LETTER Communicated by Misha Tsodyks

Learning Only When Necessary: Better Memories of Correlated Patterns in Networks with Bounded Synapses

Walter Senn
wsenn@cns.unibe.ch

Stefano Fusi
fusi@cns.unibe.ch
Department of Physiology, University of Bern, CH-30 Bern, Switzerland

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 memories. Here we show that this forgetting can be avoided by introducing additional constraints on the synaptic and neural dynamics. We consider Hebbian plasticity of excitatory synapses. A synapse is modified only if the postsynaptic response does not match the desired output. With this learning rule, the original memory performances with unbounded weights are regained, provided that (1) there is some global inhibition, (2) the learning rate is small, and (3) the neurons can discriminate small differences in the total synaptic input (e.g., by making the neuronal threshold small compared to the total postsynaptic input). 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, including a wide class of highly correlated random patterns. During the learning process, excitation becomes roughly balanced by inhibition, and the neuron classifies the patterns on the basis of small differences around this balance. The fact that synapses saturate has the additional benefit that nonlinearly separable patterns, such as similar patterns with contradicting outputs, eventually generate a subthreshold response, and therefore silence neurons that cannot 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 that is dramatically high: if one assumes that the long-term changes cannot be arbitrarily small, the memory trace decays exponentially with the number of stored patterns. The neural network remembers only the most recent stimuli, and the memory span cannot surpass a number of patterns that is proportional to the logarithm of the number of neurons (Amit & Fusi, 1994; Fusi, 2002). Slowing the learning process by changing a small fraction of synapses solves the forgetting problem, and it allows, in principle, storing an extensive number of random uncorrelated patterns, as in the case of unbounded synaptic strengths (Tsodyks, 1990; Amit & Fusi 1992, 1994; Brunel, Carusi, & Fusi, 1998). However, these studies were restricted to patterns with uniform statistics and fixed coding level (i.e., with the same average number of active neurons per pattern). Moreover, they focused on the maintenance of the memory trace, not on the dynamic mechanisms to store and retrieve information. More recent papers show that it is possible to store and retrieve real-world patterns in networks of excitatory and inhibitory neurons(see, e.g., Amit & Mascaro, 2001). However, in all these works, the internal state of each synapse has an unreasonably large number of stable states, and the synaptic dynamics are almost unaffected by the boundaries.

Here we study the dynamics of a biologically realistic network with distinct excitatory and inhibitory neurons, which is able to learn linearly separable patterns. A previous work (Amit, Wong, & Campbell, 1989) addressed the problem of separation between excitation and inhibition, but it did not consider the problems of realistic synapses whose weights are limited from above and from below. In this work, we assume that the synapses are bounded, and they do not allow arbitrarily small changes. The qualitative behavior of networks with realistic synapses does not strongly depend on the number of synaptic states that can be preserved on long timescales (Amit & Fusi, 1994; Fusi, 2002). Hence we consider the extreme case of binary synapses. As can be formally proven (Tsodyks, 1990; Amit & Fusi, 1992, 1994; Senn & Fusi, 2005), the weight assignment problem for binary synapses can be solved by a stochastic learning rule, provided that the number of input neurons is large compared to the number of patterns to be stored. To study the mean field dynamics of a stochastic model with binary synapses, we focus on the case of continuous synaptic states with multiplicative saturation. Simulations with discrete synaptic states show that this mean field description is accurate.

We consider a learning scenario in which each stimulus imposes a pattern of activities to the neurons of the network. Given a specific activity

pattern as an input to the neuron, the desired output is known and provided by a teacher (supervised learning). The teacher signal indicating the right response might come from a different cortical area (e.g., a top-down signal that encodes the class to which the current sensory stimulus belongs), or it can be provided by the sensory stimulus (e.g., when a pattern of activity is imposed to all neurons of a recurrent network, in which each neuron can be regarded as both an input neuron and an output neuron). The learning rule is designed to embed the imposed activity patterns into the synaptic matrix. After learning, each pattern seen during training can be retrieved without mistakes. In the case of a feedforward network, this means that each input pattern produces the correct response indicated by the teacher during the training. For a recurrent network, each pattern imposed by the sensory stimuli becomes a fixed point of the network dynamics. Under additional stability conditions, these fixed points can also be attractors of the network dynamics. In what follows, we restrict our analysis to only two distinct responses of the output neurons and to feedforward networks.

We show that a Hebbian learning rule with an additional stop-learning condition will find the appropriate synaptic weights to produce the desired response of a single output unit, provided that the two classes of input patterns are linearly separable. In case of unbounded synapses with the stop-learning condition, a successful learning is ensured by the classical perceptron convergence theorem (Rosenblatt, 1962; Block, 1962; Minsky & Papert, 1969; Diederich & Opper, 1987; Arbib, 1987; Hertz, Krogh, & Palmer, 1991). The perceptron learning rule embeds the patterns into the weight vector by adding or subtracting the input vector, provided that the postsynaptic neuron does not yet give the required response. This is shown to give good classification performance on real-world patterns (Amit & Mascaro, 2001). In all these cases, the weight vector becomes longer and longer as more patterns are learned. It is not clear a priori how a local algorithm could find an appropriate weight vector if the individual components are restricted within rigid boundaries. The simplicity of the convergence proof for the classical perceptron rule hides several problems that would naturally emerge in any realistic neural network. In particular, the unboundedness of the synapses allows arbitrarily large weights. Many of the parameters that control the convergence of the classical perceptron rule should be scaled to bring back the synaptic weights into a limited range (see section 4 for more details). Additional requirements are therefore necessary to guarantee the convergence when the synaptic weights are bounded. This is particularly the case when excitatory synapses have only two stable states corresponding to two different excitatory efficacies. Only in the presence of global inhibition, with

a learning rate that is small enough, and with a neuronal threshold that is small compared to the total amount of excitation, will a successful learning become possible. These constraints ensure that any set of linearly separable patterns can be learned by a Hebbian rule with a stop-learning condition and bounded synapses.

Learning with bounded synapses and the stopping condition has other interesting consequences. It is well known that in the spontaneous activity state, the total postsynaptic current produced by only the excitatory synapses is relatively high compared to 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. 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, as already pointed out in several works (see, e.g., van Vreeswijk & Sompolinsky, 1996; Amit & Brunel, 1997). Balanced excitation and inhibition emerge as a by-product 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). In the case of binary synapses, these overlaps would cause equal probabilities of being potentiated.

Surprisingly, the constraint of bounded synaptic strengths turns out to be advantageous when dealing with nonseparable 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 tends to shut down neurons that are trained with inconsistent teaching signals, as it arises during training with nonseparable patterns. This suppression mechanism prevents an erroneous activation of an output neuron.

2 The Model

2.1 Neuron Model. We consider a single postsynaptic neuron that receives excitatory inputs from N presynaptic neurons, and an inhibitory input that is proportional to the total activity of the N excitatory neurons (see

Figure 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 a global inhibition. Global inhibition is represented by an inhibitory neuron that sums all the excitatory inputs with the same weight. The activity of this inhibitory neuron is assumed to be proportional to the total (excitatory) input (linear transfer function). In a less abstract network, the inhibitory neuron would be represented by a population of inhibitory cells, with random connections from the excitatory inputs and random connections to the outputs (see, e.g., Amit & Brunel, 1997). Formally, the total postsynaptic current of the output neuron is $h = \frac{1}{N}\sum_{j=1}^{N}(G_j - g_I)\xi_j$, where ξ_j can be any value from 0 to R. Notice that the net effect of the inhibitory population can be regarded as a synaptic shift, which also allows negative weights. The components ξ_j of an input pattern can be interpreted, for instance, as the firing rate of the presynaptic neurons. 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.

2.2 Training Protocol. During training, the input neurons are repeatedly presented with all the p patterns ξ of two classes C^+ and C^-. With each presentation, the activities ξ_j are imposed to the N presynaptic neurons, and the postsynaptic neuron is clamped to the desired response (by setting $\xi_{post} = 0$ 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 should fall either above or below the threshold $\theta_\circ$, depending on whether ξ is in class C^+ or C^-.

2.3 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 (postsynaptic) current h. When the post- and presynaptic 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 postsynaptic cell is inactive ($\xi_{post} = 0$), the presynaptic neuron is active ($\xi_j > 0$), and the total synaptic input not too low ($h \geq \theta_\circ - \delta_\circ$), then 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

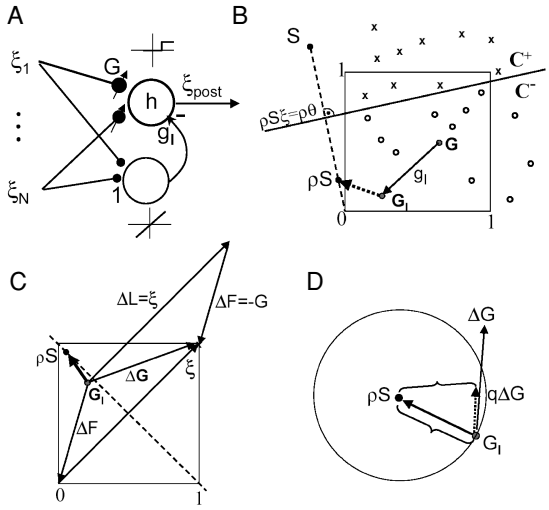

Figure 1: Neuronal architecture and sketch of the convergence proof. (A) We consider a postsynaptic neuron receiving direct excitatory input from N presynaptic neurons (ξ_j), and indirect input through a inhibitory neuron with linear input-output relationship. The excitatory weights (G_j) are subject 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 that 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 that 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 from being updated in the "correct" direction ΔL, in the sense that $(\varrho S - G_I)\Delta G > 0$. In the example, we have $(\varrho S - G_I)\Delta G < 0$; 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 $\Delta G = \Delta L + \Delta F$ toward 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 toward ϱS, provided that the learning rate q is small (distance indicated by the upper brace is smaller than that indicated by the lower brace).

writes

$$G_j^{t+1} = \begin{cases} G_j^t + q^+ \xi_j^t (1 - G_j^t), & \text{if } \xi_{post}^t = 1 \text{ and } h^t \le \theta_\circ + \delta_\circ, \\ G_j^t - q^- \xi_j^t G_j^t, & \text{if } \xi_{post}^t = 0 \text{ and } h^t \ge \theta_\circ - \delta_\circ. \end{cases} \tag{2.1}$$

The condition on 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). If the condition on h^t in equation 2.1 with $\xi_{post}^t = 0$ or 1 is met, we speak of a synaptic update. Notice that the synaptic dynamics is entirely determined by four parameters: $q^+, q^-, \theta_\circ, \delta_\circ$. The neuronal dynamics requires an additional parameter g_I setting the global inhibition.

The motivation to study learning rule 2.1 comes from a probabilistic synaptic model with binary states. In this model, the synapse stochastically flips its state on presentation of a pattern ξ, depending on the conditions on the pre- and postsynaptic activities and the total current h. Downregulated synapses ($J_j = 0$) are potentiated with probability $q^+ \xi_j$ if $\xi_{post} = 1$, $\xi_j > 0$, and $h \le \theta_\circ + \delta_\circ$. Potentiated synapses ($J_j = 1$) downregulate with probability $q^- \xi_j$ if $\xi_j > 0$, $\xi_{post} = 0$, and $h \ge \theta_\circ - \delta_\circ$. The dynamics of the expected synaptic strengths, $G_j^t = \langle J_j^t \rangle$, can be well approximated by the dynamics in equation 2.1. 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 that 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 equation 2.1) well approximates 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 & Fusi (2005).

3 Results

3.1 Linearly Separable Patterns Can Always Be Learned. Given any two sets $C^\pm$ of linearly separable patterns, a neuron endowed with global inhibition and bounded synapses obeying the mean field dynamics of equation 2.1 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 $\theta_\circ$, the learning margin $\delta_\circ$, and the 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 (see Figure 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 $\xi \in C^-$. Notice that θ and δ characterize the statistics of the patterns to be classified. They should not be confused with $\delta_\circ$ and $\theta_\circ$, which are parameters of the synaptic dynamics. Classification is then also possible for all separation vectors ϱS that are scaled by a factor ϱ, provided that the threshold and the margins are also scaled by the same factor. These different solutions correspond to output neurons that 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 will converge (to a scaled separation vector) only if the scaling factor is small enough, $\varrho \leq \epsilon\overline{g}_I/(2R)$, where ϵ is the partial separation margin of the sets $C^\pm$, $\overline{g}_I = \min\{g_I, 1 - g_I\}$ is 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 (see Figure 1B). The final weight vector to which the synaptic dynamics converges depends on the parameters $\theta_\circ$ and $\delta_\circ$ of the synaptic dynamics. To guarantee the convergence of the learning process, we need these two parameters to be small enough, $\theta_\circ \leq \varrho\theta$, $\delta_\circ \leq \varrho\delta$. Note that the threshold scaling factor ϱ is not a dynamic variable of the learning process. The parameters $\theta_\circ$ and $\delta_\circ$ are always chosen at the beginning of the learning process and are never changed. However, the learning process will actually converge only if $\theta_\circ$ and $\delta_\circ$ are chosen properly, with a size that depends on the separation margin (ϵ) of the patterns. The theorem guarantees that there is always a range of scaled thresholds for which the learning process converges: if the scaling factor ϱ and the learning rate $q = q_\pm$ are small enough, then for any global inhibition g_I between 0 and 1 (i.e., between the minimal and maximal excitatory weights), the synaptic dynamics 2.1 converges (i.e., all the patterns will be correctly classified) in at most $n_\circ = 6/(q\varrho\epsilon\overline{g}_I)$ updates of the synaptic weight vector. When the smallness conditions on q and ϱ are also considered (see the appendix), this amounts to an upper bound for the number of updates in the order of $1/\epsilon^4$. This bound is valid for any presentation order of the patterns to be learned and for any initial conditions for the synaptic states. The rigorous formulation and proof of the theorem is found in the appendix.

3.2 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 long-term

potentiation (LTP) and long-term depression (LTD), $q^+ = q^- = q$. The normalized change ΔG can be decomposed into a linear and a forgetting (saturation) part ΔL and ΔF, respectively. If the updating conditions are met, we can write equation 2.1 in the form

$$\Delta G = \Delta L + \Delta F = \begin{cases} \xi * (\mathbf{1} - G) = (1 - g_I)\xi - \xi * G_I, & \text{if } \xi \in C^+, \\ -\xi * G = -g_I\xi - \xi * G_I, & \text{if } \xi \in C^-, \end{cases} \tag{3.1}$$

where $G_I = G - g_I\mathbf{1}$ and $*$ is the component-wise product of vectors and $\Delta F = -\xi * G_I$. 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 (see Figure 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, that is, if $hN = \xi G_I < \varrho(\theta + \delta)N$, the synaptic weight vector is updated by $q\,\Delta G$ according to equation 2.1 (note that the last inequality is equivalent to the update condition $h^t \le \theta_\circ + \delta_\circ$ in equation 2.1). By subtracting this inequality from the previous one, we get $(\varrho S - G_I)\xi \ge \varrho\epsilon N$. 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,

$$(\varrho S - G_I)\Delta L \ge \varrho\epsilon\overline{g}_I N. \tag{3.2}$$

The same estimate, equation 3.2, 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 equation 3.2 would ensure that total weight vector G_I^t moves toward 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 degrees, 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, ensuring that q is small enough (see Figure 1D).

In general, the forgetting part $\Delta F = -\xi * G_I$ is not negligible. Note that this term is the same for both up- and downregulations (see equation 3.1). It arises from the synaptic saturation and tends to bring $G_I = G - g_I$ toward 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 Figure 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. 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 equation 3.2, 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 toward the target vector ϱS. Hence, provided that ϱ is small, convergence of the learning procedure is guaranteed as outlined above.

3.3 Convergence for Bounded Versus Unbounded Synapses. In the classical perceptron, the smallness of the neuronal threshold and the synaptic parameters (θ, δ, q) is not required because the synaptic weights can grow unboundedly. In fact, when increasing the number of new patterns p to be learned, the maximum synaptic weight $G_{\max}$ constantly increases (with $\sqrt{p}$ in case of random patterns; see Figure 8 and section 4). This is because each synaptic update pushes the effective weight vector G_I in the direction of the separation vector. The smaller the separation margin between the two classes to be separated, the larger the maximum weight becomes, $G_{\max} \sim 1/\epsilon$. A renormalization of the synaptic weights after learning would similarly lead to a small threshold, as it is necessary in the current framework with bounded synapses.

This renormalization is also changing the estimate of the convergence time. For unbounded synapses, the number of synaptic updates required for convergence is inversely proportional to the learning rate and the separation margin, $n_\circ \sim 1/(q\epsilon)$. It is inversely proportional to ϵ because the component of the synaptic update vector in the direction of the separation vector cannot be larger than the difference of the overlaps between the separation vector and the two classes, $|S(\xi^+ - \xi^-)| < \epsilon$ for $\xi^\pm \in C^\pm$. It is inversely proportional to q because, by definition, the length of the synaptic update vector is proportional to the learning rate. To prevent overshooting, however, q must be smaller than ϵ (see Figure 1D), yielding an upper bound of $n_\circ \sim 1/\epsilon^2$ in the case of unbounded synapses. This is in agreement with the estimate of the convergence time for the classical perceptron (see, e.g., Hertz et al., 1991). Since for learning with unbounded synapses the

maximum weight grows like $1/\epsilon$ (see Figure 8), we may obtain an a posteriori solution for the dynamics with bounded synapses by scaling all the synaptic parameters (θ, δ, q) by $\varrho = \epsilon$. As a consequence, the effective synaptic weight vector ($G_I = G - g_I\mathbf{1}$) approaches the scaled solution vector ϱS, and the learning progress per synaptic update is limited by $\varrho\epsilon$ (because $|\varrho S(\xi^+ - \xi^-)| < \varrho\epsilon$ for $\xi^\pm \in C^\pm$). Hence, q and ϵ are scaled by ϱ, and the upper bound for the number of required updates in the case of bounded synapses becomes $n_\circ \sim 1/(q\epsilon) = 1/(\varrho^2 q\epsilon) = 1/\epsilon^4$. This upper bound takes into consideration that the weight vector may need to travel in small steps from the boundary into the narrow neighborhood of the hypercube center where synaptic saturation can be neglected. If the postsynaptic neuron was already involved in a learning task, its weight vector is likely to be in this neighborhood, and learning may be as fast as without imposing these bounds.

3.4 Global Inhibition and a Small Threshold Are Necessary. To test the statement of the theorem and 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 ($j = 1, \ldots, N = 20$), uniformly distributed between 0 and $R = 40$ (to allude to realistic firing rates in terms of spikes per second). The excitatory synaptic weights G_j of the 20 synapses were randomly initialized between 0 and 1. The two classes C^+ and C^- were constructed by projecting the patterns onto a random separation vector S of length N. Each pattern was tagged according to whether the projection was above or below a separation threshold θ, which in turn was chosen to divide the patterns in two groups, each containing five patterns. In our example, we had $\theta = 17.3$, and the resulting separation margins were $\delta = \epsilon = 0.27$. In general, the model might not converge for a random choice of the parameters of the synaptic dynamics (e.g., when we use $\theta_\circ = \theta$ and $\delta_\circ = \delta$). However, our theorem guarantees that there is scaled version of the parameters ($\theta_\circ = \varrho\theta$ and $\delta_\circ = \varrho\delta$), which would allow the convergence of the learning process. In our specific example, any $\varrho \leq 0.3$ allows learning all the patterns without mistakes. As predicted, the separation of the postsynaptic current, $h^t > \theta_\circ + \delta_\circ$ and $h^t < \theta_\circ - \delta_\circ$, for patterns ξ in C^+ and C^-, respectively, is reached after a few updates of the synaptic strengths G_j^t according to equation 2.1, with $h^t = \frac{1}{N}\sum_{j=1}^{N}(G_j^t - g_I)\xi_j$ (cf. Figure 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 $(\varrho S - G_I)\Delta L > 0$, equation 3.2, while the forgetting (saturation) part may work against this progress as $(\varrho S - G_I)\Delta F$ can become negative (see Figure 2B).

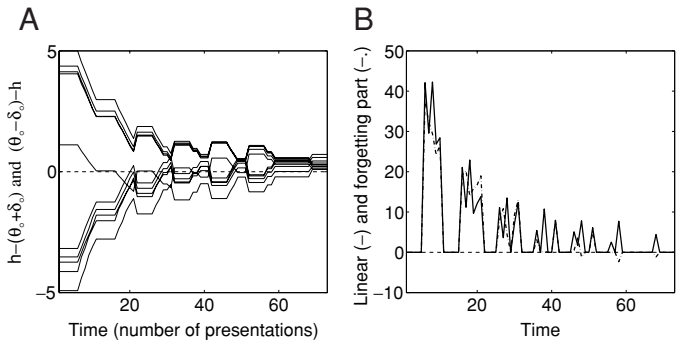

Figure 2: Any linearly separable set of patterns is learnable with limited synaptic strengths. (A) Evolution of the signed distance between the total postsynaptic current and the learning threshold, $h^t(\xi) - (\theta_\circ + \delta_\circ)$, for patterns ξ of class C^+, and $(\theta_\circ - \delta_\circ) - h^t(\xi)$, for patterns of class C^-. According to the update condition, equation 2.1, 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 led to synaptic updates). Note that the monotonic convergence of the total weight vector G_I^t toward the scaled solution vector ϱS does not imply that for all patterns, the total input $h^t(\xi)$ monotonically converges. Model parameters: $\theta_\circ = 5.2$, $\delta_\circ = 0.08$, $q = q^\pm = 2 \cdot 10^{-3}$, $g_I = 0.5$. The same set of patterns is used in Figures 3 to 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 nonsaturating and saturating part: they indicate by how much the two learning components ΔF^t and ΔG^t move the effective weight vector $G_I^t = G^t - g_I$ toward the target vector ϱS. The flat parts correspond to presentations that did not trigger synaptic updates because the patterns were already correctly implemented, and the condition on h^t in the update rule 2.1 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 48th, 58th, and 68th 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.

The value of global inhibition plays an important role, although it does not need to be finely tuned to guarantee the convergence of the learning process. As predicted by the theorem, many more learning steps are necessary if g_I is too close to the boundary 0 or 1 (see Figure 3A). In fact, the theorem predicts that the number of synaptic updates required to learn the patterns is roughly $n_\circ \propto \frac{1}{g_I} \approx \frac{1}{g_I(1-g_I)}$. The chance of finding a configuration of excitatory synapses that balance inhibition shrinks when g_I tends to a

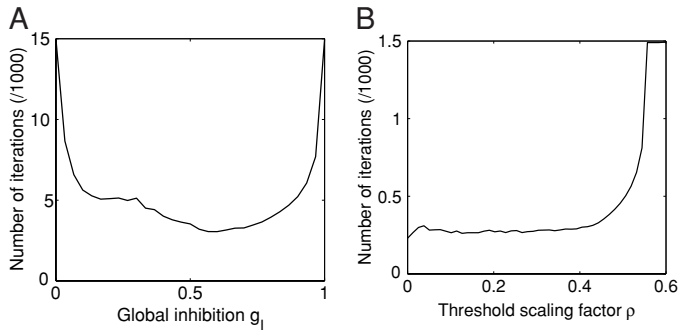

Figure 3: Learning requires global inhibition and a small scaling factor. (A) The number of iterations (in thousands) required to learn the random set of patterns is minimal if the global inhibitory strength g_I is roughly 0.5, as predicted by the theory. An inhibitory weight close to 0 or 1 urges the excitatory weights to "catch up" to the inhibitory weight, and the emerging synaptic saturation ("forgetting") strongly impairs the learning (cf. Figure 1C). The neuronal threshold, the learning margin, and the learning rate were scaled by a factor of 1/100 (yielding $\theta_\circ = 5.2 \cdot 10^{-3}$, $\delta_\circ = 8 \cdot 10^{-5}$, $q = 2 \cdot 10^{-5}$) 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. Figure 1C).

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 (see Figure 3B). The simulation result is expressed by the requirement $\varrho \leq \epsilon \overline{g}_I/(2R)$ appearing in the theorem (see also Figure 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.

Global inhibition is necessary for a simple reason. For instance, the separation of the patterns into two classes may require an output $\xi_{post} = 0$ to a pattern with a high activity level $f = \frac{1}{N}\sum_{j=1}^{N} \xi_j$ (many presynaptic neurons strongly active). This is typically not possible with excitatory synapses alone because a pattern with a high total activity would lead to a suprathreshold response. However, if the activity level f is subtracted through global inhibition, $h = \frac{1}{N}\sum G_j \xi_j - g_I f = \frac{1}{N}\sum (G_j - g_I)\xi_j$, 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). To simultaneously assign

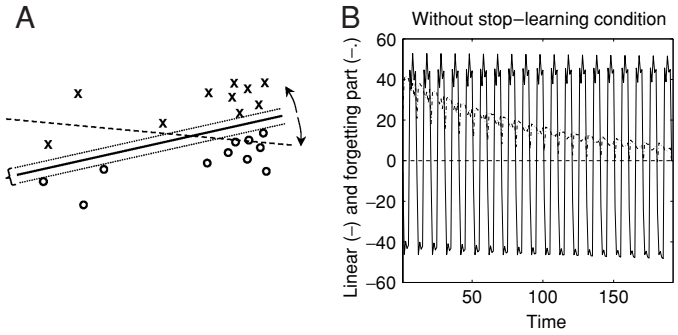

Figure 4: Individual synaptic modifications should be small and triggered only if the required response is not matched. (A) To prevent overshooting, the learning rates $q^{\pm}$ must be a fraction of the separation parameter ϵ (width of the bracelet: $2(\epsilon + \delta)$, corresponding to the separation margin between the two classes, as indicated by the parallel dotted lines). Without the stop-learning condition and without shrinking of the learning rates $q^{\pm}$ toward 0, 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 leftmost cross). 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 Figure 2B, but without stop-learning condition on the total postsynaptic current h^t in equation 2.1. The linear part oscillates because the weight vector G periodically "overlearns" the patterns, that is, is repeatedly attracted toward 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.

an output $\xi_{post} = 1$ to a pattern with low activity level, the threshold must be small. This is needed because tightly separated classes (ϵ small) require that small differences in the inputs ξ_j, independent of the size of ξ_j, may turn a subthreshold response into a suprathreshold response. After subtracting the activity level f, this becomes possible with a small threshold.

3.5 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 toward 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. Figure 4A). The requirement

of a small threshold is also confirmed by the simulations (see Senn & Fusi, 2004, in press).

Learning is also severely impaired if the stop-learning condition on the total postsynaptic current h^t in equation 2.1 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 (see Figure 4A). In fact, dropping the stop-learning condition leads to sustained oscillations in the total postsynaptic currents, and no further learning progress is achieved (see Figure 4B). Although decreasing the learning rate will reduce the amplitude of the oscillations to 0, the final position of the separation plane may still not separate the two classes of input patterns. This is because without the stopping condition, it is just the center of gravity of the patterns within each class $C^{\pm}$ that determines the final position of the separation plane, and this does not account for the outliers. Any learning rule that is able to learn tightly separated classes must incorporate some form of stopping condition.

3.6 Learning Equalizes Synaptic Strengths and Balances Inputs. In the absence of the stopping condition, the statistics of different synapses tend to reflect the statistics of LTP and LTD events. If different synapses share the same statistics, then the distribution of efficacies will also tend to be equalized (i.e., be the same across different inputs). For example, in the case of slow learning of random uncorrelated patterns with binary synapses, the asymptotic potentiation probability (G^*) for all the synapses is given by the ratio between the rate of potentiating events ($\tilde{q}^+$) divided by the total rate of events inducing potentiations or depressions ($\tilde{q}^+ + \tilde{q}^-$) (see Brunel et al., 1998, and below). This is also confirmed for the case of analog and bounded synapses. By the ongoing up- and downregulation of a synapse in the presence of the multiplicative saturation, the synaptic weights are driven toward asymptotic states where the synaptic saturation and the Hebbian learning are balanced. In the current example, this is expressed by the convergence $(\varrho S - G_I)\Delta F \to 0$ (decaying curve in Figure 4B).

In the presence of the stopping condition, the final excitatory weights reached after successful learning will always be close to the global inhibitory weight, at least when the difficulty of the task (small ϵ) requires a small threshold $\theta_\circ$ and a small learning margin $\delta_\circ$. The tighter the two classes C^+ and C^- are separated, the less distortion by synaptic saturation can be afforded, and the more uniform the synaptic 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

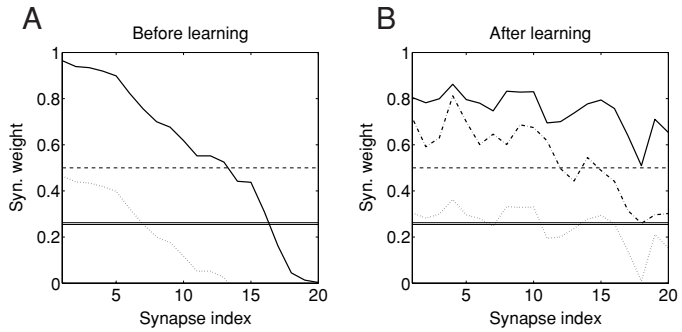

Figure 5: Balancing and equalization of the synaptic weights through learning. (A) The initial synaptic strengths G_j (solid line) span the whole possible interval between 0 and 1, scaled up by N. The two narrowly separated black lines represent the learning thresholds $\theta_\circ \pm \delta_\circ$, 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 Figure 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).

threshold (small scaling factor ϱ), depending on the separation margin of the classes to be learned (see Figures 1B and 1C). Hence, the balancing of excitation and inhibition, and the equalization of the synaptic weights, appears as a by-product of learning. This is also confirmed by our simulations. Due to their random initial values, the weights span the whole possible range of values before learning (see Figure 5A). After a few synaptic updates evoked by the initially incorrectly classified patterns, the weights all adopted roughly the same value (see Figure 5b, solid line). If the stop-learning condition is discarded, the weights are less equalized because for the small set of random patterns ($p = 10$), the asymptotic strengths varies considerably (see Figure 5b, dashed-dotted line). Weight equalization (but not necessary the balancing by inhibition) would also emerge without the stopping condition, but the number of patterns and the number of synaptic updates must be large.

3.7 Conflicting Patterns Shut Down Neuronal Activity. An interesting property of (multiplicative) synaptic saturation is that it tends to stabilize

the synaptic weights (van Rossum, Bi, & Turrigiano, 2000; Rubin, Lee, & Sompolinsky, 2001). This property can be advantageous when dealing with similar patterns requiring different outputs, since in these cases, it leads to a uniform distribution of the synaptic weights, which depends on the ratio between the probability of inducing LTP and the probability of inducing LTD. If this distribution is below the threshold of activation of the output cell, the neuron will no longer respond to these patterns, and therefore will not try to make an impossible classification of stimuli that would produce contradictory responses.

To be more concrete, we stimulate our neuron with a set of input patterns such that each synapse gets repeatedly potentiated and depressed. According to the update rule, equation 2.1, the equilibrium weight of synapse j is then determined by the equation

$$\Delta G_j = \tilde{q}^+(1 - G_j) - \tilde{q}^- G_j = 0, \tag{3.3}$$

where $\tilde{q}^\pm$ represent the effective rates of up- and downregulations. 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 equation 3.3 for G_j gives the unique equilibrium weight $G^* = \tilde{q}^+/(\tilde{q}^+ + \tilde{q}^-)$. For slow learning (small $q^\pm$) this expression does not depend on the specific order of presentation of patterns (Brunel et al., 1998). The equilibrium weight G^* is an attracting fixed point of equation 3.3, 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 equation 3.3 always drive the synapse to the unique steady state G^*. If the equilibrium weight is smaller than 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 depressed only 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 general, any attempt to train the output neuron(s) to respond with different outputs to too similar input patterns will eventually lead to a subthreshold activation. The case of a single pattern for which contradicting outputs are required is formally treated in the appendix (see theorem 2).

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 (see Figure 6A). The downward

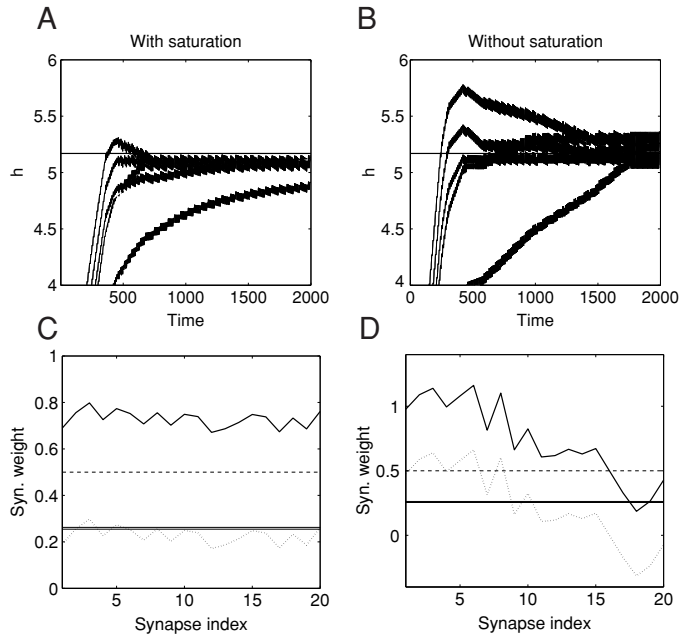

Figure 6: Synaptic saturation suppresses neuronal activity in response to conflicting patterns. (A) Evolution of the total postsynaptic current h^t in response to the five patterns trained with conflicting outputs, that is, requiring once the output $\xi_{post} = 0$ and once $\xi_{post} = 1$ for the same input patterns. After a transient response (around update 200), the total postsynaptic currents of the five patterns becomes subthreshold (horizontal line represents the neuronal threshold $\theta_\circ$). (B) Without synaptic saturation (modeled by setting $\Delta F = 0$ in equation 3.1), 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 Figure 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 the suppression of any neuronal activity (C). The final weight distribution when the upper synaptic bound was relieved does not show the equalization, and therefore does not lead to the activity suppression (D).

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 (see Figure 6C). In fact, without synaptic saturation (mimicked by cancelling the forgetting part $\Delta F = -\xi * G_I$

in the update rule, equation 3.1), the suppression effect vanishes and the total postsynaptic currents incoherently become either sub- or suprathreshold (see Figure 6B). This is also reflected in the uncontrolled growth of the synaptic weights beyond the upper boundary (see Figure 6D). Hence, teaching the neuron to respond with different outputs to the same patterns will uniformly depress the synaptic weights and silence the neuron.

3.8 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 and (2) the small learning rate is replaced by small transition probabilities between stable discrete states. If these conditions are satisfied, the expected values of the binary synapses are well described by the analog synaptic variables introduced in the model. As a consequence, the convergence of the stochastic learning process with binary synapses is well predicted by the deterministic one for analog synapses. More precisely, one proves that the stochastic algorithm is likely to converge within some finite number of updates, $n_\circ(\epsilon)$, which is bounded above by some power of $1/\epsilon$. The probability of not converging within these updates shrinks as $1/N$ when the number of neurons N increases while the separation margin ϵ is kept fixed (for a rigorous proof and simulations with highly correlated patterns, see Senn & Fusi, 2005).

Simulations with binary synapses projecting to a single output cell confirm that the stochastic learning rule is successful (see Figure 7). The parameters of the learning dynamics are the same as in the simulations of the deterministic example (see Figure 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 noisier, and it takes a larger number of presentations than in the case with continuous synapses (see Figure 7A). With an increasing number of neurons, however, the prediction of the synaptic dynamics by the mean field equation, 2.1, becomes more reliable. The redundancy in the synaptic encoding speeds up learning until it approaches the convergence speed of continuous-valued (bounded) synapses. In fact, the number of presentations per pattern, required to correctly classify the stimuli, shrinks with the increasing number of presynaptic neurons toward an asymptotic value (see Figure 7B).

4 Discussion

We showed that despite the synaptic boundedness and despite restricting plasticity to the excitatory synapses, any set of linearly separable patterns

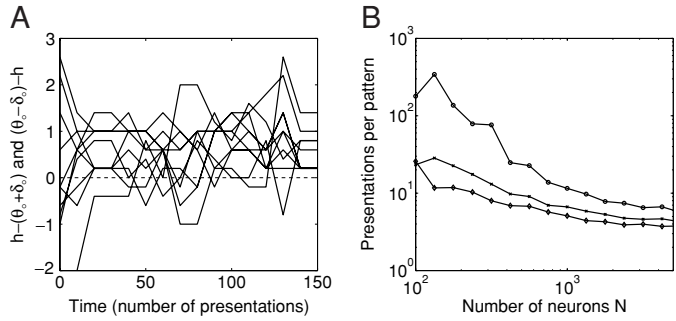

Figure 7: Convergence of stochastic learning in the case of binary synapses. (A) Total synaptic input current h^t as a function of time, evaluated for all 10 random, linearly separable patterns. Same parameters as in Figure 2, except for the number of neurons, which is $N = 100$ instead of $N = 20$. 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.

can be learned with a Hebbian rule incorporating a stop-learning condition. These biologically plausible restrictions, however, require (1) some global inhibition, (2) a small learning rate, and (3) a threshold that is small compared to 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 by-product 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.

4.1 Possible Implementations of the Stop-Learning Mechanism. The stop-learning condition is necessary to protect past memories when the same or similar patterns are insistently presented. There are many ways of implementing such a stopping mechanism. It could be inherent to the individual synapse, governed by the postsynaptic activity, or it can 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 above some critical level or below some minimal concentration (see Fusi, 2003, for a spike-driven synaptic dynamics implementing this mechanism). High calcium concentration might indicate that a neuron, which is supposed to be active, is already responding as imposed by the sensory stimulus or by the teacher, and that learning should stop. A similar mechanism can be obtained by reducing the synaptic change when the postsynaptic neuron is spending a large fraction of its time in the refractory period (Amit & Mongillo, 2003). Unfortunately, experimental data leave the question of such an intrinsic nonmonotonicity open (see, e.g., Cho, Aggleton, Brown, & Bashir, 2001). Another possibility would be that the stop-learning signal is carried by an external signal, for instance, related to the reduction of dopamine release, as observed after successful reinforcement learning (see, e.g., Fiorillo, Tobler, & Schultz, 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 therefore easy to classify (Rainer, Lee, & Logothetis, 2004).

4.2 Global Inhibition Sets the Equilibrium Distribution of the 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 the neocortex, but also in the hippocampus, may form a large network, tightly coupled through gap junctions (see, e.g., Amitai et al., 2002). In our framework, such a global inhibition defines a range, far from saturation, into which the excitatory weights will tend during the learning process. Since we restrict synaptic plasticity to excitatory synapses, 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. Nonglobal inhibition may lead to a strong and unequal forgetting across the synapses due to unequal synaptic saturation, unless inhibition is also plastic.

The supervised learning scenario with the stop-learning condition urges the total postsynaptic currents to be clustered around the neuronal threshold. Since the latter must be small relative to the synaptic bounds, the excitatory current will be balanced by inhibition after learning. Since inhibition is global, the excitatory synaptic strengths, moreover, become equalized. Weight equalization and balancing by inhibition will always emerge from the stopping condition, even when the synapses are not bounded, as long as the threshold is small and the inhibitory weights are equal. If the stopping condition is discarded and the synapses are bounded, weight equalization arises (for large N and p) because the weights tend to an asymptotic state,

which, for uniform random patterns, depends on only the ratio between the effective rate for up- and downregulation. This asymptotic state, however, is not necessarily related to the global inhibition. In fact, assuming that it is dominated by the inhibitory weight has distinct computational advantages in terms of neuronal silencing when learning does not converge (see below).

Weight equalization was also shown to emerge in an unsupervised learning scenario where the postsynaptic activity is not imposed by a teacher signal, and hence no stopping condition can be defined explicitly (Rumsey & Abbott, 2003). Instead of stopping any synaptic modification when the desired output activity is reached, a slow anti-Hebbian term is shown to smooth out large fluctuations in the synaptic weight structure caused by Hebbian learning. Bounded synapses may also contribute to some synaptic equalization in an unsupervised learning scenario, but because of the missing activity-dependent feedback, synaptic bounds are themselves not sufficient.

4.3 Slow Learning Prevents Fast Forgetting. Slow learning becomes important if the set of patterns to be learned is large. This is because 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 be able to operate with a limited number of discrete states. In a discrete-valued synaptic model, slow learning is achieved by making a selection of a small number of synapses to be modified. Stochastic selection is an unbiased way to choose the synapses to be changed (Tsodyks, 1990; Amit & Fusi, 1992, 1994) and it can be naturally implemented by exploiting the variability in the neural activity (Fusi, Annunziato, Badoni, Salamon, & Amit, 2000; Fusi, 2002). There is an optimal learning rate that allows learning uncorrelated random patterns in a single shot and forgetting slowly. Below this learning rate, every pattern should be presented more than once (Fusi, 1995; Brunel et al., 1998). The advantage is that the synaptic resources are equally distributed among all the patterns to be stored. Our binary perceptron exploits slow learning for the same reason, and the stopping condition introduces an extra selection mechanism. Interestingly, slow learning is observed in some cortical areas: for instance, in inferotemporal and perirhinal cortex (Miyashita, 1993; Yakovlev, Fusi, Berman, & Zohary, 1998; Erikson & Desimone, 1999) where the internal representations of sensory stimuli form in tens or hundreds of repetitions of the same pattern. The introduction of other internal synaptic states would allow the same memory span and a much reduced number of presentations (Fusi, Drew, & Abbott, 2005).

4.4 Small Neuronal Thresholds Allow Separating 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 (Braitenberg & Schütz, 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, that is, with decreasing separation margin between the two classes to be learned (ϵ). Interestingly, this is needed also for the classical perceptron and for many other classical learning rules like the Hopfield prescription. However, the requirement is veiled by the unboundedness of the synapses and by the fact that usually the neuronal threshold is set to 0. Indeed, as the number of patterns (p) increases, the separation margin ϵ typically becomes smaller, and more iterations are needed to converge. As a consequence, for the classical perceptron learning (i.e., with unbounded synapses), the maximum synaptic weight $G_{\max} = \max_{i=1,\ldots,N} G_i$ increases when more patterns (p) have to be learned or, more generally, when the separation margin ϵ goes to zero (see Figure 8). To enforce that the maximum synaptic weights remains finite, say 1, all the synaptic weights, the threshold (θ), and the learning margin (δ) should be scaled by the same factor, namely, the maximum weight $G_{\max}$. For random uncorrelated patterns, this maximum weight $G_{\max}$ roughly increases as the square root of the number of random patterns, $\sqrt{p}$ (see Figure 8, vertical and bottom horizontal axes), but also as the inverse of the separation margin, $1/\epsilon$ (see Figure 8, vertical and top horizontal axes). The growth of $G_{\max}$ with $\sqrt{p}$ and $1/\epsilon$ implicitly confirms the theoretical result that in the limit of large N, the separation margin ϵ shrinks as $1/\sqrt{p}$ (Köhler, Diederich, Kinzel, & Opper, 1990, equation 7). Notice that in models without the stopping condition, for example, in the Hopfield model where the synaptic weights are explicitly set as the sum over p patterns (Hertz et al., 1991), the maximum weight grows even linearly with p. Clipping the synapses to a finite range at the end of the learning process as in Sompolinsky (1987) and Amit and Mascaro (2001) does not help, because this would require a buffer for temporarily storing the unbounded synaptic weights during the learning process, which itself would require growing synapses. Our need of a proper scaling of the threshold is not a unique feature of our model, but will be shared by any model with bounded synaptic weights. We conclude that depending on the difficulty of the learning task, learning with bounded synapses requires some fine discrimination around the balance between excitation and inhibition. The correct tuning of the threshold-to-synaptic strength ratio could be performed by additional homeostatic processes (see, e.g., Desai, Cudmore, Nelson, & Turrigiano, 2002). Homeostatic plasticity may also tune the global

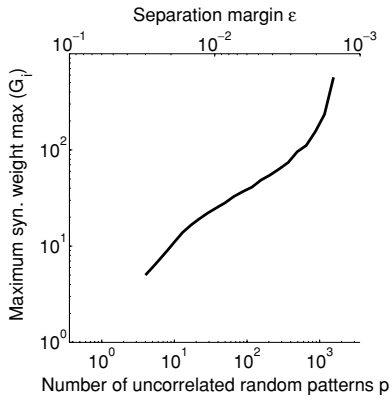

Figure 8: Maximum synaptic weight $G_{\max}$ as a function of the number of random uncorrelated binary patterns p (bottom scale) and their separation margin ϵ (top scale) for the classical perceptron with unbounded synapses. All the scales are logarithmic. According to the graph, the maximum weight after learning, $G_{\max} = \max_{i=1,\ldots,N}\{G_i\}$, increases with the square root of the number of patterns, $G_{\max} \sim \sqrt{p}$, and with the inverse of the separation margin, $G_{\max} \sim 1/\epsilon$. The number of neurons (fixed to $N = 2000$) was chosen such that for all the training sets, the number of patterns was well below the maximal capacity, $p < 2N$. The maximum weight increases because the number of synaptic updates required to find an appropriate weight vector also grows with increasing complexity of the separation task (growing p and decreasing ϵ). Recall that with each of these synaptic updates, a component of the solution vector is added to the weight vector, causing the latter to steadily grow in the direction of the solution vector. If the synaptic weights are bounded, then the neuronal threshold should be scaled with a quantity growing as $G_{\max}$, and hence growing with the difficulty of the classification task. This shows that the threshold scaling in our perceptron model with bounded synapses is the counterpart of the unlimited weight growth in the classical perceptron, and therefore cannot be avoided. The initial values of the synaptic weights were randomly chosen between -1 and 1.

inhibition (g_I) to dominate over the excitatory equilibrium weight (G^*), such that neurons silence themselves in response to unstructured input.

4.5 Silencing Uncertain Neurons Allows Dealing with Nonseparable Patterns. Suppressing the activity of a neuron trained to respond with different outputs to the same input patterns is an important property when dealing with nonseparable patterns. This problem emerges when the maximal storage capacity is surpassed or the input patterns are inherently nonseparable. For example, as the number of random input patterns increases

($p > 2N$), the chance that they are nonseparable, and therefore not classifiable by a neuron, also increases (Cover, 1965). Nonseparable patterns would uniformize the synaptic weights by means of the synaptic saturation, and their response would be suppressed by the global inhibition. The same suppression mechanism can also be exploited to improve the classification of more complex data sets like Latex deformed characters (Senn & Fusi, in press). Because patterns that are incorrectly classified typically evoke a subthreshold response, the classification performance can be improved by considering the response of several output neurons in parallel. If these neurons behave in a different way (e.g., because of the stochastic selection in the learning rule for binary synapses), then some output units would respond correctly, while those that respond incorrectly are actually silent. A similar mechanism has already been applied in Amit and Mascaro (2001), where the authors consider several output units, each randomly connected to a subset of input units. They also correctly classify a large number of Latex deformed characters. In general, an additional second layer would be required to judge whether the number of activated neurons is significant for a correct classification of the input pattern.

Appendix

A.1 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, there is some global inhibition, and 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 1. *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 $G^0 \in [0, 1]^N$, the synaptic dynamics 2.1 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 $n_\circ$ counts only the number

of presentations that trigger synaptic updates, that is, those for which the update conditions in equation 2.1 are satisfied. Since the patterns satisfying these conditions are not known a priori, however, an online algorithm needs to cycle repeatedly 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. 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 2.1 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 equation 2.1 into the linear and forgetting part according to equation 3.1. Recall that the condition for a synaptic update is satisfied if either $h = \frac{1}{N} G_I \xi \le \varrho(\theta + \delta)$ or $h = \frac{1}{N} G_I \xi \ge \varrho(\theta - \delta)$ 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 $(\varrho S - G_I)\xi \ge \varrho\epsilon N$. 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 $-(\varrho S - G_I)\xi \ge \varrho\epsilon N$. Defining the linear part in the learning rule 3.1 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, equation 3.2, presented previously in the main text,

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

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 equation 3.2) 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 the Cauchy-Schwartz inequality twice in the form $x\,y \le \|x\|\,\|y\|$, with equality if $x = y$, we get sequentially

$$\begin{aligned}(\varrho S - G_I)\Delta F &= G_I(\xi * G_I) - \varrho S(\xi * G_I) = (\sqrt{\xi} * G_I)^2 \\ &\quad - \varrho(\sqrt{\xi} * S)(\sqrt{\xi} * G_I) \\ &\ge \|\sqrt{\xi} * G_I\|(\|\sqrt{\xi} * G_I\| - \varrho\,\|\sqrt{\xi} * S\|). \end{aligned} \tag{A.1}$$

When forgetting supports learning. In the case of $\|\sqrt{\xi} * G_I\| \geq \varrho \, \|\sqrt{\xi} * S\|$, the parentheses on the right-hand side of equation A.1 is nonnegative, and one immediately concludes from it 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 toward ϱS. In fact, since $\Delta G = \Delta L + \Delta F$, we obtain from $(\varrho S - G_I)\Delta F \geq 0$, together with equation 3.2, that for any ϱ,

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

When forgetting counteracts learning. We next consider the case that $\|\sqrt{\xi} * G_I\| \leq \varrho \, \|\sqrt{\xi} * S\|$. Inserting this into equation A.1 while neglecting the term $\|\sqrt{\xi} * G_I\|$ in the parentheses on the right-hand side, we get the estimate

$$(\varrho S - G_I)\Delta F \geq -\|\sqrt{\xi} * G_I\| \, \varrho \, \|\sqrt{\xi} * S\| \geq -\varrho^2 \, \|\sqrt{\xi} * S\|^2 \geq -\varrho^2 RN. \tag{A.3}$$

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 A.3, together with equation 3.2, that

$$\begin{aligned} (\varrho S - G_I)\Delta G &\geq \varrho N(\epsilon \overline{g}_I - \varrho R) \geq \varrho \epsilon \overline{g}_I N/2, \\ &\text{provided } \|\sqrt{\xi} * G_I\| \leq \varrho \, \|\sqrt{\xi} * S\|. \end{aligned} \tag{A.4}$$

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 equations A.2 and A.4, 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

$$\begin{aligned}\|\varrho S - G_I^{t_\mu+1}\|^2 - \|\varrho S - G_I^{t_\mu}\|^2 &= -2q\left(\varrho S - G_I^{t_\mu}\right)\Delta G^{t_\mu} + q^2\|\Delta G^{t_\mu}\|^2 \\ &\leq \ldots\ldots \leq qN(qR^2 - \varrho\epsilon\overline{g}_I) \\ &\leq -q\varrho\epsilon\overline{g}_I N/2. \end{aligned} \tag{A.5}$$

Note that by definition of ΔG (see equation 3.1), we have $\|\Delta G^{t_\mu}\|^2 \leq R^2N$. 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 A.5, we get an estimate of the telescope sum,

$$\begin{aligned}\|\varrho S - G_I^t\|^2 - \|\varrho S - G_I^0\|^2 &= \|\varrho S - G_I^t\|^2 - \|\varrho S - G_I^{t-1}\|^2 \\ &\quad + \|\varrho S - G_I^{t-1}\|^2 - \|\varrho S - G_I^{t-2}\|^2 \\ &\quad + - \ldots \leq -n_t q\varrho\epsilon\overline{g}_I N/2, \end{aligned} \tag{A.6}$$

where n_t is the number of synaptic updates up to the ith presentation of a pattern. From equation A.6, we immediately obtain

$$0 \leq \|\varrho S - G_I^t\|^2 \leq \|\varrho S - G_I^0\|^2 - n_t q\varrho\epsilon\overline{g}_I N/2. \tag{A.7}$$

Since $\|\varrho S - G_I^0\|^2 \leq (\varrho^2 + g_I^2 + 1)N \leq 3N$, we conclude from equation A.7 that $\|\varrho S - G_I^t\|^2 \leq 0$ 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 $G^0 \in [0,1]^N$.

A.2 Neuronal Silencing with Conflicting Patterns. To illustrate how the perceptron with bounded synapses deals with strongly nonseparable patterns, we consider a learning scenario with a single pattern $\xi \in C^\pm$, for which both outputs 0 and 1 are required. In the course of "learning", the response to this pattern ξ will eventually be suppressed, provided that the global inhibition is strong enough. This shunting property is due to the multiplicative synaptic saturation, and it is not present in the classical perceptron with unbounded synapses (recall that a mean field description of binary synapses naturally leads to a multiplicative saturation). For simplicity, we assume that the input patterns are also binary.

Theorem 2. *Let us repeatedly present a pattern $\xi \in \{0,1\}^N$ to the perceptron, and alternately require the output $\xi_{post} = 0$ and 1. Define the asymptotic synaptic strength by $G^\infty = \frac{q^+}{q^- + q^+}$, where q^+ and q^- are the learning rates for the up- and downregulation, respectively. Assume that the asymptotic strength is dominated by the global inhibition g_I, that is, $(G^\infty - g_I)\overline{\xi} \le \theta_\circ - \kappa$, where $\theta_\circ$ is the neuronal threshold, $\kappa > 0$, and $\overline{\xi} = \frac{1}{N}\sum_{i=1}^N \xi_i$ is the mean activity of ξ. Assume that $q^- \le \kappa$. Then for any initial condition of the synaptic weight vector $G^0 \in [0,1]^N$, the response to pattern ξ becomes subthreshold, $h^t = \frac{1}{N}\sum_{i=1}^N (G_i^t - g_I)\xi_i < \theta_\circ$, after $t \approx \frac{-2\log\kappa}{q^-}$ presentations.*

The theorem can easily be generalized with the same proof to the case of $C^+ = C^-$, where $C^\pm$ contains more than only one pattern ξ. Interestingly, the estimate of the convergence time holds independent of the size of the LTP rate q^+. The basic idea of the proof is to show that a sequence up- and downregulations of a synaptic weight G_j^t (with $\xi_j > 0$) will always bring this toward its asymptotic state G^∞. The same convergence property is also shown to hold for the case of stimulations with random patterns in the absence of the stopping condition in the learning rule (Brunel et al., 1998).

Proof. We show that whenever the total current at time $t-1$ is suprathreshold, $h^{t-1} \ge \theta_\circ$, it will decrease within the subsequent two updates by at least $q^-\kappa$. Let us first assume that at time $t-1$, the total postsynaptic current is above the LTP threshold, say, $h^{t-1} = \theta_\circ + h_+^{t-1} > \theta_\circ + \delta_\circ$, with some $h_+^{t-1} > \delta_\circ$. When considering pattern ξ as a member of class C^- at time $t-1$, no potentiation is triggered due to the stop-learning condition, and we have $h^t = h^{t-1}$. When considering pattern ξ in the next time step t as belonging to class C^+, a downregulation is triggered. Since by assumption we have $(G^\infty - g_I)\overline{\xi} \le \theta_\circ - \kappa$, we can subtract this inequality from $h^t = \frac{1}{N}\sum_{i=1}^N (G_i^t - g_I)\xi_i = \theta_\circ + h_+^{t-1}$ and obtain

$$\frac{1}{N}\sum \left(G_i^t - G^\infty\right)\xi_i \ge \kappa + h_+^{t-1}. \tag{A.8}$$

According to equation 2.1, the expected change of synapse j at an LTD step is $\Delta G_i^t = -q^- G_i^t \xi_i$. With this and equation A.8, total change in the postsynaptic current at time t can be estimated by

$$h^{t+1} - h^t = \frac{1}{N}\sum \Delta G_i^t \xi_i \le -\frac{q^-}{N}\sum \left(G_i^t \xi_i - G^\infty\right)\xi_i \le -q^-\left(\kappa + h_+^{t-1}\right). \tag{A.9}$$

The first inequality holds because $G^\infty \geq 0$, and the second inequality holds because of equation A.8 and because for $\xi_j \in \{0, 1\}$ we have $\xi_i^2 = \xi_i$. Since $h^t = h^{t-1}$, we conclude that $h^{t+1} - h^{t-1} \leq -q^-(\kappa + h_+^{t-1})$.

We next show that the same inequality still holds when, for instance, a downregulation is immediately followed by an upregulation, and when the total current is not yet subthreshold. Such a situation can arise when $\theta_\circ \leq h^{t-1} \leq \theta_\circ + \delta_\circ$. Let us set again $h^{t-1} = \theta_\circ + h_+^{t-1}$. According to the learning rule 2.1, the synaptic weight after a downregulation at time $t-1$ is decreased by $q^- G_i^{t-1}\xi_i$, and it becomes $G_i^t = G_i^{t-1}(1 - q^-\xi_i)$. After a subsequent upregulation at time t, this weight is increased by $q^+(1 - G_i^t)\xi_i = q^+(1 - G_i^{t-1}(1 - q^-\xi_i))\xi_i$. By summing up the contributions of the individual components and using that $\xi_i^2 = \xi_i$, we can compute the difference in the total current after first a down- and then an upregulation as

$$
\begin{aligned}
h^{t+1} - h^{t-1} &= -\frac{1}{N}\sum q^- G_i^{t-1}\xi_i + \frac{1}{N}\sum q^+\left(1 - G_i^{t-1}(1 - q^-\xi_i)\right)\xi_i \\
&= -(q^+ + q^-)\frac{1}{N}\sum G_i^{t-1}\xi_i + q^+\overline{\xi}_i + q^+q^-\frac{1}{N}\sum G_i^{t-1}\xi_i \\
&\leq -(q^+ + q^-)\left(G^\infty\overline{\xi} + \kappa + h_+^{t-1}\right) + q^+\overline{\xi} + q^+q^-\frac{1}{N}\sum G_i^{t-1}\xi_i \qquad \text{(A.10)} \\
&\leq -(q^+ + q^-)\left(\kappa + h_+^{t-1}\right) + q^+q^- \qquad \text{(A.11)} \\
&\leq -q^-\left(\kappa + h_+^{t-1}\right). \qquad \text{(A.12)}
\end{aligned}
$$

To get the first term in equation A.10, we were plugging in the inequality $\frac{1}{N}\sum G_i^{t-1}\xi_i \geq G^\infty\overline{\xi} + \kappa + h_+^{t-1}$. This latter inequality is derived in the same way as equation A.8. To get the first term in equation A.11, we were substituting $G^\infty = q^+/(q^+ + q^-)$, and could cancel the term $q^+\overline{\xi}_i$. To get the second term in equation A.11, we used that both G_i^{t-1} and ξ_i are between 0 and 1. Inequality A.12 holds if $q^- \leq \kappa$. The case where first an up- and then a downregulation occurs leads to the same inequality, A.12.

Taken together, equations A.9 and A.12 state that whenever the total synaptic current at time $t-1$ is above threshold, $h^{t-1} = \theta_\circ + h_+^{t-1} \geq \theta_\circ$, it decreases within the next two presentations at least by the amount $q^-(\kappa + h_+^{t-1})$. Since the reduction in h^{t-1} is due to the reduction of h_+^{t-1}, also $\kappa + h_+^{t-1}$ is reduced within the two next time steps by the same amount, $\kappa + h_+^{t+1} \leq$

$(\kappa + h_+^{t-1})(1-q^-)$. We conclude that the sequence $\kappa + h_+^{2t'}$ $(t' = 0, 1, 2, \ldots)$, as long as $h_+^{2t'} \geq 0$, is bounded above by a geometric series,

$$\kappa + h_+^{2t'} \leq a_{t'} = \left(\kappa + h_+^0\right)(1-q^-)^{t'}.$$

This geometric series $\{a_{t'}\}_{t'=0,1,\ldots}$ decays below κ (and therefore $h^{2t'} = \theta_\circ + h_+^{2t'}$ decays below $\theta_\circ$) after $t'_\circ$ steps, where $t'_\circ = \log(\frac{\kappa}{\kappa + h_+^0})/\log(1-q^-) \leq \log\kappa/\log(1-q^-) \approx -\log\kappa/q^-$. The last inequality holds because $h^0 \leq 1$ and therefore $\kappa + h_+^0 \leq 1$, and the approximation holds for small q^-. Hence, for any initial conditions with $h^0 \geq \theta_\circ$, the total current $h^{2t'}$ becomes subthreshold after $t = 2t' \geq 2t'_\circ \approx -2\log\kappa/q^-$ presentations.

Acknowledgments

This work was supported by the SNF grant 3152-065234.01, the Silva Casa foundation, and the EU grant IST-2001-38099 ALAVLSI. We thank Nicolas Brunel for helpful comments and for checking the theorem, Massimo Mascaro for many discussions about the paper (Amit & Mascaro, 2001) that inspired this work, and Joe Brader for interesting discussions and proofreading.

References

Amit, D. J., & Brunel, N. (1997). A model of spontaneous activity and local delay activity during delay periods in the cerebral cortex. *Cerebral Cortex*, *7*, 237–252.

Amit, D. J., & Fusi, S. (1992). Constraints on learning in dynamic synapses. *Network: Computation in Neural Systems*, *3*, 443–464.

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

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

Amit, D. J., Wong, K., & Campbell, C. (1989). Perceptron learning with sign-constrained weights. *J. Phys.*, *A22*, 2039–2045.

Amit, Y., & Mascaro, M. (2001). Attractor networks for shape recognition. *Neural Computation*, *3*, 1415–1442.

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*. Berlin: Springer Verlag.

Block, H. (1962). The perceptron: A model for brain functioning. *Reviews of Modern Physics*, *34*, 123–135.

Braitenberg, V., & Schütz, A. (1991). *Anatomy of the cortex*. Berlin: Springer Verlag.

Brunel, N., Carusi, F., & Fusi, S. (1998). Slow stochastic learning in attractor neural networks. *Network: Computation in Neural Systems*, *9*, 123–152.

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. (1995). Prototype extraction in material attractor neural networks with stochastic dynamic learning. In S. K. Rogers & D. W. Ruck (Eds.), *Proceedings of SPIE 95, Applications and Science of Artificial Neural Networks* (Vol. 2, pp. 1027–1038).

Fusi, S. (2002). Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. *Biol. Cybernetics*, *87*, 459–470.

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

Fusi, S., Annunziato, M., Badoni, D., Salamon, A., & Amit, D. J. (2000). Spike-driven synaptic plasticity: Theory, simulation, VLSI implementation. *Neural Computation*, *12*, 2227–2258.

Fusi, S., Drew, P., & Abbott, L. (2005). Cascade models of synaptically stored memories. *Neuron*, *45*, 599–611.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison-Wesley.

Köhler, H., Diederich, S., Kinzel, W., & Opper, M. (1990). Learning algorithm for a neural network with binary synapses. *Z. Phys. B—Condensed Matter*, *78*, 333–342.

Minsky, M. L., & Papert, S. A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

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

Parisi, G. (1986). A memory which forgets. *Journal of Physics A—Mathematical & General*, *19*(10), L617–620.

Rainer, G., Lee, H., & Logothetis, N. (2004). The effect of learning on the function of monkey extrastriate visual cortex. *PLoS Biol*, 2(2), E44.

Rosenblatt, F. (1962). *Principles of neurodynamics*. New York: Spartan Books.

Rubin, J., Lee, D., & Sompolinsky, H. (2001). Equilibrium properties of temporally asymmetric Hebbian plasticity. *Phys. Rev. Lett.*, *86*(2), 364–367.

Rumsey, C., & Abbott, L. (2003). Equalization of synaptic efficacy by activity- and timing-dependent synaptic plasticity. *Journal of Neurophysiology*, *91*(5), 2273–2280.

Senn, W., & Fusi, S. (2004). Slow stochastic learning with global inhibition: A biological solution to the binary perceptron problem. *Neurocomputing*, *58–60*, 321–326.

Senn, W., & Fusi, S. (2005). Convergence of stochastic learning in perceptrons with binary synapses. *Phys. Rev. E*, *7*(5).

Sompolinsky, H. (1987). The theory of neural networks: The Hebb rule and beyond. In L. Van Hemmen & J. Morgestern (Eds.), *Heidelberg Colloquium on Glassy Dynamics*. Berlin: Springer.

Tsodyks, M. (1990). Associative memory in neural networks with binary synapses. *Modern Physics Letters*, *B*4, 713.

van Rossum, M., Bi, G., & Turrigiano, G. (2000). Stable Hebbian learning from spike timing-dependent plasticity. *Journal of Neuroscience*, *20*, 8812–8821.

van Vreeswijk, C., & Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, *274*, 1724–1726.

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.

Received January 7, 2004; accepted March 22, 2005.