thousands) required for convergence as a function of the scaling factor ϱ, with the same learning rate q as in a). As predicted by the theory, learning is impaired if the neuronal threshold, compared to the total (excitatory) synaptic strength, is not small ($\varrho > 0.5$, cf. Fig. 1c).

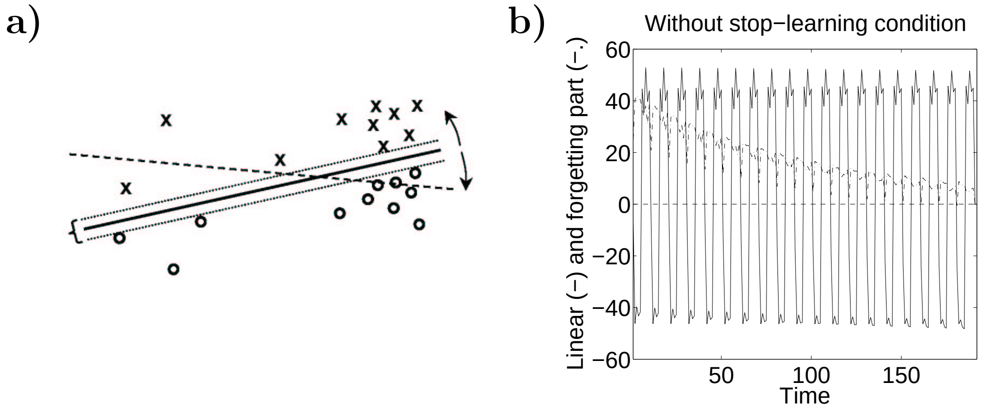

Figure 4: Individual synaptic modifications should be small and triggered only if the required response is not matched. **a)** To prevent overshooting, the learning rates $q^{\pm}$ must be a fraction of the separation parameter ϵ (width of the bracelet: $2(\epsilon + \delta)$, corresponding to the separation margin between the two classes, as indicated by the parallel dotted lines). Without stop-learning condition the weight vector G^t would be repeatedly attracted by the clusters (as appearing on the right), while patterns not in these clusters start to get misclassified (as the cross most left). The dashed line shows the separation hyperplane after learning the cluster of crosses. A subsequent learning of the cluster of circles would move the hyperplane up again (arrows). **b)** Same plot as in Fig. 2b, but without stop-learning condition on the total postsynaptic current h^t in (1). The linear part oscillates because the weight vector G periodically 'overlearns' the patterns, i.e. is repeatedly attracted towards one cluster of patterns and thereby starts to misclassify other patterns. In contrast, the forgetting part slowly converges, showing that the final weight vector oscillates close to the main diagonal where synaptic saturation is minimal and the weights are roughly equalized.

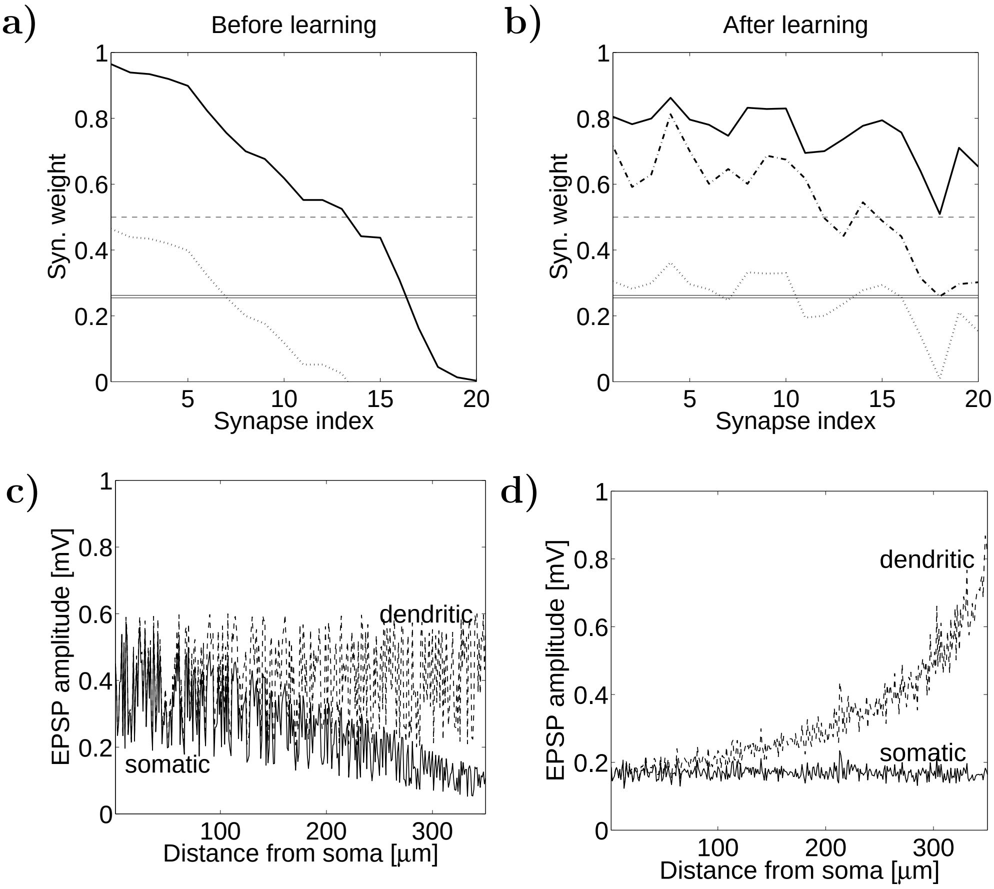

Figure 5: Balancing and equalization of the synaptic weights through learning. **a)** The initial synaptic strengths G_j (solid line) span the whole possible interval between 0 and 1, scaled up by N. The two narrowly separated black lines represent the learning thresholds $\theta_\circ \pm \delta_\circ$, divided by the average presynaptic activity of all patterns, $R/2$, to be comparable with the individual synaptic weights. The dashed line at $g_I = 0.5$ represents the global inhibitory weight (dotted line: $G_j - g_I$). **b)** After faithful learning of the set of 10 patterns in 27 synaptic updates (69 presentations, see Fig. 2) the excitatory synaptic strengths G_j became roughly equal (solid line). Subtracting global inhibition (dotted line) makes the effective synaptic weights fluctuating around the threshold. If the stop-learning condition is not imposed, the weights equalize much less (dashed-dotted line, shown after 200 synaptic updates). **c, d)** Learning on a dendritic tree. **c)** Initial synaptic strength measured locally at the dendritic site ('dendritic EPSP amplitude', G_j, uniformly distributed), and in the soma ('somatic EPSP amplitude', $G_j a_j$, decreasing with distance, see text). **d)** After faithful learning the somatically measured synaptic strengths are equalized, while the dendritically measured strengths increase with distance from the soma. Further parameter values: $\rho = 0.08$, $\theta_\circ = 1.4$, $\delta_\circ = 6 \cdot 10^{-4}$, $q^\pm = 5 \cdot 10^{-5}$.

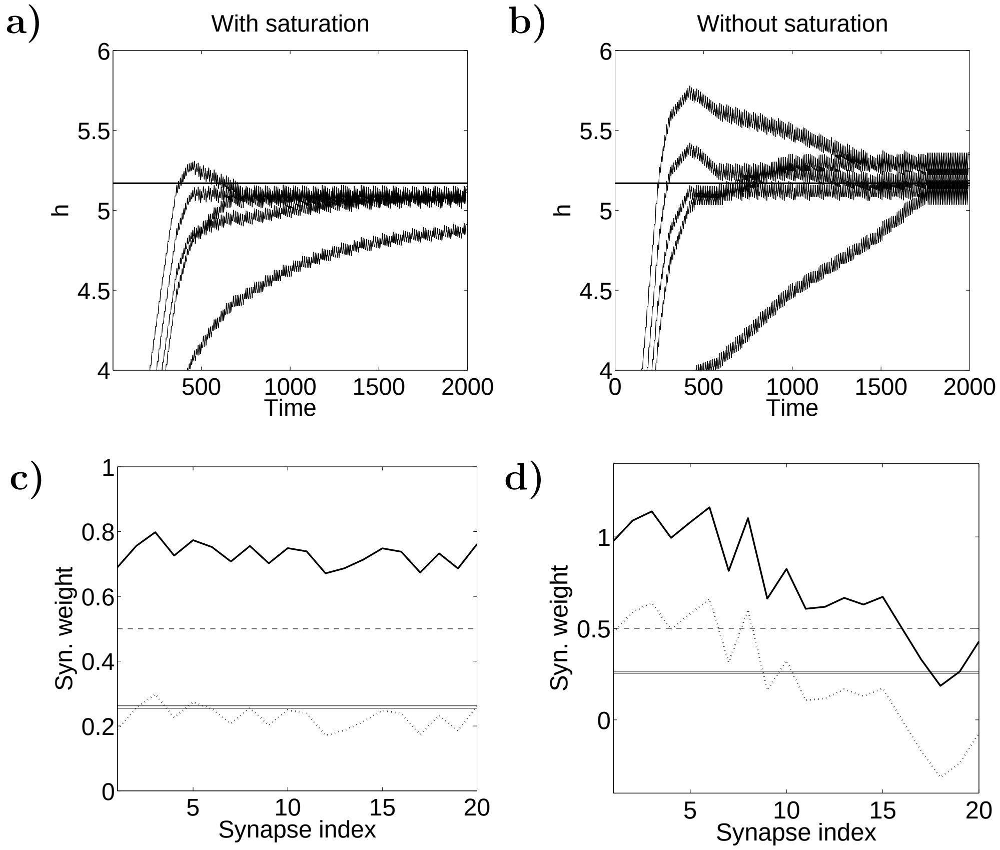

Figure 6: Synaptic saturation suppresses neuronal activity in response to conflicting patterns. **a)** Evolution of the total postsynaptic current h^t in response to the 5 patterns trained with conflicting outputs, i.e. requiring once the output $\xi_{post} = 0$ and once $\xi_{post} = 1$ for the same input patterns. After a transient response (around update 200) the total postsynaptic currents of all the 5 patterns becomes subthreshold (horizontal line represents the neuronal threshold $\theta_\circ$). **b)** Without synaptic saturation (modelled by setting $\Delta F = 0$ in Eq. 2) the postsynaptic currents do not become subthreshold. **c, d)** The final distribution of the synaptic weights G_j (solid lines) corresponding to the simulations in a) and b) with and without saturation, respectively (same initial weights as in Fig. 5a). Dashed line: global inhibition, g_I; double solid line: neuronal threshold scaled by the presynaptic mean activity, $2\theta_\circ/R$; dotted line: $G_j - g_I$. Learning the contradicting outputs homogenizes the weights in the presence of synaptic saturation, and leads to the uniform dominance of inhibition, and therefore to the suppression of any neuronal activity (c). The final weight distribution when the upper synaptic bound was relieved does not show the homogenization, and therefore does not lead to the activity suppression (d).

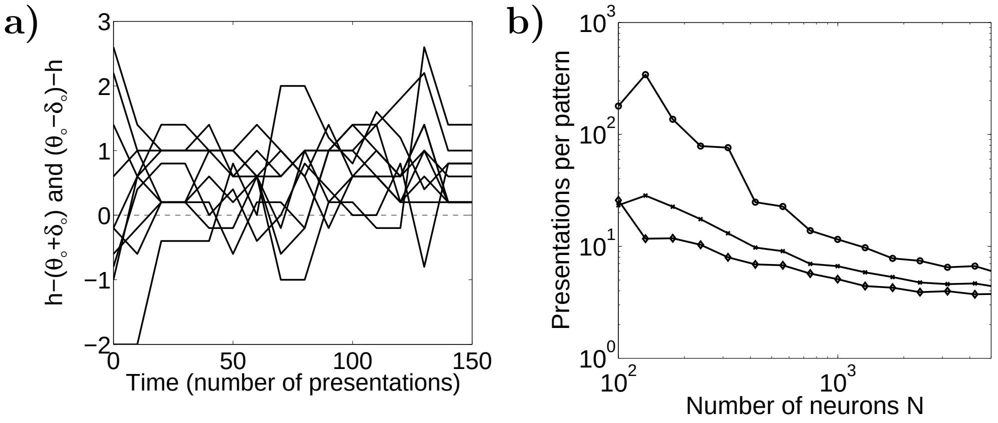

Figure 7: Convergence of stochastic learning in the case of binary synapses. **a)** Total synaptic input current h^t as a function of time, evaluated for all 10 random, linearly separable patterns. Same parameters as in Fig. 2, except for the number of neurons which is $N = 100$ instead of $N = 20$. The learning process converges in about 150 presentations (15 presentations per stimulus). **b)** Number of presentations per pattern required for convergence, as a function of the number of neurons N, for $p = 10$, 20, 40 random binary 0/1 patterns with coding level $f = 1/4$. Other parameters: $q^{\pm} = .05$, $g_I = 0.5$. The classes are constructed to be linearly separable. The neuronal threshold $\theta_\circ$ and the learning margin $\delta_\circ$ are chosen to yield a maximal separation of the classes after projecting the patterns to a solution vector S.