Convergence of stochastic learning in perceptrons with binary synapses

Walter Senn and Stefano Fusi*
Department of Physiology, University of Bern, CH-3012 Bern, Switzerland

(Received 18 September 2004; revised manuscript received 27 December 2004; published 16 June 2005)

The efficacy of a biological synapse is naturally bounded, and at some resolution, and is discrete at the latest level of single vesicles. The finite number of synaptic states dramatically reduce the storage capacity of a network when online learning is considered (i.e., the synapses are immediately modified by each pattern): the trace of old memories decays exponentially with the number of new memories (palimpsest property). Moreover, finding the discrete synaptic strengths which enable the classification of linearly separable patterns is a combinatorially hard problem known to be NP complete. In this paper we show that learning with discrete (binary) synapses is nevertheless possible with high probability if a randomly selected fraction of synapses is modified following each stimulus presentation (slow stochastic learning). As an additional constraint, the synapses are only changed if the output neuron does not give the desired response, as in the case of classical perceptron learning. We prove that for linearly separable classes of patterns the stochastic learning algorithm converges with arbitrary high probability in a finite number of presentations, provided that the number of neurons encoding the patterns is large enough. The stochastic learning algorithm is successfully applied to a standard classification problem of nonlinearly separable patterns by using multiple, stochastically independent output units, with an achieved performance which is comparable to the maximal ones reached for the task.

DOI: 10.1103/PhysRevE.71.061907 PACS number(s): 87.19.La, 87.18.Sn, 87.10.+e, 05.45.−a

I. INTRODUCTION

The strength of biological synapses can only vary within a limited range, and there is accumulating evidence that some synapses can only preserve a restricted number of states (some seem to have only two [1]). These constraints have dramatic effects on networks performing as classifiers or as associative memories. Networks of neurons connected by bounded synapses whose efficacy cannot be changed by an arbitrarily small amount, share the palimpsest property (see, e.g., [2–5]): new patterns overwrite the oldest ones, and only a limited number of patterns can be remembered. The more synapses changed on each stimulus presentation, the faster is forgetting. The loss in synaptic structure caused by fast forgetting can be avoided by changing only a small fraction of synapses, randomly chosen at each presentation. Hebbian learning with stochastic selection permits the classification and memorization of an extensive number of random uncorrelated patterns, even if the number of synaptic states is reduced to two [4,6]. However, additional mechanisms must be introduced to store more realistic patterns with correlated components.

The stochastic algorithm we investigate here is based on the classical perceptron learning rule: the synapses are (stochastically) changed only when the response of the postsynaptic cell is not the desired one. In biology, this "stop-learning" property might be the expression of some regulatory synaptic mechanisms or the expectation of a reward signal. We show that some global inhibition, a small synaptic transition probability (the "learning rate") and a small neuronal threshold are sufficient to learn and memorize a linearly separable set of patterns with an arbitrarily high probability, provided that the number of neurons encoding the patterns is large to allow for the necessary redundancy required by the binary synapses. Global inhibition is required because plasticity in our model is restricted to excitatory synapses. Since the synaptic strengths are bounded, classifying tightly separated patterns is only possible if the postsynaptic neuron can finely discriminate between the inputs generated by the two classes. This fine discrimination is achieved by choosing a small neuronal threshold and inhibitory synaptic strengths far from saturation.

In general, finding binary weights for a threshold linear unit (a "perceptron") which should separate two sets of patterns is a combinatorially hard and NP complete problem [7,8]. The difficulty of the weight assignment problem for binary synapses is also reflected in the reduced storage capacity ($=p_{\max}/N=0.83$, relating the maximal number of patterns, $p_{\max}$, which can be stored in a network of N neurons, see [9]) compared to the capacity in case of continuous-valued synapses ($p_{\max}/N=2$, see [10,11]). No convergence theorem exists for a purely local learning algorithm with binary weights which asserts that linearly separable patterns with appropriate constraints can be learned in a finite number of presentations (see Appendix A). With our stochastic algorithm, the concergence is asserted with a high probability within a finite time. The probabilistic convergence time depends polynomially on the difficulty of the task (i.e., polynomially in $1/\epsilon$, where ϵ is the separation margin between the two sets of patterns): the tighter the separation between the two sets of patterns to be learned, the more presentations are required until the perceptron is expected to correctly classify the patterns. The probability of not converging within a specific number of presentations shrinks as $1/N$ when N increases while ϵ is kept fixed. Although the original problem of separating any linearly separable sets (i.e., with fixed N and *arbitrarily small* separation margin ϵ) with binary

*Electronic address: {wsenn,fusi}@cns.unibe.ch; URL: http://www.cns.unibe.ch/~{wsenn,fusi}

1539-3755/2005/71(6)/061907(12)/$23.00

weights is NP complete, the reduced problem of separating patterns with arbitrary large N and *fixed* separation margin ϵ is unlikely to fall in this complexity class. Hence, our probabilistic convergence theorem is neither a solution of the NP-completeness problem, nor is it a contradiction to the reduced storage capacity of binary synapses (arbitrary linearly separable sets can only be separated if they are embedded in a high enough N-dimensional space with fixed ϵ). Nevertheless, since the neurons in the brain are working in parallel, and since their number is abundant compared to the number of (substantially different) patterns to be classified, the stochastic algorithm may represent a biological "solution" of the binary weight assignment problem.

An interesting feature of bounded synapses is their self-stabilizing property. When presenting similar patterns with opposing outputs, the excitatory synaptic weights converge towards a unique steady state which depends on the learning rates and the rates of presenting the patterns. If this steady state excitatory weight is dominated by the global inhibitory weight, the neuron ceases to respond to patterns for which contradictory outputs are required. This suppression mechanism strongly improves the classification power of the network. In fact, using our stochastic perceptron learning algorithm for classifying preprocessed LATEX deformed letters to train multiple perceptrons, we obtain performances (~95% correct) close to the maximal ones reached (cf. [12] and citations therein). Instead of producing responses which are wrong with high probability, the postsynaptic currents become subthreshold during the course of the training and the neurons stay silent.

The presented algorithm is also important for neuromorphic hardware implementations of learning networks. The analog values representing the synaptic weights cannot be easily stored for long time scales (days or months), unless a digital approach is adopted. In fully analog VLSI chips, synaptic memories can be implemented by floating gates, which allow storing analog values with a resolution of a few bits (up to 4–5) [13]. Given that (1) the qualitative behavior of networks with discrete synapses does not change much when the number of preserved states increases [5], (2) the floating gate technology requires high voltages and sometimes nonstandard technologies, bistable (binary) synapses seem to be the simplest and the most efficient solution. The stochastic algorithm presented here, without the stopping condition, has been implemented by a spike-driven synaptic dynamics which can exploit the irregularities of the pre- and postsynaptic spike trains to generate activity-dependent random transitions between the two stable states [14–16]. After learning, in the absence of further stimulus presentations, the memories can be preserved indefinitely, and they are very robust also to the disrupting action of nonstimulus dependent spontaneous activity.

The paper is organized as follows: After presenting the neuron model, the learning rule and the formal theorem, we give an extended outline of the proof (Sec. III B). We then test the predicted finite convergence time and its dependency on the synaptic transition probability for sets of uncorrelated, linearly separable patterns (Sec. III C). To explore the benefits of the stochastic learning and the synaptic saturation we apply our algorithm to the classification of nonlinearly separable patterns with multiple perceptrons (Sec. III D). The discussion addresses the putative reasons for the good performance on nonseparable data sets, and hints to literature on a biologically more realistic, spike-driven implementation of the current algorithm. Appendix A explains why the "directed drift" argument previously used to "prove" the convergence of a similar stochastic algorithm for binary synapses [17] fails. Appendix B, finally, gives the rigorous proof of our theorem.

II. MODEL

A. Network model

We consider a network of N input neurons, each connected to M output neurons. All the input neurons feed a population of inhibitory cells, which in turn, project onto the output neurons. Neuron i is active $(a_i=1)$ if the total postsynaptic current h_i is above a threshold $\theta_o \in \mathbf{R}$, and inactive $(a_i=0)$ otherwise, $a_i=\mathcal{H}(h_i-\theta_o)$. The total postsynaptic current h_i is the weighted sum of the synaptic inputs from the network and some global inhibition, $h_i=(1/N)\Sigma_{j=1,j\neq i}^{N}(J_{ij}-g_I)a_j$, with a fixed inhibitory synaptic weight $g_I \in (0,1)$. The excitatory synaptic weight J_{ij} from the presynaptic neuron j to the postsynaptic neuron i is a stochastic variable, as explained below, and takes on the binary values 0 or 1.

During training, for each stimulus the input neurons are *clamped* to a specific pattern of activities $a_j=\xi_j^\mu$. A pattern of desired activities is imposed by an instructor to the output neurons $(a_i=\xi_i^\mu)$. The goal of learning is to modify the synapses in such a way that the desired output is produced by the input also in the absence of the instructor, i.e., when the output activity is entirely determined by the weighted sum of the inputs. In particular, if there are two possible desired outputs for each stimulus ($a_i=0$ or 1), then the goal is to find values $J_{ij}=0$ or 1 such that

$$\frac{1}{N}\sum_{j\neq i}(J_{ij}-g_I)\xi_j^\mu \begin{cases} > \theta_o + \delta_o & \text{if } \xi_i^\mu = 1 \\ < \theta_o - \delta_o & \text{if } \xi_i^\mu = 0 \end{cases} \tag{1}$$

for all patterns ξ^μ. The parameter $\delta_o \geqslant 0$ represents some learning margin.

Notice that in a recurrent network each unit can be regarded as both an input and an output neuron, so $N=M$. The same formalism and results also apply to the case of recurrent networks. In particular, condition (1) guarantees that each pattern ξ^μ is a fixed point of the network dynamics $a_i=\mathcal{H}[h_i(a)-\theta_o]$. These fixed points are also attractors in the limit of large N with a strictly positive (fixed) δ_o. In what follows we consider the case of a recurrent network. We also drop the index μ of ξ^μ because the same considerations apply to any generic pattern ξ.

B. Local stochastic learning rule

When the neuronal activities are clamped with a fixed binary activity pattern ξ, synapses stochastically switch their states depending on the pre- and postsynaptic activities, and depending on the total postsynaptic current. A synapse which is depressed $(J_{ij}=0)$ will be potentiated with probability q^+, provided that (1) the pre- and postsynaptic neurons are active, $\xi_j = \xi_i = 1$, but (2) the total postsynaptic current is not too large, $h_i \leqslant \theta_o + \delta_o$, with some learning margin $\delta_o \geqslant 0$. In turn, a synapse which is potentiated $(J_{ij}=1)$ will be depressed with probability q^-, provided that (1) the presynaptic neuron is active, $\xi_j = 1$ and the postsynaptic neuron inactive, $\xi_i = 0$, but (2) the total postsynaptic current is not too much below threshold, say $h_i \geqslant \theta_o - \delta_o$. The factors q^+ and q^- represent sufficiently small learning rates for potentiation and depression, respectively. The dynamics of the synaptic strengths evolves in discrete time steps, according to the sequential clamping of the network with different activity patterns. In summary, upon presentation of a pattern ξ^t at time t the synapses from an active presynaptic neuron j (i.e., with $\xi_j^t = 1$) to a postsynaptic neuron i change according to

$$J_{ij}(t+1) = \begin{cases} J_{ij}(t) + \zeta_j^+(1 - J_{ij}(t)), & \text{if } \xi_i^t = 1, h_i^t \leqslant \theta_o + \delta_o \\ J_{ij}(t) - \zeta_j^- J_{ij}(t), & \text{if } \xi_i^t = 0, h_i^t \geqslant \theta_o - \delta_o, \end{cases} \tag{2}$$

where $\zeta_j^\pm$ are binary random variables which are 1 with probability $q^\pm$ and 0 with probability $1 - q^\pm$, respectively. The saturation factors arise because a synapse will only be potentiated provided it is currently depressed, hence the factor $[1 - J_{ij}(t)]$, and a synapse will only be depressed provided it is currently potentiated, hence the factor $J_{ij}(t)$. We speak of a *synaptic update* for the postsynaptic neuron i if the synapses targeting i undergo a stochastic potentiation or depression, respectively, i.e., if the conditions in one of the lines in Eq. (2) are satisfied. The condition on the total postsynaptic current h_i^t in (2) is referred to as a *stop-learning condition* since it prevents h_i^t from increasing or decreasing more than just required to reproduce the correct output ξ_i^t.

C. On-line learning scenario

We consider a set $\mathcal{C}$ of p binary activity patterns $\xi = (\xi_1, \ldots, \xi_N)$ with $\xi_j \in \{0,1\}$. The patterns are repetitively presented to the network, such that each cycle of p patterns covers the whole set $\mathcal{C}$. When presenting pattern $\xi^t \in \mathcal{C}$ at time t, the N neurons will be clamped to the activities $\xi_1^t, \ldots, \xi_N^t$, and the total postsynaptic currents h_i^t are calculated by the neurons. Applying the learning rule (2), a synapse will stochastically potentiate (depress) if the conditions for potentiation (depression) are satisfied (i.e., if it is not yet potentiated or depressed, respectively, and if the corresponding conditions on ξ_i, ξ_j, and h_i are satisfied). Learning stops (converges) for each pattern if the learning thresholds are surpassed, $h_i > \theta_o + \delta_o$ if $\xi_i = 1$ and $h_i < \theta_o - \delta_o$ if $\xi_i = 0$, and the total currents h_i therefore faithfully reproduce the clamped activities in the sense of (1).

III. RESULTS

A. Learning linearly separable patterns with binary synapses

A necessary condition for a set of patterns $\mathcal{C}$ to consist of local attractors is that each of its patterns ξ satisfies the self-consistency conditions (1). In turn, these self-consistency conditions require that for each neuron i the subset of patterns ξ which activate i $(\xi_i = 1)$ is linearly separable from the subset of patterns which do not activate i $(\xi_i = 0)$. In other words, they require that $\mathcal{C}$ is *componentwise linearly ϵ-separable* for some $\epsilon > 0$, i.e., that for each neuron (component) i there is a separation vector $S = S^{(i)} \in [-1,1]^N$ with $S_i = 0$ and a separation threshold $\theta = \theta_i \in \mathbf{R}$ such that $\xi S > (\theta + \epsilon)N$ for all $\xi \in \mathcal{C}$ with $\xi_i = 1$, and $\xi S < (\theta - \epsilon)N$ for all $\xi \in \mathcal{C}$ with $\xi_i = 0$. The following theorem states that for fixed $\epsilon > 0$ and large N the componentwise linear ϵ separability is also sufficient for a class of patterns to be learned by the network with the stochastic synaptic updates. Under these conditions we show that for sufficiently small scaling factors $\varrho = \varrho_i > 0$ and sufficiently small transition probabilities (learning rates) $q = q_i^\pm > 0$ the synaptic dynamics (2) with neuronal threshold $\theta_o = \varrho\theta$ and learning margin $\delta_o = \varrho\delta$ is likely to converge within a finite number of presentations. For simplicity we assume that θ_o and δ_o are the same for all neurons i in the network, but any thresholds and learning margins below some value would also be admissible. The learning rate and the scaling factor depend on ϵ (and the choice of the global inhibition), but are kept fixed during the learning process.

Theorem: *Let $\mathcal{C}$ be an arbitrarily large set of componentwise, linearly $(\epsilon + \delta)$-separable patterns $\xi \in \{0,1\}^N$ with separability threshold θ (and $\epsilon > 0$, $\delta \geqslant 0$). Fix any inhibitory strength $g_I \in (0,1)$, any scaling factor $\varrho \leqslant \epsilon \overline{g_I}/16$, and any learning rate $q \leqslant ((\varrho\epsilon)^2 \overline{g_I}/8)^2$, where $\overline{g_I} = \min\{g_I, 1 - g_I\}$. Consider a recurrent network with neuronal threshold $\theta_o = \varrho\theta$, learning margin $\delta_o = \varrho\delta$, and global inhibition g_I. Then, for any repeated presentation of the patterns $\xi \in \mathcal{C}$ and any initial condition $J_{ij} \in \{0,1\}^{N^2}$, the synaptic dynamics (2) converges for large N (with fixed ϵ) with arbitrarily high probability in at most $n_o = 8/[q\overline{g_I}(\varrho\epsilon)^2]$ synaptic updates for each neuron. Fixing the separation margin ϵ, the probability of not converging within n_o updates scales as $1/N$.*

A formal proof of the theorem is given in Appendix B. The assumption that the separation parameter ϵ is fixed while N (and perhaps also p) increases yields the necessary redundancy for encoding the patterns across the N neurons. This redundancy may not be present if the number of patterns p is arbitrarily growing with N. In fact, the probability that $p = cN$ randomly chosen patterns are (componentwise) linearly separable is below 0.5 for $c > 2$, and for fixed $c > 2$ it drops to 0 with increasing N [10]. In turn, the patterns are likely to be separable for large N if $c < 2$ (in the limit of large but fixed N the expected separation margin ϵ drops as $1/\sqrt{p}$, see [18], Eq. (7)). In any case, increasing the number of neurons (N) while fixing the number of random binary patterns (p), makes it likely that the patterns become linearly separable. Note that an appropriate learning rate q and an appropriate scaling factor ϱ are not required to be in the order of $1/N$. These parameters only scale with a power of the separation margin ϵ and with the distance $\overline{g_I}$ of the global inhibitory weight from its boundaries 0 and 1, but they do not depend on the network size N.

B. Outline of the proof

Since during the learning process the neuronal activities are clamped to fixed values (ξ_j) we may discard the recurrent connections and consider each neuron individually. Picking out any postsynaptic neuron i we have to show that learning stops for the synapses projecting onto that neuron. Dropping the index i we abbreviate $J^t=[J_{i1}(t),\dots,J_{iN}(t)]$, and the total synaptic strength onto neuron i is written as $J_I^t=J^t-g_I\mathbf{1}$. The set of patterns $\mathcal{C}$ splits into the two subsets $\mathcal{C}^+$ and $\mathcal{C}^-$ composed of patterns which either activate or do not activate neuron i, $\xi_i=1$ or $\xi_i=0$, respectively.

1. Controlling synaptic saturation

The general strategy of the convergence proof is to approximate the discrete-valued synaptic dynamics by the mean field dynamics with analog synaptic strengths, as treated in [19,20]. The "mean field" at time t is defined by the expectation values $\langle\langle J_I^t\rangle\rangle$ of the total synaptic weight vector $J_I^t=J^t-g_I1$ across all trajectories $J_I^{t'}$ up to $t'=t$. As a first step one proves that for tightly separated patterns (small ϵ) $\langle\langle J_I^t\rangle\rangle$ converges for $t=1,2,\dots$ to a scaled solution vector ϱS separating the classes $\mathcal{C}^+$ and $\mathcal{C}^-$. The convergence is enforced by the "teacher" who "tells" whether the desired output ($\xi_i^t=1$ or 0) could be reproduced by the neuron [in which case no synaptic update occurs because the condition on h_i^t in Eq. (2) is not satisfied] or could not be reproduced (in which case a distorted fraction of the input pattern ξ^t is added or subtracted to the expected weight vector, depending on whether the output should be 1 or 0, respectively).

If there would be no synaptic saturation, the convergence $\langle\langle J_I^t\rangle\rangle\to\varrho S$ would follow as in the classical perceptron convergence proof (see, e.g., [21]; compare also Sec. III B 4 below). In the case of synaptic saturation, however, a distortion of the expected update vector arises (the "forgetting" part) which drives the expected excitatory weight vector $\langle\langle J^t\rangle\rangle$ steadily away from the boundary. Without the stop-learning condition, when learning infinitely occurs, synaptic saturation drives this expected weight vector towards some asymptotic state where the learning effort is balanced by the synaptic saturation. In the presence of the stopping condition, this asymptotic state may not be reached. Instead, after a successful learning, the weight modifications stop when the distribution of the postsynaptic currents is narrowly clustered around the neuronal threshold, with a peak just above $\theta_o+\delta_o$ and a peak just below $\theta_o-\delta_o$. If the threshold scaling factor ϱ is small, $\theta_o=\varrho\theta$ and $\delta_o=\varrho\delta$ are both small, and the final distribution of the postsynaptic currents will be close to 0. This is only possible if most of the components of $\langle\langle J_I^t\rangle\rangle$ become small. In fact, learning tightly separable patterns drives the expected total weight vector towards the scaled solution vector, $\langle\langle J_I^t\rangle\rangle=\langle\langle J^t\rangle\rangle-g_I\mathbf{1}\to\varrho S\approx\mathbf{0}$, such that after learning all the components components are small, $\langle\langle J_I^t\rangle\rangle\approx\mathbf{0}$. As a consequence, the expected excitatory weight vector approaches the global inhibitory weight vector, $\langle\langle J^t\rangle\rangle\approx g_I\mathbf{1}$. If the global inhibitory strength is in the middle of the maximal and minimal synaptic strength, $g_I=0.5$, learning pushes the expected excitatory weight vector towards the center of the hypercube, $\langle\langle J^t\rangle\rangle\to\mathbf{0.5}$.

The benefit of choosing the global inhibition around 0.5 is that both the synaptic saturation *and* the learning effort, are pushing the expected weight vector away from the boundary towards the hypercube center. As $\langle\langle J^t\rangle\rangle$ approaches the hypercube center, however, synaptic saturation starts to counteract the learning because saturation tends to drive the weight vector into a uniform equilibrium state in which any synaptic structure acquired by the learning is flattened out. Fortunately, when the excitatory weight vector is close to the hypercube center, forgetting becomes negligible, while the effect of the learning (the linear part) remains finite. Synaptic saturation can therefore be controlled by choosing a small scaling factor ϱ which gates the dynamics of the expected excitatory weight vector towards, but not on to, the center of the hypercube. Far from the synaptic bounds learning is dominated by the linear part, as in the classical perceptron learning without synaptic bounds, and the expected excitatory weight vector may converge towards a possible solution vector, $\langle\langle J^t\rangle\rangle\to\varrho S+g_I\mathbf{1}$. Of course, to prevent overshooting, a small threshold scaling factor ϱ also requires a small learning rate q.

2. Problem of synaptic correlations

There is a problem, though, with this strategy of proof because a description of the dynamics of $\langle\langle J^t\rangle\rangle$ as a function of the expected total current $\langle\langle h^t\rangle\rangle$ would require that the trajectories J^t follow arbitrarily close the trajectory of $\langle\langle J^t\rangle\rangle$ as the network size increase. However, due to the stochasticity in the synaptic updates there are always trajectories which strongly deviate from the mean (and which actually do not converge), and only "typical" trajectories J^t with typical synaptic updates may follow the mean $\langle\langle J^t\rangle\rangle$ until convergence. Typical trajectories remain close to each other only if their synapses are updated at exactly the same time steps. This is because the stop-learning condition introduces correlations among the synaptic states which may sum up in time and become large. These correlations produce a variance in the total postsynaptic current h^t at time t which can be as large as $1/N+q$, with $q=q^\pm$ the learning rate(s), whereas without stop-learning condition the variance in h^t is bounded by $1/N$. Such a variance is too large to be neglected because the total number of updates necessary for the convergence increases with the inverse of the learning rate $(1/q)$, and the expected deviations of the total postsynaptic current from the mean can therefore accumulate throughout the learning process up to 1. The dynamics of $\langle\langle J^t\rangle\rangle$ can therefore only be described by tracking subclasses of the full distribution of possible trajectories.

3. Restriction of the synaptic dynamics to subclasses

To take account of the synaptic correlations we partition the space of all possible trajectories into subclasses of trajectories following the same update sequence, i.e., satisfying the same update inequalities imposed onto $h^{t'}$ at any time step $t'\leqslant t$. Instead of tracing the mean $\langle\langle J^t\rangle\rangle$ across all possible trajectories, we trace the individual means $\langle J^t\rangle$ across trajectories of the same subclass. Restricted to such a subclass with the same update sequences, one can show that the "subclass variance" of h^t is small enough and, in fact, shrinks to 0 as N becomes large. Our proof, however, moves along a slightly different path. We show that with high probability an individual subclass mean uniformly converges within a finite time. By definition of the subclass, all trajectories simultaneously satisfy or do not satisfy the update conditions, and the dynamics of the subclass mean therefore stops if and only if each individual trajectory within the subclass stops. Since the subclasses together cover all possible trajectories, the convergence is assured with high probability for all trajectories.

4. Convergence of each subclass mean

To prove the subclass convergence we show that each time step t the subclass mean $\langle J_I^t \rangle$ strictly moves with a minimal positive step size towards the scaled solution vector ϱS. Since the initial distance $\| \langle J_I^0 \rangle - \varrho S \|$ is finite, the convergence with such strictly positive step sizes must stop. To show that the distance from $\langle J_I^t \rangle$ to ϱS decreases each time step by a minimal amount we have to show that (1) the update vector $\langle \Delta J^t \rangle$ forms an angle strictly smaller than 90° with the direction from $\langle J_I^t \rangle$ to ϱS, and that (2) the update vector $\langle \Delta J^t \rangle$ is not too long (to prevent overshooting). The first requirement directly follows from the learning rule when the saturation is small: If the synaptic saturation in Eq. (2) would be neglected, the expected update vector would be $\langle \Delta J^t \rangle = \pm q \xi^t$, depending on whether the condition for a long-term potentiation [LTP, upper line in Eq. (2)] or a long-term depression [LTD, lower line in Eq. (2)] is satisfied, respectively. In the case of LTP, e.g., the separability assumption states that $\xi^t \varrho S > \varrho(\theta + \epsilon) N$ and the condition on h^t in Eq. (2) states that $N\langle h^t \rangle = \langle J_I^t \rangle \xi^t \leqslant \varrho \theta + \varrho \delta$. Combining these two inequalities yields $(\varrho S - \langle J_I^t \rangle) q \xi^t \geqslant q \varrho \epsilon N$. The same estimate is also obtained in the case of LTD. If we now take synaptic saturation into account, the expected update vector in case of LTP, e.g., becomes $\langle \Delta J^t \rangle = q \xi^t (\mathbf{1} - \langle J^t \rangle)$, and this can be written as $\langle \Delta J^t \rangle = q \xi^t (\mathbf{1} - g_I \mathbf{1}) - q \xi^t \langle J_I^t \rangle$. The additional factor $(\mathbf{1} - g_I \mathbf{1})$ does not harm since the components are identical. But the additional forgetting part, $\langle \Delta F \rangle = -\xi^t \langle J_I^t \rangle = -\xi^t \langle J^t - g_I \mathbf{1} \rangle$, may well distort the update vector. Fortunately, the distortion is small if each component of the expected weight vector $\langle J^t \rangle$ is close to $g_I \approx 0.5$. Moreover, because of the negative sign in front of $\xi^t \langle J^t - g_I \mathbf{1} \rangle$, the forgetting part actively drives $\langle J^t \rangle$ towards this hypercube center at 0.5 where $\langle \Delta F \rangle \approx 0$. We conclude that for small ϱ (defining the final distance of $\langle J^t \rangle$ from $g_I 1$, cf. Sec. III B 1 above) and g_I close to 0.5, the angle between $\langle J_I^t \rangle$ and ϱS is strictly below 90°, $(\varrho S - \langle J_I^t \rangle) q \xi^t \geqslant q \varrho \epsilon N$. It remains to be shown that the update vector is short, more precisely, that $\| \langle \Delta J^t \rangle \|^2 < q \varrho \epsilon N$.

5. Synaptic correlations are small within large subclasses

Unfortunately, the smallness of the expected change $\| \langle \Delta J^t \rangle \|^2$ is not evident. It can be large if the synaptic correlations evoked by the update condition imposed on h^t are strong. In fact, due to this update condition, a subclass defined by an update sequence $\Delta J_j^{t'}$ up to time t could be composed of only a single trajectory. In this case the above norm square would just count the number of synaptic transitions at time t, $\| \langle \Delta J^t \rangle \|^2 = \Sigma_j \langle \Delta J_j^t \rangle^2 = \Sigma_j | \Delta J_j^t |$, and this can be in the order of qN or even larger. Note that the expected (subclass) transition at a specific synapse j, in the case of a single trajectory in the subclass, is equal to $\langle \Delta J_j^t \rangle = \Delta J_j^t = 0$ or ± 1. Only if there are many trajectories within a subclass, such that they are faithfully sampling the transition probability q at each synapse, will the expected transitions become small, say $\langle \Delta J_j^t \rangle^2 \leqslant q^{3/2}$, and $\| \langle \Delta J^t \rangle \|^2$ would be smaller than $q^{3/2} N$. The requirement of a large number of equivalent trajectories, i.e., trajectories giving the same increment $\Delta h^t = (1/N) \Sigma_j \Delta J_j \xi_j$ and therefore staying in the same subclass, is a redundancy requirement onto the synaptic encoding. We next show that with large N the equivalent trajectories becomes numerous and that in each subclass up to time t there are enough trajectories such that $\| \langle \Delta J^t \rangle \|^2$ is small.

6. Redundancy assures large subclasses and hence good approximation of the dynamics by the subclass means

To assure that most of the subclasses contain many trajectories one might simply duplicate the synapses between individual neurons, or equivalently, duplicate the neurons which encode an individual component of the patterns. However, such an explicit coding scheme is not necessary. Instead, the smallness of $\| \langle \Delta J^t \rangle \|^2$ is obtained from increasing the number of (presynaptic) neurons, N, while encoding the p patterns such that the separation margin ϵ does not shrink. Two cases are considered: (1) In the case that only a few synapses satisfy the update condition on ξ_j, ξ_{post} and h^t in the learning rule [Eq. (2)], the subclass mean $\langle \Delta J_j^t \rangle^2$ might still be large for these synapses (perhaps even 1). But since $\langle \Delta J_j^t \rangle^2 = 0$ for all other synapses which do not satisfy the update conditions, we obtain $\| \langle \Delta J^t \rangle \|^2 \leqslant q^{3/2} N$. (2) In the case that many synapses satisfy the update conditions, there will be many different stochastic transitions with the same effect on the total postsynaptic current h^t, and therefore giving trajectories within the same subclass. Averaging over these many trajectories one gets also in this case $\| \langle \Delta J^t \rangle \|^2 \leqslant q^{3/2} N$, as argued above. To get $q^{3/2} N \leqslant \varrho q \epsilon N$ as required above we need a second time that the learning rate q is small. We conclude that each time step the distance from the subclass mean $\langle J_I^t \rangle$ to the solution ϱS shrinks by at least a fixed amount. This completes the convergence proof.

C. Classifying uncorrelated random patterns

To test the statements of the theorem we trained a single binary perceptron with stochastic learning on random uncorrelated binary patterns, as in [22]. In Fig. 1(a) we show the number of iterations per pattern until learning for a fixed postsynaptic neuron stops as a function of the number of neurons N of the input layer. We considered $p = 10$, 20, and 40 random uncorrelated binary patterns, generated with a probability of 1/4 for a neuron to be active. As expected, the finite size effects decrease with N and the number of iterations tend asymptotically to a value which depends only on the number of patterns. This is because the separation margin of the two classes decreases with increasing number of patterns. Figure 1(b) shows, the number of iterations per pattern needed for convergence as a function of the scaling factor ϱ

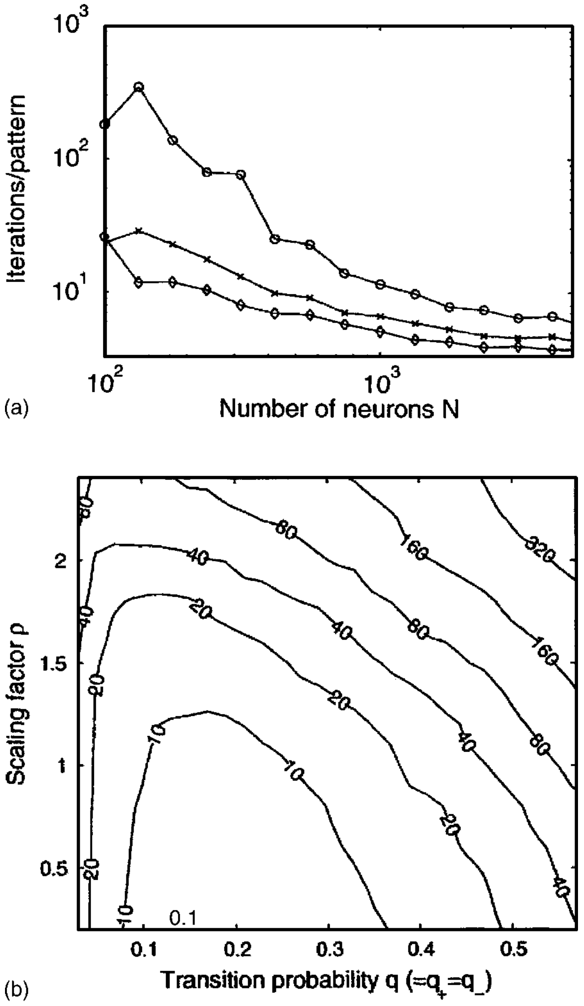

FIG. 1. Convergence time as a function of different parameters. (a) Number of iterations per pattern as a function of the number of neurons N for $p=10$, 20, and 40 uncorrelated random patterns ($q=0.05, f=1/4, \theta=0.01$); (b) Number of iterations per pattern as a function of the learning rate (synaptic transition probability) q and the scaling factor ϱ. Convergence is only guaranteed if ϱ and q are small. Note that for small q the convergence time only increases because the "step size" decreases, and not because the combinatorial problem becomes difficult as is the case for large ϱ and large q.

and the transition probabilities $q=q^{+}=q^{-}$ for random uncorrelated binary patterns ($p=10$, $N=100$). If learning is too fast or ϱ is too large, the number of iterations grows very quickly and eventually becomes impossible to converge. Note that there is an optimal learning rate for each threshold. While a large learning rate can prevent the convergence, a very small learning rate will only slow down the learning process, although convergence remains guaranteed. The long convergence time is due to the fact that there is a minimal distance from the initial weight vector to the set of possible solution vectors which needs to be crossed and which leads to a scaling of $1/q$ of the number of required iterations. The convergence time also increases if the global inhibitory strength g_I approaches 0 or 1 (not shown, but see [19,20]).

D. Classifying nonlinearly separable patterns with multiple perceptrons

We also trained the perceptron on more complex, LATEX deformed characters [Fig. 2(a)], preprocessed as in [12]. The goal is to classify $32n$ nonlinearly separable patterns organized in n classes ($n=10-200$). As the patterns are not linearly separable and therefore not classifiable by a single perceptron, each class is learned by a group of ten independent perceptrons. For each of the n classes the $N=2010$ input neurons project to all ten output neurons within the group. Each output neuron of one specific group is trained to be selective to the 32 patterns of one class. For example, during the learning phase the first group of output neurons is activated (clamped to 1) only when one of the 32 samples of the letter "a" are presented. It is silent (clamped to 0) for any other pattern. The second group of output neurons is trained to respond only to "b"s, the third group to the "c"s, etc. All the patterns are repeatedly presented in a fixed order, and every time the synapses are randomly updated according to the stochastic learning rule with $q^{+}=q^{-}=0.01$. Learning might stop after a finite number of iterations (e.g., for a small number of classes), or there might be always errors. In the latter case the simulation is stopped after 300 repetitions per pattern. In the test phase, only the presynaptic neurons are clamped, and the activities of the output neurons are obtained from thresholding the total postsynaptic currents. The input pattern is classified by a majority rule: the group of neurons with the most active neurons determines to which class the input pattern is assigned. The neuronal threshold θ_o is set to $5/N$, and the margin δ_o for stopping the learning is also $5/N$ (corresponding to a difference of five neurons). On average each pattern activates 50 neurons of the input layer, but the coding levels vary over a wide range (from 10 to 100 neurons).

Figures 2(b) and 2(c) show the distribution of the total postsynaptic currents generated by the different patterns across those output neurons which should get activated (solid line), and across those which should not get activated (dashed line). Before learning [Fig. 2(b)] the two distributions are very similar because the initial synaptic weights are random and not correlated with the patterns. After learning [Fig. 2(c)], they are well separated by the neuronal threshold, allowing a classification without errors. In the present example, learning converged because only a small number of classes were used (26 classes corresponding to the letters of the alphabet). Although the patterns are highly correlated, they are apparently still linearly separable.

In the case of nonlinearly separable patterns, the situation is more complex. When each pattern is presented for classification there are three possible outcomes: (1) no output unit is activated: the pattern is *not classified*; (2) the majority of active output units belong to the correct class: the pattern is correctly classified; (3) the majority of active output units belong to the wrong class: the pattern is *misclassified*. The results are shown in Fig. 3, where we plot the fraction of misclassified (a) and nonclassified (b) patterns as a function of the number of classes. In the case shown in the figure, the fraction of misclassified patterns is very small compared to the fraction of nonclassified patterns. The ratio between these

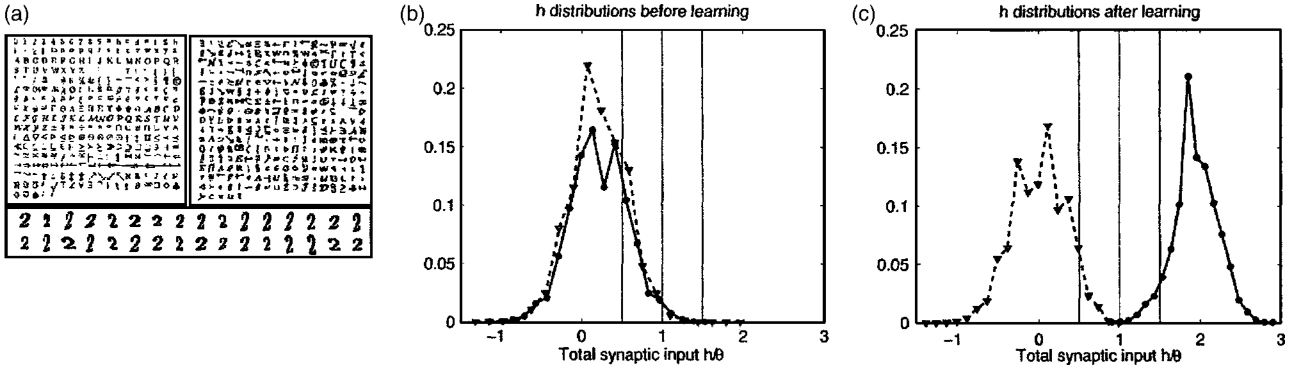

FIG. 2. Classification of nonlinearly separable patterns: (a) Deformed LATEXcharacters used for the benchmark test. Left: prototype letters; right: random sample of deformed letters; bottom: sample of deformed letters of class "2" (reproduced from [12]). (b),(c) Distributions of the total postsynaptic currents h_i evoked by the patterns belonging to the two different classes $\mathcal{C}^+$ (solid line, pooling together all output units which should get activated) and $\mathcal{C}^-$ (dashed line, pooling together all output units which should not get activated), averaged over the 26×10 output neurons representing the letters of the alphabet in groups of 10. While before learning both classes evoked subthreshold currents (b), the two classes are well separated after the training (c), with patterns $\xi \in \mathcal{C}^+$ evoking suprathreshold currents (solid line) and patterns $\xi \in \mathcal{C}^-$ evoking subthreshold currents (dashed line). Vertical lines represent the neuronal threshold θ_o, flanked with the stop-learning thresholds $\theta_o \pm \delta_o$.

two quantities depends on the statistics of the patterns and on the ratio between q^+ and q^-. As the number of depressing events (patterns which satisfy the condition for LTD) increases compared to the number of potentiating events, the distribution of the total synaptic current drifts to lower values, thereby inactivating a larger number of output units. This increases the fraction of nonclassified patterns, but lowers the fraction of misclassified patterns. The network becomes more undecisive, but also more reliable in the classification task.

IV. DISCUSSION

We have shown that any set of linearly separable patterns can be learned by our local stochastic learning rule with discrete-valued synapses. The restriction of the synaptic plasticity to excitatory synapses makes global inhibition necessary. Moreover, a tight separation margin between the two classes of patterns requires a small learning rate (implemented in the form of small synaptic potentiation and depression probabilities) and a small neuronal threshold.

A. Slow learning and redundancy

In this paper we fix the learning rate and the neuronal threshold in advance, depending on the difficulty of the classification task. However, they might also be adjusted by some homeostatic mechanism operating during the learning process. For instance, the threshold might slightly decrease if the clamped activity is not correctly predicted by the total postsynaptic current, and it might slightly increase in the other cases. Note that decreasing the neuronal threshold is equivalent to increasing all the excitatory synaptic weights, as it arises in biological neurons through homeostatic plasticity [23].

The probability for the learning process to converge within some fixed number of presentations increases (as $1-1/N$) as the number of neurons (N) tends to infinity. While the number of neurons increases, the separation margin (ϵ) must remain strictly positive (i.e., bounded away from 0). This is a form of redundancy which is necessary for learning with discrete synaptic weights. Together with the slow learning, it represents the price for solving a combinatorially difficult task with a learning rule which is purely local in space and time. This is consistent with the fact that the maximal storage capacity of networks with binary synapses is smaller than the one for synapses with continuous weights (see [9,10] and Sec. I). Moreover, the redundancy implies a solution of the original, NP-complete weight assignment problem for classifying linearly separable patterns.

B. Spike-driven stochastic implementation

The stochasticity in the synaptic modification is more than just a slowing down of the learning process. The stochastic selection of the synapses spatially decorrelates the synaptic updates which in turn allows for an optimal redistribution of the synaptic resources and to classify nonlinearly separable patterns.

The stochastic selection mechanism can be implemented in terms of a detailed spike-driven synaptic dynamics by exploiting the irregularity of the spike trains. A synaptic modification, for instance, could only be triggered upon coincidences of some pre- and postsynaptic spikes within a fixed time window or by the accumulation of coincidences of presynaptic spike and high postsynaptic depolarization [5,14,24,25]. The stopping mechanism can be implemented in terms of these dynamic features of biological synapses [26]. In all these cases the load of generating the noise to drive the stochastic selection mechanism is transferred outside the synapse and is delegated to the collective behavior of the interacting neurons which may show highly irregular spiking patterns [15]. By this "out-sourcing" of the noise-generating machinery it becomes possible to control arbitrarily small transition probabilities.

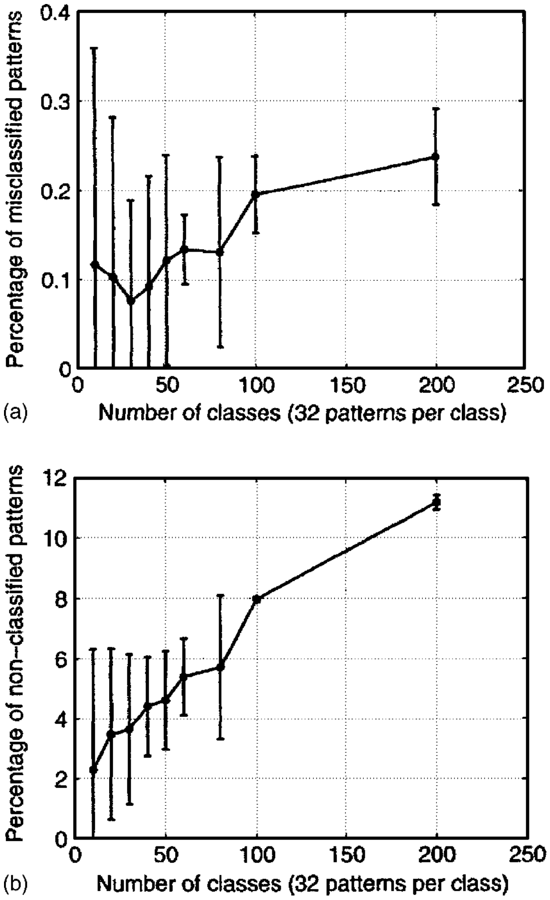

FIG. 3. (a) Fraction of misclassified patterns (for which none of the $n \times 10$ output units is activated) and (b) fraction of nonclassified patterns (for which the majority rule applied to the n groups of ten output units gives the wrong classification) as a function of the number n of classes of preprocessed LATEXdeformed characters. Each class contains 32 different patterns [see example of class "2" in Fig. 2(a)]. Both fractions increase as a power law as the number of classes increases. Note that the number of misclassified patterns is an order of magnitude smaller than the number of nonclassified patterns. The misclassification is kept small by exploiting the fact that the neurons tend to shut down when exposed to nonlinearly separable patterns (see Discussion). Each point in the two panels represents the average performance across ten different choices of n classes ($n=10,20,\ldots,200$ as indicated on the abscissa; for $n=100$ and 200 we only chose two class sets). The error bars indicate the standard deviation.

C. Classification of nonlinearly separable patterns

A single output neuron can only classify linearly separable patterns. However, when the stochastic learning rule is applied to a network of multiple output units, it becomes possible to discriminate between patterns which are not linearly separable. We showed that the classification performances on a large complex data set (LATEXdeformed characters) are surprisingly good, better than the ones of recent models which in general are more complex and require parameter tuning. For example the best performance on 293 classes in [12] is 60% correct, while in our case we have 94.5% correct when the same number n of output units per class is used ($n=20$). The authors of [12] need a complex and unnatural boosting technique to achieve a comparable performance. A first ingredient for our good performance resides in the fact that each output unit experiences a different realization of the stochastic process generated by the binary random variables ζ when updating the synapses [see Eq. (21)]. This means that each output unit will end up in classifying the patterns according to a different hyperplane. When the information of all output units is combined, the classification of nonlinearly separable patterns becomes possible. A weaker form of stochasticity which is based on the quenched randomness of the connections is exploited in [12].

A second ingredient for the good performance is related to the read-out from the output units and depends on the statistics of active output units: in order to read the relevant information, only the output units with a reliable response should be considered for the majority evaluation, while the other units should be silent (and therefore will only contribute to a nonclassification, not to a misclassification, see Fig. 3). The average fraction of active output neurons can be controlled by changing the ratio between q^+ and q^-, which in turn determines the asymptotic distribution when learning cannot converge. The existence of an asymptotic distribution and a fast convergence towards this distribution are guaranteed by the discreteness and the boundedness of the synapses. Although we did not present a theory for multiple output units in the case of nonlinearly separable patterns, the above reasoning is proven to apply to a simplified scenario [20]: when contradictory patterns are presented (i.e., when the very same pattern belongs to two different classes), the output units will be likely shut down, provided that the ratio of the effective LTD rate $\tilde{q}^-$ over the effective LTP rate $\tilde{q}^+$ is large enough. This ratio in general will depend on the statistics of the patterns and on the number of classes. The tuning of these effective LTD and LTP rates might be realized by other mechanisms like homeostatic plasticity [23].

D. Discrete versus continuous synapses

For simplicity, but also because of the direct applications to learnable VLSI networks performing a memory task, we were focusing on the case of binary synapses. However, the convergence theorem similarly holds with a general, discrete-valued finite set of synapses, and even with discrete-valued neuronal activities. Whether on the macroscopic level synapses can be modified in discrete or continuous steps remains to be further investigated [1]. Naturally, the synaptic strengths are always bounded. Due to this boundedness, phenomena such as balancing of excitation emerge during the learning process [20]. The same phenomena are also expressed for the potentiation probabilities of finite, discrete-valued synapses. The main difference in terms of learning is that (finite) discrete-valued synapses require slower learning than continuous-valued (bounded) synapses. The slower learning, on the other hand, is compensated by an increased memory stability endowed by the discreteness of the synaptic states.

ACKNOWLEDGMENTS

This work was supported by SNF Grant No. 3152A0-105966/1, the Silva Casa Foundation, and European Union

Grant No. IST-2001-38099 ALAVLSI. We would like to thank M. Mascaro for providing us with the dataset of preprocessed LATEXdeformed characters and for useful inspiring discussions about the role of the stopping condition. We are grateful to E. Amaldi for discussions about NP-complete problems, and to an anonymous referee for many helpful comments.

APPENDIX A: DIRECTED DRIFT FAILS

A form of the present stochastic algorithm (directed drift) was studied in [17], and arguments proving the convergence were given. However, the directed drift argument fails in the way it is used in [17]. To expose the problem, and to motivate the notions of typicality and redundancy introduced in the formal convergence proof below, we present a simple example.

According to the directed drift argument the distance $\|J^t - S\|$ from the synaptic state vector J^t to the solution vector S would show a nonvanishing negative drift, and therefore would shrink to 0 with high probability within a fixed time. Unfortunately, in general this is not true. Let us assume that the single pattern $\xi=(1,\ldots,1)$ with an even number of N components has to be learned with output $\xi_0=1$. Assuming a threshold $\theta_o=1/2$, a possible solution vector is $S=(1,\ldots,1,0,\ldots,0)$, where the number of 1's is $N/2$. Let us consider the synaptic state $J^t=(1,\ldots,1,0,\ldots,0)$ with a slightly smaller number $N/2-n$ of 1's, say $n=\lfloor \epsilon N \rfloor$ with some small $\epsilon > 0$. This synaptic state leads to the independent stochastic potentiation of synapses $N/2-n+1,\ldots,N$ with probability q. In this example the distance from J^t to S increases, $\|J^{t+1}-S\| > \|J^t - S\|$, with high probability. This is because only a potentiation of the first few (n) synapses $N/2-n+1,\ldots,N/2$ brings J^t closer to S, while the potentiation of all the remaining ($N/2$) synapses $N/2+1,\ldots,N$ moves J^t away from S. If $N \gg n$ the probability that J^t moves away from S is therefore arbitrarily close to 1.

What is always true, however, is that the *expected* weight vector $\langle\langle J^t \rangle\rangle$ converges to S, provided that q is small enough. In fact, the expected synaptic state at the next time step is $\langle J^{t+1}\rangle=(1,\ldots,1,q,\ldots,q)$, and its distance to S is smaller than the distance from J^t to S, $\|\langle J^{t+1}\rangle - S\|^2 = n(1-q)^2 + (N/2)q^2 < \|J^t - S\|^2 = n$, provided that q is smaller than 2ϵ. Hence, the drift argument must be applied to the expectation values, and it has to be assured that the stochastic dynamics closely follows the dynamics of the expectation values. This is the case for typical sequences since, for fixed ϵ and large N, the stochastically selected components are reliably sampling the subsets of size $(N/2-n)/N$, n/N, and $N/(2N)$, respectively. Beside this redundancy argument, the proof must deal with the problem of synaptic saturation (forgetting), as it also arises in the case of bounded synapses with continuous strengths [20].

APPENDIX B: PROOF OF THE THEOREM

1. Typical trajectories

Let us fix an arbitrary sequence of patterns $\{\xi^t\}_{t=0,1,\ldots}$ which repeatedly cycles through the p patterns $\xi \in \mathcal{C}^{\pm}$. Let the pattern $\xi=\xi^t$ have aN nonvanishing components (a is the coding level of the pattern). Let us assume that ξ gives rise to a stochastic *update*, i.e., that the condition on the total postsynaptic current $h=(J-g_I 1)\xi/N$ in the learning rule (2) is satisfied. Synapse j is selected with probability $q\xi_j$. Let ζ be the vector indicating that synapse j got selected ($\zeta_j=1$) or not ($\zeta_j=0$). The mean and variance of ζ_j are given by $\langle\langle \zeta_j \rangle\rangle = q\xi_j$ and $\mathrm{var}\zeta_j = q\xi_j(1-q\xi_j) \leqslant q\xi_j$, respectively. A stochastic update vector ζ is called *typical* if $|1\zeta - 1\langle\langle\zeta\rangle\rangle| \leqslant (q^2/2)aN$, i.e., if the number of selected synapses does not deviate too much from its expectation value. Since $1\langle\langle\zeta\rangle\rangle \leqslant qaN$ and $\mathrm{var}(1\zeta) \leqslant qaN$ we conclude with the Chebyshev inequality that any randomly sampled update is typical with probability larger than $1-\epsilon_o$, i.e.,

$$\mathcal{P}\left\{ |1\zeta - 1\langle\langle\zeta\rangle\rangle| \leqslant \frac{q^2}{a}aN \right\} \geqslant 1 - \frac{4\,\mathrm{var}(1\zeta)}{(q^2 aN)^2} \geqslant 1-\epsilon_o, \quad \text{with} \quad \epsilon_o = \frac{4}{q^3 aN}, \tag{B1}$$

Where, $\mathcal{P}$ denotes the probability measure on $\{0,1\}^N$ induced by $\mathcal{P}\{\zeta \in \{0,1\}^N | \zeta_j = 1\} = q\xi_j$. A trajectory is called typical if each synaptic update is typical. The following property will be used several times in the sequel. Let $x \in [-1,1]^N$. By applying Chebyshev's inequality twice we get

$$\mathcal{P}\left\{ \zeta \text{ is typical and } |x\zeta - x\langle\langle\zeta\rangle\rangle| \leqslant \frac{q^2}{2}aN \right\} \geqslant 1 - 2\frac{4\,\mathrm{var}(x\zeta)}{(q^2 aN)^2} \geqslant 1-2\epsilon_o. \tag{B2}$$

Note that the factor of 2 arises because we require two conditions, $|1\zeta - 1\langle\langle\zeta\rangle\rangle| \leqslant (q^2/2)aN$ as imposed in (B1), and also $|x\zeta - x\langle\langle\zeta\rangle\rangle| \leqslant (q^2/2)aN$. Each of these conditions is satisfied with probability larger than $1-\epsilon_o$.

2. Trajectories with identical update sequences

Each trajectory J^t of (2) specifies a binary sequence $\alpha(J)$ with $\alpha(J)^t=1$ or 0, depending on whether or not the condition for a synaptic update on h^t is satisfied. Let us choose some sequence α^t of 0's and 1's. Let $\mathcal{T}_t^\alpha$ denote the set of typical trajectories J having the same update sequence α up to time t, i.e., $\alpha(J)^{t'}=\alpha^{t'}$ for $t'=1,\ldots,t$. We will show that there is a t_o such that for all $t>t_o$ the set $\mathcal{T}_t^\alpha$ is either empty or $\alpha^t=0$, i.e., for $t>t_o$ no synaptic update takes place. Let us assume that the set $\mathcal{T}_t^\alpha$, and therefore also $\mathcal{T}_{t+1}^\alpha$, is not empty.

Let us assume that the condition on h^t for a stochastic update is satisfied, i.e., that $\alpha^t=1$. Let us write (2) in the form $J^{t+1}=J^t+\Delta J^t$ with $\Delta J^t = \zeta * (1-J^t)$ if $\xi^t \in \mathcal{C}^+$ and $\Delta J^t = -\zeta * J^t$ if $\xi^t \in \mathcal{C}^-$. Here, "$*$" denotes the componentwise product of vectors. Let $\langle J^t \rangle$ be the expected synaptic strength across the trajectories in $\mathcal{T}_t^\alpha$ at time t. Similarly, $\langle \Delta J^t \rangle$ denotes the expected change of J^t when averaging over $\mathcal{T}_{t+1}^\alpha$. We decompose $\langle \Delta J^t \rangle$ into a *linear* part ΔL and a *forgetting* part ΔF. Setting $J_I = J - g_I 1$ and dropping the time index we obtain from (2);

$$\langle \Delta J\rangle = \Delta L + \Delta F = \begin{cases} (1-g_I)\langle\zeta\rangle - \langle \zeta * J_I\rangle, & \text{if } \xi\in\mathcal{C}^+ \\ \quad -g_I\langle\zeta\rangle - \langle\zeta * J_I\rangle, & \text{if } \xi \in \mathcal{C}^-, \end{cases} \tag{B3}$$

where $\Delta F=-\langle\zeta * J_I\rangle$ and $\Delta L=(1-g_I)\langle\zeta\rangle$ or $\Delta L=-g_I\langle\zeta\rangle$, depending on whether $\xi\in\mathcal{C}^+$ or $\xi\in\mathcal{C}^-$, respectively.

3. Learning based on the linear part

According to the update and separability condition for the case $\xi\in\mathcal{C}^+$ we have $\xi J_I<\varrho(\theta+\delta)N$ and $\xi\varrho S>\varrho(\theta+\delta+\epsilon)N$, respectively, and therefore $(\varrho S-J_I)\xi\geqslant\varrho\epsilon N$. Similarly, for the case $\xi\in\mathcal{C}^-$ we have the two conditions $\xi G_I>\varrho(\theta-\delta)N$ and $\xi\varrho S<\varrho(\theta-\delta-\epsilon)N$, respectively, and therefore $-(\varrho S-J_I)\xi\geqslant\varrho\epsilon N$. We thus obtain

$$(\varrho S-J_I)(\pm\xi)\geqslant\varrho\epsilon N, \quad \text{for } \xi\in\mathcal{C}^{\pm}. \tag{B4}$$

Averaging this inequality over the ensemble $\mathcal{T}_t^{\alpha}$ yields $(\varrho S-\langle J_I\rangle)(\pm\xi)\geqslant\varrho\epsilon N$, depending on whether ξ is in class $\mathcal{C}^+$ or $\mathcal{C}^-$, respectively. This is correct because averaging is a convex operation. In particular, setting $y=\varrho S-J_I$ we have $\langle y\xi\rangle\geqslant\min(y\xi)\geqslant\varrho\epsilon N$. Let us abbreviate $x=\varrho S-\langle J_I\rangle$. Since $\langle\langle\zeta\rangle\rangle=q\xi$ and $\langle\langle\zeta_j\rangle\rangle\geqslant q$ for aN components we conclude that

$$\pm x\langle\langle\zeta\rangle\rangle\geqslant\varrho q\epsilon aN, \quad \text{for } \xi\in\mathcal{C}^{\pm}. \tag{B5}$$

Note that $\langle\langle.\rangle\rangle$ denotes the average over the whole ensemble of possible updates ζ at time t. We would like to replace $\langle\langle\zeta\rangle\rangle$ in the above estimate by $\langle\zeta\rangle$, where the latter average is taken over the ensemble $\mathcal{T}_{t+1}^{\alpha}$. Since $|x_j|\leqslant 1$ we can combine (B2) and (B5). For a typical update we therefore have with probability larger than $1-2\epsilon_o$;

$$\pm x\zeta\geqslant\varrho q\epsilon aN-\frac{q^2}{2}aN=qaN\left(\varrho\epsilon-\frac{q}{2}\right)\geqslant\frac{q\varrho}{2}\epsilon aN, \tag{B6}$$

provided $q\leqslant\varrho\epsilon$. Averaging over $\mathcal{T}_{t+1}^{\alpha}$ we obtain from (B6) that with probability larger than $1-2\epsilon_o$;

$$\pm x\langle\zeta\rangle\geqslant\frac{q\varrho}{2}\epsilon aN, \quad \text{for } \xi\in\mathcal{C}^{\pm}. \tag{B7}$$

To get an upper bound for ϵ_o we first note that $|\varrho S-J_I|\leqslant 1$, and conclude from (B4) that $a\geqslant\varrho\epsilon$. With the expression for ϵ_o given in (B1) we obtain $\epsilon_o\leqslant 4/(q^3\varrho\epsilon N)$.

4. Controlling the forgetting part

To control the effect of the forgetting term we write $\Delta F=-\langle\zeta * J_I\rangle=-\langle\zeta\rangle * \langle J_I\rangle-C$ with $C=\langle\zeta * J_I\rangle-\langle\zeta\rangle * \langle J_I\rangle$. Setting $x=\varrho S-\langle J_I\rangle$, inserting the expression for ΔF and applying the Cauchy-Schwartz inequality ($xy\leqslant\|x\|\,\|y\|$, with equality if and only if $x=y$) we get

$$\begin{aligned} x\Delta F &= \langle J_I\rangle(\langle\zeta\rangle * \langle J_I\rangle)-\varrho S(\langle\zeta\rangle * \langle J_I\rangle)-xC \\ &= (\sqrt{\langle\zeta\rangle} * \langle J_I\rangle)^2-\varrho(\sqrt{\langle\zeta\rangle} * S)(\sqrt{\langle\zeta\rangle} * \langle J_I\rangle)-xC \\ &\geqslant \|\sqrt{\langle\zeta\rangle} * \langle J_I\rangle\|(\|\sqrt{\langle\zeta\rangle} * \langle J_I\rangle\|-\varrho\|\sqrt{\langle\zeta\rangle} * S\|)-xC. \end{aligned} \tag{B8}$$

To estimate xC we write $xC=\langle u\zeta\rangle-v\langle\zeta\rangle$ with $u=x * J_I$ and $v=x * \langle J_I\rangle$. Note that $|u_j|\leqslant 1$ and $|v_j|\leqslant 1$. For a typical update we get from (B2) that $|v\zeta-v\langle\langle\zeta\rangle\rangle|\leqslant(q^2/2)aN$ with probability larger than $1-2\epsilon_o$. Averaging again over the ensemble $\mathcal{T}_{t+1}^{\alpha}$ gives with probability larger than $1-2\epsilon_o$:

$$|v\langle\zeta\rangle-v\langle\langle\zeta\rangle\rangle|\leqslant\frac{q^2}{2}aN. \tag{B9}$$

Similarly, we get from (B2) that $|u\zeta-u\langle\langle\zeta\rangle\rangle|\leqslant q^2/2aN$ with probability larger than $1-2\epsilon_o$. Since $\langle u\rangle=v$ we obtain by averaging $|\langle u\zeta\rangle-v\langle\langle\zeta\rangle\rangle|\leqslant(q^2/2)aN$. Combining this inequality with (B9) yields for a typical update with probability larger than $1-4\epsilon_o$:

$$|xC|=|\langle u\zeta\rangle-v\langle\zeta\rangle|\leqslant q^2aN. \tag{B10}$$

5. When forgetting supports learning

In the case of $\|\sqrt{\langle\zeta\rangle} * \langle J_I\rangle\|\geqslant\varrho\|\sqrt{\langle\zeta\rangle} * S\|$ one immediately gets from (B8) and (B10) that $x\Delta F\geqslant-q^2aN$. Since $\langle\Delta J\rangle=\Delta L+\Delta F$ we obtain together with (B7) that with probability larger than $1-6\epsilon_o$;

$$x\langle\Delta J\rangle\geqslant qaN\left(\frac{\varrho}{2}\epsilon\overline{g_I}-q\right)\geqslant\frac{q\varrho}{4}\epsilon\overline{g_I}aN, \tag{B11}$$

provided that $\|\sqrt{\langle\zeta\rangle} * \langle J_I\rangle\|\geqslant\varrho\|\sqrt{\langle\zeta\rangle} * S\|$ and $q\leqslant(\varrho/4)\epsilon\overline{g_I}$. Recall that $\overline{g_I}=\min\{g_I,1-g_I\}$.

6. When forgetting counteracts learning

Let us now assume that $\|\sqrt{\langle\zeta\rangle} * \langle J_I\rangle\|\leqslant\varrho\|\sqrt{\langle\zeta\rangle} * S\|$. Setting $y=S * S$ we conclude from (B2) that with probability larger than $1-2\epsilon_o$:

$$\|\sqrt{\langle\zeta\rangle} * S\|^2=y\langle\zeta\rangle\leqslant y\langle\langle\zeta\rangle\rangle+\frac{q^2}{2}aN\leqslant qaN(1+q/2).$$

From (B8) and (B10) we therefore get $x\Delta F\geqslant -\varrho^2\|\sqrt{\langle\zeta\rangle} * S\|^2-xC\geqslant-qaN(2\varrho^2+q)$. Since $\langle\Delta J\rangle=\Delta L+\Delta F$ we obtain together with (B7) that with probability larger than $1-4\epsilon_o$,

$$\begin{aligned} x\langle\Delta J\rangle &\geqslant qaN\left(\frac{\varrho}{2}\epsilon\overline{g_I}-(2\varrho^2+q)\right) \\ &= q\varrho aN\left(\frac{\epsilon}{4}\overline{g_I}-2\varrho\right)+qaN\left(\frac{\epsilon\varrho}{4}\overline{g_I}-q\right) \\ &\geqslant\frac{q\varrho}{8}\epsilon\overline{g_I}aN, \end{aligned} \tag{B12}$$

provided that $\|\sqrt{\langle\zeta\rangle} * \langle J_I\rangle\|\leqslant\varrho\|\sqrt{\langle\zeta\rangle} * S\|$, $\varrho\leqslant(\epsilon/16)\overline{g_I}$ and $q\leqslant(\epsilon\varrho/8)\overline{g_I}$.

7. Equivalent updates within a subclass

We next show that $\|\langle\Delta J\rangle\|^2$ is bounded by $q^{3/2}N$. Since $\Delta J=\zeta * (1-J)$ and $\Delta J=-\zeta * J$ for $\xi\in\mathcal{C}^{\pm}$, respectively, we have $\|\langle\Delta J\rangle\|^2\leqslant\|\langle\zeta\rangle\|^2$. We will show that the expectation value $\langle\zeta_j\rangle$ is in the order of q for most components j.

To give such an upper bound for $\langle\zeta_j\rangle$ across the trajectories in $\mathcal{T}^{\alpha}_{t+1}$ we exploit the synaptic redundancy according to which many different synaptic states J^{t+1} lead to the same total postsynaptic current $h^{t+1}=(1/N)J_I^{t+1}\xi^{t+1}$. Let us consider the case of a synaptic potentiation at time t; the case of a depression being treated similarly. The states of the trajectories $J\in\mathcal{T}^{\alpha}_{t+1}$ at time $t+1$ then have the form $J^{t+1}=J^t+\Delta J^t$ with $\Delta J^t=\zeta*(1-J^t)$. Let $\mathcal{Z}^{\alpha}$ be the set of update vectors ζ at time t corresponding to the trajectories in $\mathcal{T}^{\alpha}_{t+1}$. This is the ensemble over which the expectation value $\langle\zeta\rangle$ is taken. We decompose $\mathcal{Z}^{\alpha}$ into the subsets $\mathcal{Z}^{\alpha}_J$ of updates ζ starting from a fixed state J^t at time t. This set is further decomposed into sets of *equivalent* update vectors ζ giving the same total current h^{t+1}, $\mathcal{Z}^{\alpha}_J(l)=\{\zeta\in\mathcal{Z}^{\alpha}_J\,|\,\Delta J^t\xi^{t+1}=l\}$ for $l\in\mathbf{N}$. By applying twice the convexity of the averaging process we get $\langle\zeta_j\rangle\leqslant\max{}_J\langle\zeta_j\rangle_J\leqslant\max{}_{J,l}\langle\zeta_j\rangle_l$, where the different brackets $\langle.\rangle$, $\langle.\rangle_J$, and $\langle.\rangle_l=\langle.\rangle_{J,l}$ refer to the expectation values across $\mathcal{Z}^{\alpha}$, $\mathcal{Z}^{\alpha}_J$, and $\mathcal{Z}^{\alpha}_J(l)$, respectively.

To estimate $\langle\zeta_j\rangle_l$ we first note that ζ_j at time t does not affect the total postsynaptic current h^{t+1} if synapse j is already potentiated, $J_j^t=1$. For such components j the condition $\zeta\in\mathcal{Z}^{\alpha}_J(l)$ therefore does not represent any restriction on the value of ζ_j, and the average across $\mathcal{Z}^{\alpha}_J(l)$ is the same as the average across all realizations, $\langle\zeta_j\rangle_l=\langle\langle\zeta_j\rangle\rangle=q$, provided that $\mathcal{Z}^{\alpha}_J(l)\neq\emptyset$. If these components j are numerous we conclude that $\|\langle\zeta\rangle\|^2\leqslant q^{3/2}N$. If they are not numerous, we will use the redundancy argument to show that $\langle\zeta_j\rangle_l$ is small. This leads to the following case distinctions (1) and (2).

8. Equivalent updates provide redundancy

Set $z=(1-J^t)*\xi^{t+1}$ and let $\mathcal{I}_{1/0}$ be the set of components j with $z_j=1$ or $z_j=0$, respectively. Let $\tilde{a}N$ be the number of components in $\mathcal{I}_1$, i.e., the number of components which may have been potentiated at time t $(J_j^t=0)$, and which are activated by the pattern at time $t+1$ $(\xi_j^{t+1}=1)$. Only for these components will the state of the random variable ζ_j at time t have the chance to affect the total postsynaptic current h^{t+1}. We consider the two partially overlapping cases: (1), $\tilde{a}\leqslant q^{3/2}(1-q^{1/2})/(1-q^2)$, and (2), $\tilde{a}\geqslant q^2$.

(1) In the first case, when $\tilde{a}\leqslant q^{3/2}(1-q^{1/2})/(1-q^2)$, we have $\langle\zeta_j\rangle_l=\langle\langle\zeta_j\rangle\rangle=q$ for the numerous components $j\in\mathcal{I}_0$, as outlined above, and trivially $\langle\zeta_j\rangle_l\leqslant 1$ for $j\in\mathcal{I}_1$. We therefore obtain

$$\|\langle\Delta J^t\rangle\|^2\leqslant\|\langle\zeta\rangle_t\|^2\leqslant q^2(1-\tilde{a})N+\tilde{a}N\leqslant q^{2/3}N,$$

$$\text{provided that } \tilde{a}\leqslant q^{3/2}\frac{1-q^{1/2}}{1-q^2}. \tag{B13}$$

(2) In the second case, when $\tilde{a}\geqslant q^2$, we again engage the typicality of an update. Since $\mathrm{var}(z\zeta)\leqslant q\tilde{a}N$ we have according to Chebyshev's inequality, analogously to (B2)

$$\mathcal{P}\{|z\zeta-z\langle\langle\zeta\rangle\rangle|\leqslant q^2\tilde{a}N\}\geqslant 1-\frac{\mathrm{var}(z\zeta)}{(q^2\tilde{a}N)^2}\geqslant 1-\tilde{\epsilon}_o,$$

$$\text{with } \tilde{\epsilon}_o=\frac{1}{q^2\tilde{a}N}\leqslant\frac{1}{q^5N}. \tag{B14}$$

The probability that an update is typical in the sense of (B1), and that it satisfies (B14), is larger than $1-\epsilon_o-\tilde{\epsilon}_o$. By definition of z we have $z\langle\langle\zeta\rangle\rangle\leqslant q\tilde{a}N$, and formula (B14) states that with this high probability

$$\Delta J^t\xi^{t+1}=z\zeta=((1-J^t)*\xi^{t+1})\zeta\leqslant q\tilde{a}N(1+q)\leqslant 2q\tilde{a}N. \tag{B15}$$

By definition of $\mathcal{Z}^{\alpha}_J(l)$ we have $z\zeta=l$ for $\zeta\in\mathcal{Z}^{\alpha}_J(l)$, and we conclude from (B15) that with the same high probability $l\leqslant 2q\tilde{a}N$ if the set $\mathcal{Z}^{\alpha}_J(l)$ is non-empty. In this case we can write $\mathcal{Z}^{\alpha}_J(l)=\{\zeta\in\{0,1\}^N\,|\,\zeta z=l\}$ and drop the constraint of $\zeta\in\mathcal{Z}^{\alpha}_J$. This is because $\mathcal{Z}^{\alpha}_J(l)$ is either empty, or the condition $\zeta z=l$ already implies that $\zeta\in\mathcal{Z}^{\alpha}_J$. Hence, $\mathcal{Z}^{\alpha}_J(l)$ consists of all binary vectors ζ which have exactly $l\leqslant 2q\tilde{a}N$ 1's among the $\tilde{a}N$ nonvanishing components of z. Since the individual variables ζ_j are stochastically independent, the relative frequency of $\zeta_j=1$ for a fixed j across the set $\mathcal{Z}^{\alpha}_J(l)$ is therefore $l/(\tilde{a}N)$. This is the desired redundancy according to which different synaptic update vectors ζ lead to the same total postsynaptic current h^{t+1}. Since $a>0$ we conclude that with probability larger than $1-\epsilon_o-\tilde{\epsilon}_o$ we have for $j\in\mathcal{I}_1$:

$$\langle\zeta_j\rangle_l=\frac{l}{\tilde{a}N}\leqslant\frac{2q\tilde{a}N}{\tilde{a}N}=2q. \tag{B16}$$

Since for $j\in\mathcal{I}_0$ the value of ζ_j does not affect h^{t+1} we again have $\langle\zeta_j\rangle_l=\langle\langle\zeta_j\rangle\rangle=q$ for these components as explained above. Together with (B16) we get that $\langle\zeta_j\rangle\leqslant\max{}_l\langle\zeta_j\rangle_l\leqslant 2q$ for any component j, and therefore $\|\langle\Delta J^t\rangle\|^2\leqslant\|\langle\zeta\rangle_l\|^2\leqslant 4q^2N$. If we assume that $q\leqslant 1/16$ we obtain together with (B13) that, independent of a, we have with probability larger than $1-\epsilon_o-\tilde{\epsilon}_o$:

$$\|\langle\Delta J^t\rangle\|^2\leqslant\|\langle\zeta\rangle_l\|^2\leqslant q^{3/2}N. \tag{B17}$$

9. Learning in the general case stops

We finally show that the distance from $\langle J_I\rangle$ to ϱS decreases with each synaptic update at least by some fixed quantity. Let t_μ denote the time(s) when pattern ξ^μ is presented and the synapses are updated $(\alpha^{t_\mu}=1)$. At a subsequent time step $t_\mu+1$ there is $\langle J_I^{t_\mu+1}\rangle=\langle J_I^{t_\mu}\rangle+\langle\Delta J^{t_\mu}\rangle$. Recalling the abbreviation $x^{t_\mu}=\varrho S-\langle J_I^{t_\mu}\rangle$ and combining (B11) and (B12) we estimate $x^{t_\mu}\langle\Delta J^{t_\mu}\rangle\geqslant q\varrho\epsilon\overline{g_I}aN/8$ with probability larger than $1-6\epsilon_o$. With (B17) we obtain with probability larger than $1-7\epsilon_o-\tilde{\epsilon}_o$;

$$\begin{aligned}\|x^{t_\mu+1}\|^2-\|x^{t_\mu}\|^2&=-2x^{t_\mu}\langle\Delta J^{t_\mu}\rangle+\|\langle\Delta J^{t_\mu}\rangle\|^2\\&\leqslant qN(q^{1/2}-\varrho\epsilon\overline{g_I}a/4)\leqslant-q\varrho\epsilon\overline{g_I}aN/8,\end{aligned} \tag{B18}$$

provided that $q^{1/2}\leqslant\varrho\epsilon\overline{g_I}a/8$. Since we may safely assume that the coding level of a pattern is larger than the scaled separation margin parameter, $a\geqslant\varrho\epsilon$, this and the previous conditions on q are satisfied if $q\leqslant[(\varrho\epsilon)^2\overline{g_I}/8]^2$. Recall that we also require $\varrho\leqslant\epsilon\overline{g_I}/16$.

We next sum up the contributions of all the updates up to time t evoked by the different patterns, $\langle J_I^t \rangle = J_I^0 + \Sigma_{t'_\mu < t} \langle \Delta J^{t'_\mu} \rangle$. Applying iteratively estimate (B18), and using again that $a \geqslant \varrho \epsilon$, we obtain that with probability larger than $1 - n_t(7\epsilon_o + \tilde{\epsilon}_o)$:

$$0 \leqslant \|x^{t_\mu}\|^2 \leqslant \|x^0\|^2 - n_t q (\varrho \epsilon)^2 \overline{g_I} N/8, \tag{B19}$$

where n_t is the number of synaptic updates up to the tth presentation of a pattern. Since $\|x^0\| = \|\varrho S - J_I^0\|^2 \leqslant N$ we get from (B19) that with high probability we would have $\|x^{t_\mu}\|^2 < 0$ after $n_t > n_o = 8/(q\overline{g_I}(\varrho\epsilon)^2)$ updates. Since this is not possible we conclude that with high probability there cannot be more than n_o synaptic updates, i.e., $\alpha^t = 0$ for all t larger than $t_o = pn_o$, provided $\mathcal{T}_t^\alpha$ is not empty. The estimate $t_o = pn_o$ holds because within each cycle there is at least one of the totally p patterns which leads to a synaptic update—otherwise the patterns would have been already learned. Hence, the trajectories become stationary with high probability after at least t_o time steps. Since $\epsilon_o \leqslant 4/(q^3 \varrho \epsilon N)$ and $\tilde{\epsilon}_o \leqslant 1/(q^5 N)$ the probability of not converging after n_o updates (t_o time steps) is in the order of $O(1/N)$.

[1] C. Petersen, R. Malenka, A. Roger, A. Nicoll, and H. JJ, PNAS 95, 4732 (1998).

[2] G. Parisi, J. Phys. A **19**, L (1986).

[3] D. J. Amit and S. Fusi, **3**, 443 (1992).

[4] D. J. Amit and S. Fusi, Neural Comput. **6**, 957 (1994).

[5] S. Fusi, Biol. Cybern. **87**, 459 (2002).

[6] M. Tsodyks, Mod. Phys. Lett. B **4**, 713 (1990).

[7] E. Amaldi, in *Artificial Neural Networks*, edited by T. Kohonen (North-Holland, Amsterdam, 1991), Vol. 1, pp. 55–60.

[8] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, New York, 1999).

[9] W. Krauth and M. Mezard, J. Phys. (France) **50**, 3057 (1989).

[10] T. M. Cover, IEEE Trans. Electron. Comput. **14**, 326 (1965).

[11] E. Gardner, Europhys. Lett. **4**, 481 (1987).

[12] Y. Amit and M. Mascaro, Neural Comput. **13**, 1415 (2001).

[13] C. Diorio, P. Hasler, B. Minch, and C. Mead, IEEE Trans. Electron Devices **43**, 1972 (1996).

[14] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, and D. J. Amit, Neural Comput. **12**, 2227 (2000).

[15] E. Chicca and S. Fusi, in *World Congress on Neuroinformatics*, edited by F. Rattay (ARGESIM/ASIM, Vienna, 2001), pp. 468–477.

[16] G. Indiveri, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2002), Vol. 15, pp. 1091–1098.

[17] S. Venkatesh, J. Comput. Syst. Sci. **46**, 198 (1993).

[18] H. Köhler, S. Diederich, W. Kinzel, and M. Opper, Z. Phys. B: Condens. Matter **78**, 333 (1990).

[19] W. Senn and S. Fusi, Neurocomputing **58–60**, 321 (2004).

[20] W. Senn and S. Fusi, Neural Comput. (to be published).

[21] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).

[22] D. J. Amit and S. Fusi, Neural Comput. **6**, 957 (1994).

[23] G. Turrigiano and S. Nelson, Nat. Rev. Neurosci. **5**, 97 (2004).

[24] P. Del Giudice, S. Fusi, and M. Mattia, J. Physiol. (Paris) **97**, 659 (2003).

[25] D. J. Amit and G. Mongillo, Neural Comput. **15**, 565 (2003).

[26] S. Fusi, Rev. Neurosci. **14**, 73 (2003).