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

Walter Senn, Stefano Fusi [1]

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

Abstract

Networks of neurons connected by plastic all-or-none synapses tend to forget quickly 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 linearly separable patterns can be learned by networks with binary excitatory synapses in a finite number of presentations provided that, given any global non-zero inhibition: (1) the binary synapses are changed with some small enough probability (slow learning) only when the output neuron does not give the desired response (as in the classical perceptron rule) (2) the neuronal threshold separating the total synaptic inputs corresponding to different classes is small enough.

Key words: synaptic plasticity, learning

1 Introduction

Activity dependent synaptic plasticity is known to play an important role in learning and in the last decade many experimentalists studied successfully the protocols to induce long term modifications. Much less is known about the maintenance of the synaptic efficacy, for instance whether the synapse can preserve a continuous or a discrete set of efficacies, or whether the change in the efficacy is itself continuous or discrete. Given the accumulating evidence that some synapses can preserve only a very limited number of states (some seem to have only two [1]), it is natural to question about the implications for a network with binary synapses performing as an associative memory or as a classifier.

Networks of neurons connected by realistic synapses (i.e. with bounded efficacies which do not permit infinitesimally small changes) share the *palimpsest* property (see e.g. [2]): new patterns overwrite the oldest ones, and only a limited number of patterns can be remembered. Moreover, if each stimulus changes a large number of synapses, both learning and forgetting rates are fast. The oblivion

[1] *E-mail addresses:* {wsenn,fusi}@cns.unibe.ch

provoked by fast forgetting can be avoided by changing only a small fraction of synapses, chosen randomly at each presentation. Stochastic selection permits storage of an extensive number of random uncorrelated patterns even if the number of synaptic states is reduced to two. However, additional mechanisms must be introduced to store patterns with more realistic and complex statistics. 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' condition might be the expression of some regulatory synaptic mechanism and it turned out to be sufficient to memorize linearly separable patterns.

2 The model

2.1 The neuronal dynamics

We consider a single postsynaptic neuron which receives excitatory inputs from N neurons and an inhibitory input which is proportional to the total activity of the N excitatory cells. The output neuron is either active or inactive depending on whether the total synaptic current is above or below a threshold θ. The total synaptic input h is the weighted sum of the synaptic inputs ξ_j, which are assumed to be either 0 (inactive input) or analog variables in the range between ν_0 and ν_{max}: $h = \frac{1}{N}\sum_{j=1}^{N}(J_j - g_I)\xi_j$, where $g_I \in (0,1)$ is an arbitrary, but global inhibitory synaptic weight. The excitatory weights J_j are binary.

2.2 The synaptic dynamics

The binary synapses flip randomly upon each presentation and modify their internal state to encode the pre- and the postsynaptic activity depending on whether the appropriate conditions on the total synaptic input are met. Depressed synapses ($J_j = 0$) are potentiated with probability $q_+\xi_j$, when the pre and the post-synaptic cells are both active and the total synaptic input is not too large ($h \leq \theta_\circ + \delta_\circ$, where $\delta_\circ \geq 0$ is a learning margin). Potentiated synapses ($J_j = 1$) are depressed with probability $q_-\xi_j$ when the pre-synaptic neuron is active, the postsynaptic cell is inactive and the total synaptic input is not too low ($h \geq \theta_\circ - \delta_\circ$). The factors q_+ and q_- control the learning and forgetting rates of the network. This rule can be formally summarized by introducing a binary random variable $\zeta_j^{\pm}$ which is 1 with probability $q_\pm\xi_j$ and 0 with probability $1 - q_\pm\xi_j$:

$$J_j^{t+1} = \begin{cases} J_j^t + \zeta_j^+ \,(1 - J_j^t)\,, & \text{if post. active, } \xi_j^t > \nu_\circ, \text{ and } h^t \leq \theta_\circ + \delta_\circ\,, \\ J_j^t - \zeta_j^- \, J_j^t\,, & \text{if post. inactive, } \xi_j^t > \nu_\circ, \text{ and } h^t \geq \theta_\circ - \delta_\circ\,. \end{cases} \tag{1}$$

The perceptron's learning dynamics can be approximately described in terms of the probability G_j^t that synapse j is potentiated. The same description applies to continuous bounded synapses. Following each stimulation, the probabilities G are updated according to:

$$G_j^{t+1} = \begin{cases} G_j^t + q_+ \, \xi_j^t(1 - G_j^t)\,, & \text{if post. active, } \xi_j^t > \nu_\circ, \text{ and } h^t \leq \theta_\circ + \delta_\circ\,, \\ G_j^t - q_- \, \xi_j^t G_j^t\,, & \text{if post. inactive, } \xi_j^t > \nu_\circ, \text{ and } h^t \geq \theta_\circ - \delta_\circ\,. \end{cases} \tag{2}$$

Since it can be proven that the correlations between the synaptic states J_j^t and J_k^t are bounded by a quantity which scales as $q = q_+ = q_-$, the dynamics of G^t is a good approximation for the expected dynamics of J^t in the limit of large N and small q, $G^t \approx \langle J^t \rangle$.

2.3 The learning scenario

During training the network is repeatedly presented with all the p patterns which are to be learned. At each presentation, the pattern of activity is imposed to the N neurons and the output is clamped to the desired response. The patterns belong to class C^+ when the output neuron should be active and to class C^- otherwise.

3 Results

Any sets $C^\pm$ of linearly separable patterns can always be learned in a finite number of presentations by a perceptron with binary excitatory synapses and global inhibition provided that the threshold $\theta_\circ$ separating the two classes is small and that learning is slow (small transition probabilities). More precisely, we assume that there is a separation vector S (not necessarily binary and positive) with $\|S\| = N$ and a threshold θ such that $\xi S > (\theta + \delta + \epsilon)N$ for $\xi \in C^+$, and $\xi S < (\theta - \delta - \epsilon)N$ for $\xi \in C^-$, with $\delta + \epsilon > 0$. We consider an output neuron with a scaled threshold $\theta_\circ = \varrho\theta$, a learning margin $\delta_\circ = \varrho\delta$, and any global inhibition g_I between 0 and 1. Then, the synaptic dynamics (2) converges in at most $n_\circ = 6/(q\varrho\epsilon\overline{g}_I)$ synaptic updates, provided that the learning rate and the scaling factor are small, $q \leq \varrho\epsilon\overline{g}_I/(2\nu_{max}^2)$ and $\varrho \leq \epsilon\overline{g}_I/(2\nu_{max})$, where $\overline{g}_I = \min\{g_I, 1 - g_I\}$. This is valid for any presentation order of the patterns to be learned and for any initial condition.

An upper bound for the number of presentations, t, until learning stops is $t = pn_\circ = 6p/(q\varrho\epsilon\overline{g}_I)$. A similar upper bound on the number of synaptic updates for the stochastic dynamics (1) holds with probability $1 - O(1/N)$. Importantly, the separation margin $\delta + \epsilon$ is assumed to be independent of N. This represents a form of redundancy in the coding of the p patterns with N neurons. A trivial way to meet the requirement $\delta + \epsilon =$ constant while N is growing is to duplicate each component ξ_j of an initial pattern ξ. However this is usually unnecessary.

3.1 The sketch of the proof

The idea behind threshold scaling and global inhibition is to keep the expected synaptic strength $\langle J^t \rangle \approx G^t$ away from the lower and upper boundaries of the synaptic efficacies in order to prevent the weight vector G^t from being distorted by 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 (2) is met we have:

$$\Delta G = \Delta L + \Delta F = \begin{cases} \xi * (\mathbf{1} - G) = (1 - g_I)\xi - \xi * G_I \ , & \text{if } \xi \in C_+ \ , \\ -\xi * G = -g_I\xi - \xi * G_I \ , & \text{if } \xi \in C_- \ , \end{cases} \tag{3}$$

where $G_I = G - g_I\mathbf{1}$ and '$*$' is the componentwise product of vectors.

The linear term $\Delta L = (1 - g_I)\xi$ and $\Delta L = -g_I\xi$ (for $\xi \in C^{\pm}$, respectively) is the learning component which is parallel to the pattern to be learned (Fig. 1a), and it is the only term which is present in the case of the classical perceptron learning with analog unbounded synapses. This component always brings G^t towards a solution vector. In fact, if $\xi \in C^+$ we have $\xi\varrho S > \varrho(\theta + \delta + \epsilon)N$, and if the update condition on h is met, we have $\xi\overline{G} < \varrho(\theta + \delta)N$. Hence,

$$(\varrho S - G_I)\Delta L \geq \varrho\epsilon\overline{g}_I N \,, \tag{4}$$

for the case of $\xi \in C^+$, but the same estimate also holds for $\xi \in C^-$ with the corresponding definition of ΔL. Were the forgetting part negligible, we would have $\Delta G \approx \Delta L$, and (4) would ensure that G_I^t moves towards ϱS, provided that the learning rate q is small. In fact, if the angle between $\rho S - G_I$ and ΔG is smaller than 90°, the weight vector $G_I + q\Delta G$ at the next time step is always closer to ρS than G_I for a small enough value of q (Fig. 1c).

The forgetting part $\Delta F = -\xi G_I$ is due to the non-linearities of synaptic saturation and tends to bring G towards the oblivion configuration (all synapses with uniform expectation values, equal to g_I). 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 $(\varrho S - G_I)\Delta F \geq -\varrho^2\nu_{max}N$. Hence, provided that the scaling factor ϱ is small, convergence of the learning procedure is guaranteed as outlined above.

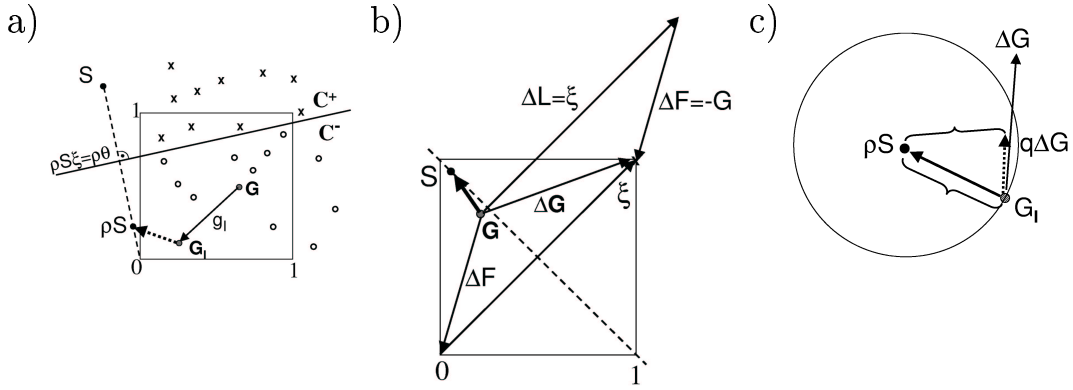

Fig. 1. Sketch of the prove. **a)** The set of patterns C^+ and C^- are assumed to be linearly separable, with a separation vector S and a threshold θ. Since S may contain negative components and components larger than 1, it cannot in general be approximated by the vector G of the synaptic potentiation probabilities. Only if the solution vector S (and with it the threshold θ) is squeezed and if some global inhibition is present is it possible to approximate a possible solution vector, $\varrho S \approx G_I = G - g_I\mathbf{1}$. **b)** Global inhibition and a small threshold are also required if the solution vector S lies within the unit hypercube. This is because synaptic saturation (ΔF) may prevent the weight vector G from being updated in the 'correct' direction ΔL (in the sense that $(\varrho S - G_I)\Delta G > 0$). This is not possible if G is close to the main diagonal and far from $\mathbf{0}$ and $\mathbf{1}$ (small ϱ and g_I between 0 and 1). **c)** A positive scalar product $(\varrho S - G_I)\Delta G > 0$ ensures that the G_I moves towards ϱS, provided that the learning rate q is small.

3.2 Simulations

We trained a binary perceptron with stochastic learning with both random uncorrelated binary patterns, as in [4], and a more complex data set of LaTeX deformed characters, preprocessed as in [5]. In Fig. 2A we show the number of iterations per pattern needed to converge to a solution as a function of the scaling factor ϱ 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 it becomes impossible to converge. If learning is too slow the number of iterations scales as $1/q$, but the convergence is always guaranteed. In Fig. 2B we show again the number of iterations per pattern for $10, 20, 40$ random uncorrelated binary patterns (prob. of active neuron: $f = 1/4$) 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 easily explained by looking at the variance of h (see Fig.2C when the same sequence of the same patterns is presented many times (every time with a different realization of the ζ variables of Eq. 1). The uncorrelated component of the variance (the variance that one would have if different synapses were statistically independent) scales as $1/N$ and the correlated component (the remaining part) scales in the same way and it is always negative. We also trained the perceptron on more complex LaTeX deformed characters. The network is composed of 2000 input neurons and 26 output neurons, one for each class (corresponding to the letters of the alphabet). During training the network is presented with 32 stimuli per class and for every pattern the unit corresponding to the correct class is turned on while the others are kept silent. The maximal error rate is 0.5% (percentage of the 832 patterns which are misclassified) and is usually due to false positives.

4 Conclusions

We showed 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.

The stochastic selection mechanism can be implemented in terms of a detailed spike-driven synaptic dynamics by exploiting the irregularity of the spike trains. Indeed the same mean spike frequencies might correspond to a large number of different spike train realizations: some of them might have the properties to induce long lasting modifications, while others would simply leave the synapse unchanged. Such a stochastic selection can be easily implemented by accumulating coincidences of events (e.g. pre-syn spike and high post-syn depolarization) with temporary synaptic modifications. These changes are then consolidated only if some threshold is crossed [3]. In such a case the load of generating the noise to drive the stochastic selection mechanism is transferred outside the synapse, delegated to the collective behavior of the interacting neurons [6], and allows for much more controllable small transition probabilities.

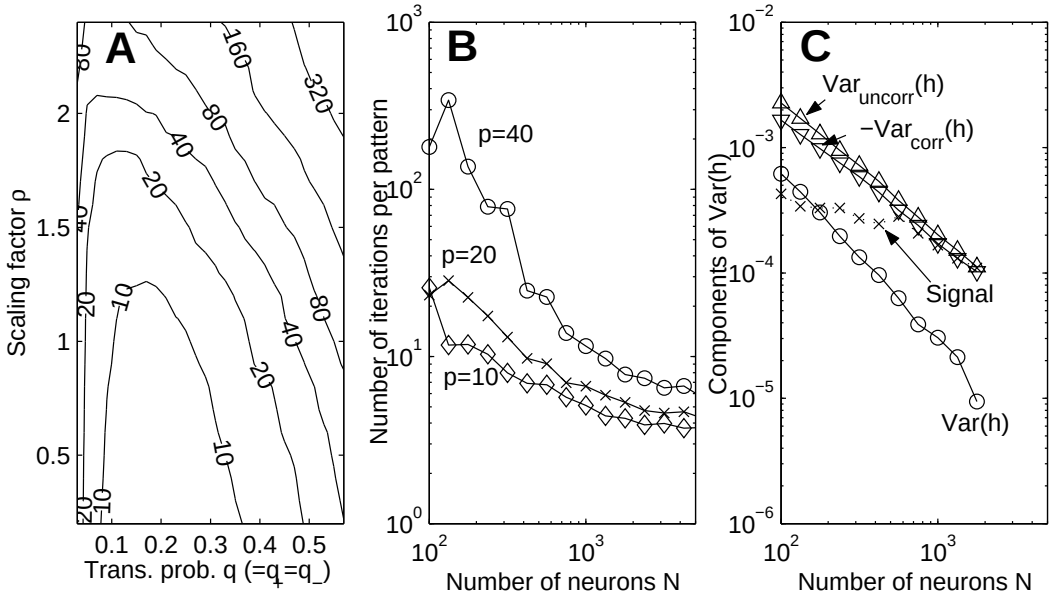

Fig. 2. Simulation results. A: number of iterations per pattern as a function of the learning rate q and the scaling factor ϱ. B: number of iterations per pattern as a function of the number of neurons N for $p = 10, 20, 40$ ($q = 0.05, f = 1/4, \theta = 0.01$). C: Correlated (down triangles) and uncorrelated 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. See the text for discussion.

Acknowledgements

This work was supported by the EU grant IST-2001-38099 ALAVLSI. We thank Massimo Mascaro for very useful discussions and for providing the LaTeX preprocessed data set.

References

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

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

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

[4] D.J. Amit and S. Fusi (1994). Dynamic learning in neural networks with material synapses *Neural Comp.*, **6**, 957.

[5] Y. Amit and M. Mascaro (2001) Attractor Networks for Shape Recognition, Neural Computation, **3**, 1415-1442

[6] E. Chicca, S. Fusi (2001), Stochastic synaptic plasticity in deterministic aVLSI networks of spiking neurons, Proc. of the World Congress on Neuroinformatics, Editor Frank Rattay, ARGESIM/ASIM Verlag, Vienna, 468-477