Available online at www.sciencedirect.com
SCIENCE DIRECT®

Neurocomputing 58–60 (2004) 321 – 326

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Slow stochastic learning with global inhibition: a biological solution to the binary perceptron problem

Walter Senn, Stefano Fusi*

Institute of Physiology, University of Bern, Bühlplatz 5, Bern 3012, Switzerland

Abstract

Networks of neurons connected by plastic all-or-none synapses tend to quickly forget previously acquired information when new patterns are learned. This problem could be solved for random uncorrelated patterns by randomly selecting a small fraction of synapses to be modified upon each stimulus presentation (slow stochastic learning). Here we show that more complex, but still linearly separable patterns, can be learned by networks with binary excitatory synapses in a finite number of presentations provided that: (1) there is non-vanishing global inhibition, (2) the binary synapses are changed with small enough probability (slow learning) only when the output neuron does not give the desired response (as in the classical perceptron rule) and (3) the neuronal threshold separating the total synaptic inputs corresponding to different classes is small enough.

Keywords: Synaptic plasticity; Learning; Memory: Classification

1. 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 [4]). These constraints have dramatic effects on networks performing as classifiers or as an associative memory. Networks of neurons connected by bounded synapses which cannot be changed by an arbitrarily small amount share the *palimpsest* property (see e.g. [2]): new patterns overwrite the oldest

* Corresponding author.
E-mail addresses: wsenn@cns.unibe.ch (W. Senn), fusi@cns.unibe.ch (S. Fusi).

0925-2312/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.01.062

ones, and only a limited number of patterns can be remembered. The more synapses changed on each stimulus presentation, the faster is forgetting. Moreover, learning to separate two classes of patterns with discrete synaptic weights is a combinatorially hard problem (the 'binary perceptron problem', see [1]). Fast forgetting can be avoided by changing only a small fraction of synapses, chosen randomly at each presentation. Stochastic selection permits the classification and memorization of an extensive number of random patterns, even if the number of synaptic states is reduced to two [2]. However, additional mechanisms must be introduced to store more realistic patterns with correlated components. The solution we study here is based on the perceptron learning rule: the synapses are changed with some probability only when the response of the post-synaptic cell is not the desired one. This 'stop-learning' property might be the expression of some regulatory synaptic mechanisms or the effect of a reward signal. Together with global inhibition, a small synaptic transition probability and a small neuronal threshold are sufficient to learn and memorize any linearly separable set of patterns.

2. The model

Neuron model: We consider a single postsynaptic neuron which receives excitatory inputs from N presynaptic neurons, and an inhibitory input which is proportional to the total activity of the N excitatory neurons. The postsynaptic neuron is either active or inactive, depending on whether the total postsynaptic current h is above or below a threshold θ_0. The total current is calculated by the weighted sum of the synaptic inputs ξ_j, $h = 1/N \sum_{j=1}^{N}(J_j - g_{\mathrm{I}})\xi_j$, where ξ_j can take on any value from (and including) 0 to R. The excitatory weights J_j are binary, either 0 or 1, and the common inhibitory weight g_{I} is set to an analog value in between the excitatory weights, $g_{\mathrm{I}} \in (0, 1)$.

Synaptic dynamics: During training the network is repeatedly presented with all the p patterns ξ of two classes C^+ and C^-. At each presentation, the activities ξ_j are clamped to the N presynaptic neurons, and the output of the postsynaptic neuron is clamped to the desired response ($\xi_{\mathrm{post}} = 0$ or 1, depending on whether ξ belongs to class C^+ or C^-, respectively). The synaptic learning rule is designed such that, after successful training, the total synaptic current h generated by a pattern ξ falls either above or below the threshold θ_0, depending on whether ξ is in class C^+ or C^-.

Upon presentation of a pattern ξ the binary synapses flip stochastically, depending on the pre- and postsynaptic activities and the total current h. Downregulated synapses ($J_j = 0$) are potentiated with probability $q_+\xi_j$ when the pre- and postsynaptic cells are both active (ξ_{post}, $\xi_j > 0$) and the total synaptic current not too large ($h \leqslant \theta_0 + \delta_0$, with a learning margin $\delta_0 \geqslant 0$). Potentiated synapses ($J_j = 1$) downregulate with probability $q_-\xi_j$ when the presynaptic neuron is active ($\xi_j > 0$), the postsynaptic cell inactive ($\xi_{\mathrm{post}} = 0$), and the total synaptic input not too low ($h \geqslant \theta_0 - \delta_0$). The factors q_+ and q_- control the learning and forgetting rates of the network.

This rule can be formally summarized by introducing a random variable $\zeta_j^{\pm}$ which, when presenting pattern ξ^t at time t, is 1 with probability $q_\pm \xi_j^t$ and 0 otherwise. If the Hebbian conditions for a synaptic potentiation are met, the synaptic state changes probabilistically according to $J_j^{t+1} = J_j^t + \zeta_j^+ (1 - J_j^t)$. If the conditions for downregulation are met, then $J_j^{t+1} = J_j^t - \zeta_j^- J_j^t$. To approximate this dynamics we consider the probabilities G_j^t that synapse j is potentiated at time t. These probabilities represent the expected synaptic weights, $G_j^t = \langle J_j^t \rangle$, and their dynamics is governed by

$$G_j^{t+1} = \begin{cases} G_j^t + q_+ \xi_j^t (1 - G_j^t), & \text{if } \xi_{\text{post}}^t > 0, \quad \xi_j^t > 0, \text{ and } h^t \leqslant \theta_0 + \delta_0, \\ G_j^t - q_- \xi_j^t G_j^t, & \text{if } \xi_{\text{post}}^t = 0, \quad \xi_j^t > 0, \text{ and } h^t \geqslant \theta_0 - \delta_0. \end{cases} \tag{1}$$

Note that, since the fluctuations of h^t for different realizations of the stochastic process ζ shrink to zero with growing N (see Fig. 2c below), the expected total current $\langle h^t \rangle$, obtained by replacing the J_j^t's with the G_j^t's, does well approximate the actual total current h^t.

3. Results

Given any two sets $C^\pm$ of linearly separable patterns, a neuron endowed with global inhibition and the stochastic learning rule described above will always learn to correctly classify the patterns in a finite number of presentations. The tighter the separation between the two classes $C^\pm$, the smaller the neuronal threshold θ_0, the learning margin δ_0, 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 (Fig. 1a). Writing this separation margin as $\delta + \varepsilon$ we have $\xi S > (\theta + \delta + \varepsilon)N$ for $\xi \in C^+$, and $\xi S < (\theta - \delta - \varepsilon)N$ for $\xi \in C^-$. Classification is then also possible by a separation vector which is scaled by a factor ϱ, provided that also the threshold and the margins are scaled by the same factor. These different solutions correspond to output neurons which would separate the patterns around different thresholds at the end of the training session (e.g. $h > \varrho\theta + \varrho\delta$ for $\xi \in C^+$). However, as we show, the synaptic dynamics can only converge to the separation vector corresponding to a small enough scaling factor, $\varrho \leqslant \varepsilon \bar{g}_{\text{I}}/(2R)$, where ϱ depends on the partial separation margin ε, the 'distance' $\bar{g}_{\text{I}} = \min\{g_{\text{I}}, 1 - g_{\text{I}}\}$ of the global inhibitory weight g_{I} from the weight boundaries 0 and 1, and the maximal input strength R (Fig. 1a). Given such a scaling factor and any global inhibition g_{I} between 0 and 1, the synaptic dynamics (1) converges (i.e. all the patterns will be classified correctly) in at most $n_0 = 6/(q\varrho\varepsilon\bar{g}_{\text{I}})$ synaptic updates, provided that the learning rate is small enough, $q \leqslant \varrho\varepsilon\bar{g}_{\text{I}}/(2R^2)$. This is valid for any presentation order of the patterns to be learned and for any initial conditions for the synaptic states. A similar upper bound on the number of synaptic updates for the stochastic dynamics holds with probability $1 - O(1/N)$. Importantly, the separation margin $\delta + \varepsilon$ is assumed to be independent of N. As N increases while p remains fixed, the constant margin implies some redundancy in the coding (e.g. many neurons encode the same stimulus activity).

Proof sketch: The idea behind the threshold scaling and the global inhibition is to keep the expected synaptic strength $G^t = \langle J^t \rangle$ away from the lower and upper boundaries of the synaptic efficacies. This prevents the weight vector G^t from being distorted by

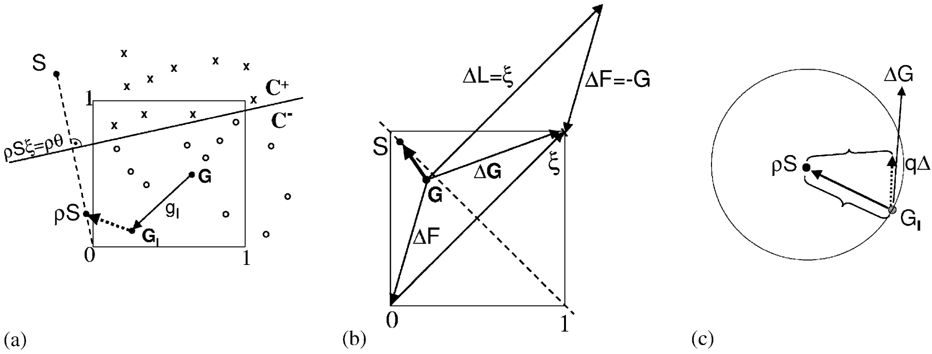

Fig. 1. Sketch of the proof: (a) Sets C^+ (crosses) and C^- (circles) of patterns ξ are assumed to be linearly separable, with a separation vector S and a threshold θ. Since S may contain negative components and components larger than 1, it cannot in general be approximated by the vector G of the synaptic potentiation probabilities. 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, $\varrho S \approx G_{\mathrm{I}} = G - g_{\mathrm{I}}\mathbf{1}$, with a weight vector G laying within the unit hypercube, far from saturation at 0 and 1. (b) Without global inhibition ($g_{\mathrm{I}} = 0$ and $G_{\mathrm{I}} = G$), synaptic saturation (ΔF) may prevent the weight vector G to be updated in the 'correct' direction ΔL, in the sense that $(\varrho S - G_{\mathrm{I}})\Delta G > 0$. In the shown example we have $(\varrho S - G_{\mathrm{I}})\Delta G < 0$, i.e. the update moves G_{I} away from the solution vector ϱS (where $\rho = 1$ in panel (b)). This is because an update of G_{I} in the desired direction ΔL is distorted by the nearby boundaries and, instead, G_{I} moves in the direction of $\Delta G = \Delta L + \Delta F$ towards the upper right corner. Such a distortion is not possible if G is close to the main diagonal and far from $\mathbf{0}$ and $\mathbf{1}$ (achieved by a small ϱ and g_{I} in between 0 and 1, see (a)). (c) A positive scalar product $(\varrho S - G_{\mathrm{I}})\Delta G > 0$ ensures that the G_{I} moves towards ϱS, provided that the learning rate q is small (distance indicated by the upper brace is smaller than that indicated by the lower brace).

synaptic saturation. The expected synaptic change ΔG^t, where $G^{t+1} = G^t + q\Delta G^t$, can be decomposed into a 'linear' and a 'forgetting' part. If the updating condition in (1) is met we have:

$$\Delta G = \Delta L + \Delta F = \begin{cases} \xi * (\mathbf{1} - G) = (1 - g_{\mathrm{I}})\xi - \xi * G_{\mathrm{I}} & \text{if } \xi \in C^+, \\ -\xi * G = -g_{\mathrm{I}}\xi - \xi * G_{\mathrm{I}} & \text{if } \xi \in C^-, \end{cases} \tag{2}$$

where $G_{\mathrm{I}} = G - g_{\mathrm{I}}\mathbf{1}$ and '$*$' is the componentwise product of vectors. The linear term $\Delta L = (1 - g_{\mathrm{I}})\xi$ in case of $\xi \in C^+$ and $\Delta L = -g_{\mathrm{I}}\xi$ in case of $\xi \in C^-$, respectively, is the learning component which is parallel to the pattern to be learned (Fig. 1b). 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 $\xi \in C^+$, for instance, we have $\xi\varrho S > \varrho(\theta + \delta + \varepsilon)N$ by assumption of the separability, and if the update condition on h is met, we have $\xi G_{\mathrm{I}} < \varrho(\theta + \delta)N$ as required in (1). By subtracting the second from the first inequality we get

$$(\varrho S - G_{\mathrm{I}})\Delta L \geqslant \varrho\varepsilon\bar{g}_{\mathrm{I}}N, \tag{3}$$

where $\bar{g}_{\mathrm{I}}$ is defined above. The same estimate is true for $\xi \in C^-$. When the forgetting part is negligible, we would have $\Delta G \approx \Delta L$, and (3) would ensure that G_{I}^t moves toward ϱS, provided that q is small (Fig. 1c). In fact, if the angle between $(\rho S - G_{\mathrm{I}})$

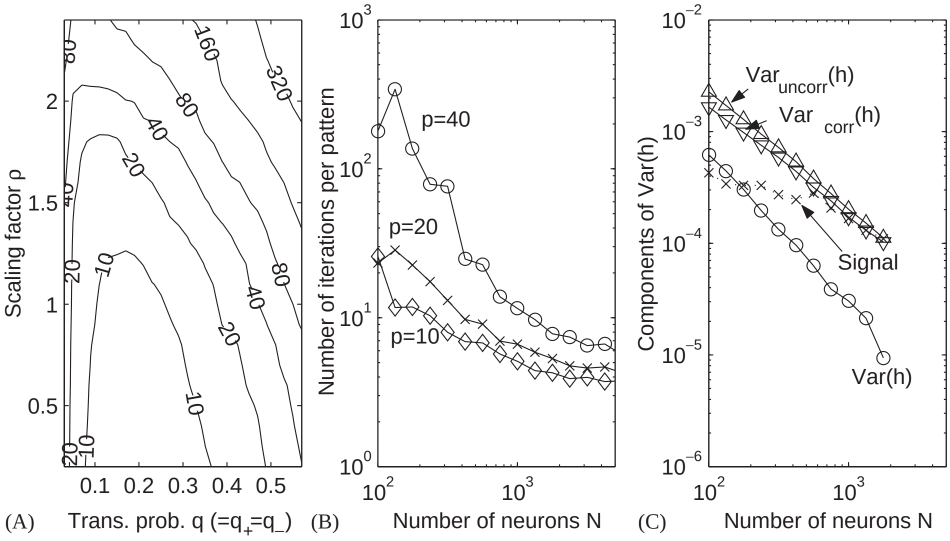

Fig. 2. Simulation results: (A) number of iterations per pattern as a function of the learning rate q and the scaling factor ϱ; (B) number of iterations per pattern as a function of the number of neurons N for $p = 10, 20, 40$ ($q = 0.05$, $f = \frac{1}{4}$, $\theta = 0.01$); (C) correlated (down triangles) and uncorrelated components (up triangles) of the variance of h as a function of N ($p = 40$) in a double log scale. The signal (crosses) expresses the square of the average distance between the h produced by the two different classes.

and ΔG is smaller than 90°, the weight vector at the next time step, $G_\mathrm{I} + q\Delta G$, is always closer to ρS than G_I was, provided q is small enough.

The forgetting part $\Delta F = -\xi G_\mathrm{I}$ in (2) is due to the non-linearities of synaptic saturation and tends to bring G_I towards 0, where $G_j = g_\mathrm{I}$ and no synaptic structure would be present. Hence it might neutralize or even counteract learning as explained in Fig. 1b. However, this negative effect is strongly reduced and can become negligible if the weight vector is close to the main diagonal, i.e. if the expectation values of all the synaptic strengths are roughly equal. It is possible to show that $(\varrho S - G_\mathrm{I})\Delta F \geqslant -\varrho^2 RN$. Hence, provided that the scaling factor ϱ is small, convergence of the learning procedure is guaranteed.

Simulations: We trained our binary perceptron with p random uncorrelated binary patterns. Fig. 2A shows the number of iterations per pattern needed to converge to a solution as a function of the scaling factor ϱ and the learning rates (transition probabilities) $q = q_+ = q_-$ (with $p = 10$ and $N = 100$). As learning becomes too fast, or ϱ too large, the number of required iterations grows very quickly and, eventually, learning will not converge. If learning is too slow the number of iterations scales as $1/q$, but the convergence is always guaranteed. Similarly, the number of required iterations grows as the global inhibition g_I becomes close to 0 or 1 (not shown). Fig. 2B shows the number of iterations per pattern for $p = 10, 20, 40$ random uncorrelated binary patterns (probability of an active neuron: $f = \frac{1}{4}$) as a function of the number of neurons N of the input layer. 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 explained by the variance of h^t, e.g. at time $t = 10$, obtained by presenting 1000 times the same sequence of patterns with different realizations of ζ (Fig. 2C). The uncorrelated component of the variance (the variance that one would have if different synapses were statistically independent, i.e. without the global stopping condition) scales as $1/N$, and the correlated component (the remaining part) scales in the same way and is always negative.

4. Conclusions

We have shown that stochastic learning allows a perceptron with binary excitatory weights to converge in a finite number of updates for any separable set of patterns, provided that there is some global inhibition, a small neuronal threshold, and slow learning. These ingredients rescue binary synapses from fast forgetting due to saturation of the potentiation probabilities. They also allow to store as many patterns as in a network with analogue unbounded synapses (proportional to N^α, with α from 1 to 2, depending on the coding level f of the patterns, instead of $\log N$, as obtained without these ingredients, see [2]). The considered stochastic selection mechanism can be implemented in terms of a detailed spike-driven synaptic dynamics by exploiting the irregularity of the spike trains [3]. Indeed, the same mean spike frequencies are realized by a large number of different spike train realizations: some of them might induce long lasting modifications while others would simply leave the synapse unchanged. In this case, the stochastic selection mechanism is a property of the network, and not of the single synapse.

Acknowledgements

This work was supported by the EU Grant IST-2001-38099 ALAVLSI and the SNF Grant 3152-065234.01. We thank J. Brader for useful remarks.

References

[1] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, New York, 1999.

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

[3] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, D.J. Amit, Spike-driven syn. plasticity: theory, simulation, VLSI impl, Neural Comput. 12 (2000) 2227–2258.

[4] C.C.H. Petersen, R.C. Malenka, R.A. Nicoll, J.J. Hopfield, All-or-none potentiation at CA3–CA1 synapses, Proc. Natl. Acad. Sci. 95 (1998) 4732.