Perceptron beyond the limit of capacity

P. del Giudice, S. Franz, M. A. Virasoro

J. Phys. France **50** (1989) 121-134 15 JANVIER **1989**, PAGE 121

Classification
Physics Abstracts
05.90 — 02.70

Perceptron beyond the limit of capacity

P. Del Giudice, S. Franz ([1]) and M. A. Virasoro ([1])*

Istituto Superiore di Sanità-Laboratorio di Fisica, Viale Regina Elena 299-Roma, Italy

([1]) Dipartimento di Fisica dell' Università di Roma « la Sapienza », Piazzale Aldo Moro 4-Roma, Italy

(*Reçu le 16 août 1988, accepté le 15 septembre 1988*)

Résumé. — Nous considérons une application entrée-sortie pour un perceptron dans laquelle les formes sont divisées en classes. L'analyse par la mécanique statistique dans le cas d'un nombre fini de classes donne les mêmes résultats que pour une seule classe ; nous calculons la limite de capacité et les paramètres d'ordre pertinents en champ moyen. Nous généralisons ensuite l'analyse à l'ensemble canonique de Derrida-Gardner dans lequel le perceptron peut être étudié au-delà de sa limite de capacité. Nous complétons l'analyse en étudiant numériquement la règle d'apprentissage du perceptron. Nous discutons finalement la relevance de ces résultats à l'émergence possible d'une catégorisation spontanée.

Abstract. — An input-output map in which the patterns are divided into classes is considered for the perceptron. The statistical mechanical analysis with a finite number of classes turns out to give the same results as the case of only one class of patterns ; the limit of capacity and the relevant order parameters are calculated in a mean field approach. The analysis is then extended to the Derrida Gardner canonical ensemble in which the perceptron can be studied beyond the limit of capacity. We complete the analysis with numerical simulations with the perceptron learning rule. The relevance of those results to the possible emergence of spontaneous categorization is finally discussed.

Introduction.

Neural networks can be trained to learn certain rules : i.e. mappings that associate an output word to any input belonging to a certain space. In some cases they generalize : starting from a subset of instances of the mapping the machine reaches a configuration in which the rule is correctly realized for all possible inputs. The network is then said to implement the rule.

In the analysis of this capability the concept of the « entropy » of a rule has proved useful [1]. By this one means the measure of the mutiplicity of network configurations that implement the given rule : the larger the entropy, the smaller the number of instances needed

(*) Signatures in alphabetic order.

during the training. In [1] a thermodynamical analysis of a boolean network through an exhaustive enumeration of all possible configurations of the network has led to a direct calculation of the entropy of all rules.

The same problem, but for a different architecture can be approached analytically using the techniques introduced by E. Gardner [2]. This author has shown how to use the replica method to calculate the volume of network configurations capable of storing a certain number of patterns. The method was originally applied to fully connected networks of the Hopfield type ; a straightforward extension of this technique allows the calculation of the entropy for a two layer feedforward architecture, the perceptron [3] (attempts to apply the same method to multilayer machines have so far been unsuccessful).

As shown in the classical work by Minsky and Pappert [4], there are rules that the perceptron cannot learn ; when a calculation *à la* Gardner is possible, this limitation reflects itself in a network configurations volume which goes to zero when one trains the machine with a too large number of examples ; for instance, for the random mapping, the perceptron will be able to learn up to $2N$ examples (N is the number of input neurons). What happens if one tries to train the system with a larger number of examples ?

In this paper we address this question ; in particular we analyze the following type of rules : a) the input and the output patterns are assumed to be grouped in a finite (when $N \to \infty$) number of classes ; b) if two inputs belong to the same class, the corresponding outputs also belong to the same class. If inside the corresponding classes the patterns to be mapped are chosen at random, then again the perceptron will be able to learn only a finite number of examples, and the rule cannot be implemented. In these conditions one may ask if the machine, although incapable to realize the correct correspondence between individuals, succeeds in associating correctly the classes.

In the first section we define the rule to be analyzed and give analytical results about the limit of capacity of the perceptron for this rule and the minimum error produced by the machine beyond this limit ; in the second section we study through numerical simulations, using the perceptron learning algorithm (the so-called δ-rule) [3, 4, 5], the properties of the perceptron beyond its limit of capacity, which are different from the corresponding « thermodynamical » properties. In this context we discuss the relevance of the initial conditions for the asymptotic behaviour of the machine.

We end the paper with some concluding remarks on the relevance of these calculations to understand categorization in neural network.

1. Replica analysis of the perceptron.

The architecture we consider is the perceptron [3] ; a neural network consisting of just one input and one output layer of binary units, directly communicating through a matrix of real valued connections.

Our notation will be the following : $\sigma_j = \pm 1\ j = 1, \ldots, N$, are the values taken by the N input units, $s_i = \pm 1$, $i = 1, \ldots, N'$, are the values taken by the N' output units, J_{ij} denotes the strength of the connection between unit j in input and unit i in output.

The relation between the values on the output units and those on the input units is the usual step function

$$s_i = \operatorname{sgn}\left(\sum_{j=1}^{N} J_{ij}\,\sigma_j\right) \quad i = 1, \ldots, N' \tag{1.1}$$

in which zero threshold is assumed for all output units.

Our rule is defined as follows : we consider R input and output patterns respectively denoted by $\{\sigma_j^\mu\}$ and $\{\xi_i^\mu\}$ (« prototypes ») which we use for the definition of R corresponding classes. Each of the σ_j^μ and ξ_i^μ are randomly chosen as ± 1 with probability 1/2 independently from each other. The rule is the mapping :

$$\{\sigma_j^{\mu\nu}\} \to \{\xi_i^{\mu\nu}\} \tag{1.2}$$

where $\{\sigma_j^{\mu\nu}\}$ and $\{\xi_i^{\mu\nu}\}$ are the set of values which determine respectively the input and the output pattern ν, $\nu = 1, \ldots, Q$ which belongs to the class μ. $P = QR$ is the total number of patterns ; we denote with α the ratio $\dfrac{P}{N}$.

The values $\sigma_j^{\mu\nu}$ and $\xi_i^{\mu\nu}$ are independently chosen with the following probability law :

$$P(\sigma_j^{\mu\nu} | \sigma_j^\mu) = \frac{1 + m_{\text{in}}\, \sigma_j^{\mu\nu}\, \sigma_j^\mu}{2} \tag{1.3a}$$

$$P(\xi_i^{\mu\nu} | \xi_i^\mu) = \frac{1 + m_{\text{out}}\, \xi_i^{\mu\nu}\, \xi_i^\mu}{2} . \tag{1.3b}$$

In this way m_{in} and m_{out}, « magnetizations », represent the average overlap between the individuals of a class and the corresponding prototypes.

We say that our rule is implementable provided a set of $\{J_{ij}\}$ exists such that

$$\xi_i^{\mu\nu} \sum_{j=1}^{N} J_{ij}\, \sigma_j^{\mu\nu} > 0 \quad \forall i, \mu, \nu . \tag{1.4}$$

1.1 THE MICROCANONICAL APPROACH. — As we stressed in the introduction we are interested in the possibility of giving a quantitative estimate of a suitably defined « complexity » with respect to a given architecture without explicit reference to a particular learning algorithm. This can be achieved in the way recently pioneered by E. Gardner [2] ; the quantity of interest is identified as the entropy density S [1] defined by

$$S = \frac{1}{N'} \lim_{N \to \infty} \frac{1}{N} \ln V(\{\xi_i^{\mu\nu}\}, \{\sigma_j^{\mu\nu}\}) \tag{1.5}$$

where $V(\{\xi_i^{\mu\nu}\}, \{\sigma_j^{\mu\nu}\})$ is the normalized volume in the space of all possible $\{J_{ij}\}$ occupied by the set of those satisfying (1.4). Explicitly :

$$V(\{\xi_i^{\mu\nu}\}, \{\sigma_j^{\mu\nu}\}) = \int d\mu\, \{J_{ij}\} \prod_{\mu\nu i} \theta\left(\xi_i^{\mu\nu} \sum_{j=1}^{N} J_{ij}\, \sigma_j^{\mu\nu}\right) \tag{1.6}$$

where the measure $d\mu\, \{J_{ij}\}$ is :

$$d\mu\, \{J_{ij}\} = \frac{\left(\prod_{ij} dJ_{ij}\right) \prod_i \delta\left(\sum_j J_{ij}^2 - J^2\right)}{\int \left(\prod_{ij} dJ_{ij}\right) \prod_i \delta\left(\sum_j J_{ij}^2 - J^2\right)} .$$

The calculation of V has several analogies with the microcanonical ensemble approach to the statistical mechanics of systems with quenched disorder. With reference to the usual example

of such systems, the spin glasses, we note that the role of quenched variables is played here by the « spin » variables ($\sigma_j^{\mu\nu}$ and $\xi_i^{\mu\nu}$) while the $\{J_{ij}\}$ are annealed.

As usual one assumes that extensive thermodynamical quantities are independent of the particular sample of quenched variables. So S is a function of m_{in}, m_{out}, α and coincides with its mean value over the σ and the ξ

$$S(\alpha, m_{\text{in}}, m_{\text{out}}) = \frac{1}{N'} \lim_{N \to \infty} \frac{1}{N} \times \left\langle \ln V\left(\left\{\xi_i^{\mu\nu}\right\}, \left\{\sigma_j^{\mu\nu}\right\}\right)\right\rangle_{\xi, \sigma} . \tag{1.7}$$

Provided the number of classes R remains finite as N goes to infinity, the result for S can be extracted from a mean field theory in the replica approach :

$$S(\alpha, m_{\text{in}}, m_{\text{out}}) = \frac{1}{N'} \lim_{N \to \infty} \lim_{n \to 0} \frac{1}{N} \frac{1}{n} \times \ln \left\langle V^n\left(\left\{\xi_i^{\mu\nu}\right\}, \left\{\sigma_j^{\mu\nu}\right\}\right)\right\rangle_{\xi, \sigma} .$$

Following the methods used in [2] the result is obtained by a saddle point with respect to the variables q^{ab} and M_μ^a (which play the role of order parameter), for which a replica symmetric ansatz is chosen ($q^{ab} = q$; $M_\mu^a = M_\mu$) [7] :

$$S(\alpha, m_{\text{in}}, m_{\text{out}}) = \frac{1}{2}\left[\frac{q}{1-q} + \ln(1-q)\right] +$$

$$+ \alpha \int \frac{dy}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \left[\frac{1 - m_{\text{out}}}{2} \ln H\left(\frac{-aM_\mu + \sqrt{qy}}{\sqrt{1-q}}\right) + \begin{pmatrix} m_{\text{out}} \to -m_{\text{out}} \\ m_{\text{in}} \to -m_{\text{in}} \end{pmatrix}\right] \tag{1.8}$$

where $a = \dfrac{m_{\text{in}}}{\sqrt{1 - m_{\text{in}}^2}}$ and :

$$H(x) = \int_x^\infty e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} .$$

The values M_μ and q are determined by the saddle point equations :

$$\frac{q}{\sqrt{1-q}} = \alpha \int \frac{dy}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \left[\frac{1 + m_{\text{out}}}{2} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \frac{\sqrt{1-q}\,x + y\sqrt{\frac{(1-q)}{q}}}{H(x)} + \begin{pmatrix} m_{\text{out}} \to -m_{\text{out}} \\ m_{\text{in}} \to -m_{\text{in}} \end{pmatrix}\right]$$

$$0 = \int \frac{dy}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \left[\frac{1 + m_{\text{out}}}{2} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \frac{(1-q)^{-1/2}}{H(x)} - \begin{pmatrix} m_{\text{out}} \to -m_{\text{out}} \\ m_{\text{in}} \to -m_{\text{in}} \end{pmatrix}\right] \tag{1.9}$$

in which

$$x = \frac{-aM_\mu + \sqrt{qy}}{\sqrt{1-q}} .$$

The equations show that $M_\mu \equiv M$ is independent from μ. The « physical » meaning of the order parameters q and M is clear from the following formulas :

$$q = \left\langle \frac{\sum_j J_{ij}^a J_{ij}^b}{J^2} \right\rangle_{\xi\sigma}, \quad a \neq b ;$$
$$M = \left\langle \frac{\xi_i^\mu \sum_j J_{ij}^a \sigma_j^\mu}{J^2} \right\rangle_{\xi\sigma} . \tag{1.10}$$

So q measures the average overlap between pairs of solutions while M is a kind of « magnetization » for the $\{J_{ij}\}$. From this interpretation it is clear that the minimum value for $V(\{\xi_i^{\mu\nu}\}\{\sigma_j^{\mu\nu}\})$ is obtained when $q \to 1$ and this determines the critical value α_c for α; so

$$S(\alpha, m_{in}, m_{out}, q, M) \to -\infty ; \quad \alpha \to \alpha_c (q \to 1) . \tag{1.11}$$

In this limit the saddle point equations take the form :

$$1 = \alpha_c \left[\int_{aM}^{\infty} \left(e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} \frac{1 + m_{out}}{2} (-aM + y)^2 \right) - \begin{pmatrix} m_{out} \to -m_{out} \\ m_{in} \to -m_{in} \end{pmatrix} \right]$$
$$0 = \int_{aM}^{\infty} \left(e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} \frac{1 + m_{out}}{2} (-aM + y) \right) + \begin{pmatrix} m_{out} \to -m_{out} \\ m_{in} \to -m_{in} \end{pmatrix} . \tag{1.12}$$

We notice that in these equations, as well as in the expression for the entropy, there is not explicit dependence on m_{in}, as long as $m_{in} \neq 0$; the dependence on m_{in} is only implicit through the factor (aM), so the saddle point equations can be numerically solved with respect to the variables (aM) and α_c as functions of m_{out}. When $m_{out} = m_{in}$ equations (1.12) become identical to those obtained by E. Gardner in [2]. The result is shown in the figures 1 and 2.

A simple modification of the calculation for S takes into account the case in which the magnetizations for the input and the output patterns $(m_{in}^\mu, m_{out}^\mu)$, as well as the number of patterns $(P^\mu = \alpha^\mu N)$ are different for each class. In this case :

$$S(\{\alpha^\mu\}, \{m_{in}^\mu\}, \{m_{out}^\mu\}) = \frac{1}{2}\left[\frac{q}{1-q} + \ln (1-q) \right] +$$
$$+ \sum_\mu \alpha^\mu \int \frac{dy}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \left[\frac{1 - m_{out}^\mu}{2} \ln H\left(\frac{a^\mu M_\mu + \sqrt{qy}}{\sqrt{1-q}} \right) + \begin{pmatrix} m_{out}^\mu \to -m_{out}^\mu \\ m_{in}^\mu \to -m_{in}^\mu \end{pmatrix} \right] . \tag{1.13}$$

The corresponding equations for the limit of capacity are

$$1 = \sum_\mu \alpha_c^\mu \left[\int_{a^\mu M^\mu}^{\infty} \left(e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} \frac{1 + m_{out}^\mu}{2} (-a^\mu M^\mu + y)^2 \right) - \begin{pmatrix} m_{out}^\mu \to -m_{out}^\mu \\ m_{in}^\mu \to -m_{in}^\mu \end{pmatrix} \right] \tag{1.14}$$

$$0 = \int_{a^\mu M^\mu}^{\infty} \left(e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} \frac{1 + m_{out}^\mu}{2} (-a^\mu M^\mu + y) \right) + \begin{pmatrix} m_{out}^\mu \to -m_{out}^\mu \\ m_{in}^\mu \to -m_{in}^\mu \end{pmatrix} . \tag{1.15}$$

Equation (1.14) can be written as $1 = \sum_\mu \alpha_c^\mu A^\mu$ with all $A^\mu > 0$. From figure 1, which can be viewed as the plot of $1/A^\mu$ *versus* m_{out}^μ, it can be seen that A^μ is a decreasing function of

m^{μ}_{out}. It follows that if one wanted to maximize the total number of learned patterns one would have to put them in those classes which share the same highest m^{μ}_{out} regardless of the way they are distributed among them.

We now come back to the simple case in which the magnetizations are the same for all classes.

When $m_{in} = 0$, then $\alpha_c = 2$ for all values of m_{out}. From figures 1 and 2 we see that for $m_{in} \neq 0$ α_c is a monotonically increasing function of m_{out} ranging from $\alpha_c = 2$ to $\alpha_c = +\infty$, and similarly for M which ranges from 0 to ∞.

So when $m_{out} \neq 1$ the limit of capacity is finite and we have to say that the rule is not implementable in the sense defined by equation (1.4). It is only for $m_{out} = 1$ (and $m_{in} \neq 0$) that direct inspection shows that $q \to 0$ for any α while at the same time $M \to \infty$.

The fact that q does not grow to 1 proves that there are always a great number of different $\{J_{ij}\}$ that contribute.

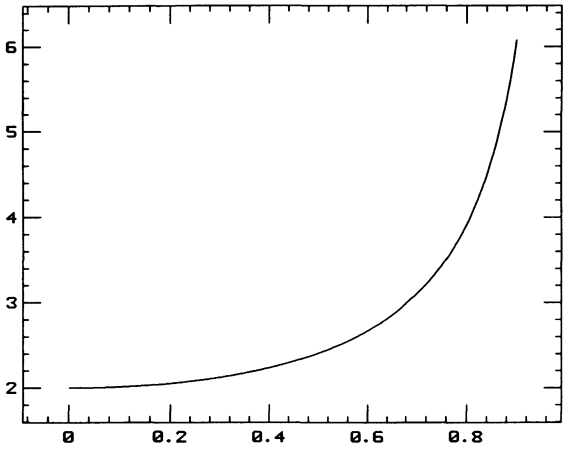

Fig. 1. — Maximum capacity *versus* the output patterns magnetization.

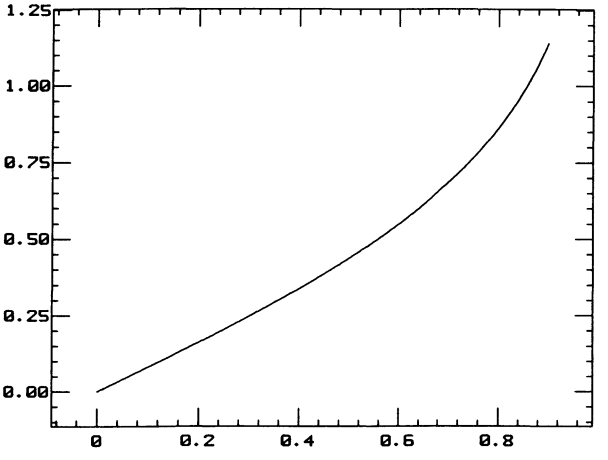

Fig. 2. — (aM) *versus* the output patterns magnetization.

The fact that $M \to \infty$ shows that generalization is occurring. Given a pattern $\sigma^{\mu\nu}$ which the machine has not been exposed to, we consider the quantity

$$x_i = \xi_i^\mu \sum_{j=1}^{N} J_{ij}\, \sigma_j^{\mu\nu} = \xi_i^\mu \sum_{j=1}^{N} J_{ij}\, \sigma_j^\mu\, \varepsilon_j^{\mu\nu} \tag{1.16}$$

having set $\sigma_j^{\mu\nu} = \sigma_j^\mu\, \varepsilon_j^{\mu\nu}$. Equations (1.3a, b), when used to derive the probability distribution of $\varepsilon_j^{\mu\nu}$ imply that the variables $(J_{ij}\, \sigma_j^\mu)$ and $\varepsilon_j^{\mu\nu}$ are statistically independent, and $P(\varepsilon_j^{\mu\nu} = \pm 1) = (1 \pm m_{\text{in}})/2$. Therefore x_i is a sum of a large number of uncorrelated terms and the central limit theorem can be used to evaluate the probability that $x_i < 0$, that is the probability of an incorrect classification of the pattern $\sigma_j^{\mu\nu}$ at point i :

$$P(x < 0) \simeq \int_{aM}^{\infty} e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} \equiv H(aM)\,. \tag{1.17}$$

As $M \to \infty H(aM) \to 0$ and generalization appears.

1.2 THE CANONICAL APPROACH. — In the calculation for S we have only included configurations $\{J_{ij}\}$ for which zero error is produced ; it is interesting to investigate the optimal property of the machine beyond the limit of capacity. This means finding the minimum error given by the perceptron for a $\alpha > \alpha_c$ and the relevant properties of the configurations producing this error.

This can be achieved if we move from the microcanonical calculations for S to a canonical ensemble approach in which we consider the error as the energy function of our system [13]. So we can introduce a partition function Z

$$Z = \int d\mu(\{J_{ij}\}) \times \exp\left[-\beta \sum_{i\mu\nu} \theta\left(-\xi_i^{\mu\nu} \sum_{j=1}^{N} J_{ij}\, \sigma_j^{\mu\nu}\right)\right] \tag{1.18}$$

and a density of free energy

$$f = \frac{1}{N'} \lim_{N\to\infty} -\frac{1}{\beta N} \ln Z \tag{1.19}$$

for which it is reasonable to assume the self-averaging property. We can calculate the minimum error E_0 (normalized to one) as the zero temperature limit of the internal energy density U divided by α :

$$U = \frac{\partial(\beta f)}{\partial f}\,; \quad E_0 = \lim_{\beta\to\infty} \frac{U}{\alpha} = \lim_{\beta\to\infty} \frac{f}{\alpha}\,. \tag{1.20}$$

Using the relation

$$e^{a\theta(-x)} = \theta(x) + \theta(-x)\, e^a$$

we see that the same tools suitable for the calculation of S can be employed here. The result for f is :

$$E_0 = \lim_{\beta\to\infty} \frac{f}{\alpha} = -\frac{1}{ax^2} + \left[\frac{1 + m_{\text{out}}}{2}\left(\int_{x+aM}^{\infty} \frac{dy}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} + \frac{1}{x^2}\int_{aM}^{x+aM} \frac{dy}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} (-aM + y)^2\right)\right] + \left[\begin{pmatrix} m_{\text{out}} \to -m_{\text{out}} \\ m_{\text{in}} \to -m_{\text{in}} \end{pmatrix}\right] \tag{1.21}$$

where $x = \lim_{\beta \to \infty} \sqrt{2\beta(1-q)}$ and M are given by the saddle point equations :

$$1 = \alpha \left\{ \left[\frac{1+m_{\text{out}}}{2} \int_{aM}^{x+aM} e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} (-aM+y)^2 \right] + \left[\begin{pmatrix} m_{\text{out}} \to -m_{\text{out}} \\ m_{\text{in}} \to -m_{\text{in}} \end{pmatrix} \right] \right\}$$
$$0 = \frac{1+m_{\text{out}}}{2} \int_{aM}^{x+aM} e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} (-aM+y) - \begin{pmatrix} m_{\text{out}} \to -m_{\text{out}} \\ m_{\text{in}} \to -m_{\text{in}} \end{pmatrix} . \tag{1.22}$$

From the equations for f valid for all β it can be seen that in order to have $E_0 > 0$ the following relation must be satisfied :

$$\lim_{\beta \to \infty} \beta(1-q) = a \textit{ finite number}$$

this means that for all $\alpha > \alpha_c$, we have $q = 1$ and zero volume for the relevant configurations $\{J_{ij}\}$.

It is possible to calculate explicitly the minimum error E_0 in the limit $\alpha \to \infty$, in which equations (1.19) show that $x \to 0$, $aM \to \infty$ in such a way that

$$\lim_{\alpha \to \infty} E_0 = \frac{1-m_{\text{out}}}{2} . \tag{1.23}$$

This corresponds to configurations $\{J_{ij}\}$ for which

$$\text{sgn}\left(\sum_{j=1}^{N} J_{ij} \sigma_j^{\mu\nu} \right) = \xi_i^{\mu} \tag{1.24}$$

for all $\{\sigma_j^{\mu\nu}\}$ belonging to the class μ ; i.e. the minimum error is achieved when the machine associate to each pattern the prototype of the corresponding class.

2. Numerical simulations.

We performed numerical simulations to investigate the behaviour of the perceptron beyond its limit of capacity. We used the δ-rule learning algorithm [4], [5] for which a convergence theorem exists so that one knows that the machine will find the solution if at least one exists.

In this section we will analyze the following questions :

1) what are the generalization properties of the perceptron for a rule (1.2) when $m_{\text{out}} = 1$?

2) what are the performance of the machine for the same rule with $m_{\text{out}} < 1$?

In particular : will the machine succesfully extract the regularities of the rule i.e. associate correctly the classes, even when it cannot learn the exact input-output correspondence ?

The following version of the δ-rule was used : after the presentation of each pair of input and output patterns, the J's were updated according to

$$J_{ij} \to J_{ij} + \eta \left(\xi_i^{\mu\nu} - \text{sgn}\left(\sum_{j=1}^{N} J_{ij} \sigma_j^{\mu\nu} \right) \right) \sigma_j^{\mu\nu} \equiv J_{ij} + \eta (\xi_i^{\mu\nu} - s_i^{\mu\nu}) \sigma_j^{\mu\nu} \tag{2.1}$$

where η is a parameter that measures the relative size of the updating term with respect to the actual J value.

The initial configuration $\{J_{ij}^0\}$ was chosen at random. The formula (2.1) shows that the output sites are independent from each other, a perceptron with N' output units is equivalent to N' perceptrons with one output unit each, in which the same patterns are presented in input. We restricted ourselves to the case $N' = 1$, so that J_{ij} becomes J_j, s_i becomes s and so on. It is easy from this case to recover the general one.

We call a time step a complete scanning of the P examples in the training set. After each time step, the following quantities were calculated :

$$E = \frac{1}{P} \sum_{\mu\nu} \frac{(\xi^{\mu\nu} - s^{\mu\nu})^2}{4}$$

and

$$m_s^{\text{tr}} = \frac{1}{P} \sum_{\mu\nu} s^{\mu\nu} \xi^{\mu} . \tag{2.2}$$

E represents the average error made over the whole set of training patterns ; m_s^{tr}, is the average overlap between the output obtained for a given input pattern and the prototype of the corresponding output class.

We also looked at quantities which caracterize the configurations of the J's produced by the algorithm :

$$q = \frac{\sum_{j=1}^{N} J_j J_j'}{\sqrt{\sum_j J_j^2 \sum_j J_j'^2}} \tag{2.3}$$

where J_j and J_j' are two configurations produced for the same set of training patterns but different initial conditions,

$$M = \frac{1}{R} \sum_{\mu=1}^{R} \left(\frac{\xi^{\mu} \sum_j J_j \sigma_j^{\mu}}{\sqrt{\sum_j J_j^2}} \right) . \tag{2.4}$$

It must be noted that q and M produced by the algorithm (2.1) are in general not the same as those introduced in section 1, since the δ-rule does not mimic the thermodynamical behaviour, as we will see in the following.

We also measured the correlation coefficient between the configuration $\{J_i\}$ obtained and the input patterns $\{\sigma_j^{\mu\nu}\}$.

The learning procedure (2.1) was iterated until either $E = 0$ was obtained (in the case $m_{\text{out}} = 1$) or a stationary regime was reached in which for all relevant quantities we could detect only fluctuations around a constant average.

After the training a test was performed on a new set of P' patterns $\{\sigma_j'^{\mu\nu}\}$ for the same rule.

We measured the magnetization m_s^{test}

$$m_s^{\text{test}} = \frac{1}{P'} \sum_{\mu\nu} s'^{\mu\nu} \xi^{\mu} . \tag{2.5}$$

The results for $m_{\text{out}} = 1$ are the following : although small variations of η are irrelevant, we can distinguish two different regimes according to the order of magnitude of $\hat{\eta} = \dfrac{\eta m_{\text{in}}}{|J_j^0|}$.

When $\hat{\eta} \sim 1$ the updating terms in the J's, when different from zero, are much bigger than the initial term J_j^0 ; the rule is learned after few patterns are presented to the machine, with a very small probability of error. The resulting configuration for the J's is

$$J_j = 2\,\eta \sum_{\mu} \xi^{\mu}\,\sigma_j^{\mu 1} + J_j^0 \tag{2.6}$$

where $\{\sigma_j^{\mu 1}\}$ denotes the first pattern belonging to the class μ for which the machine gives the wrong answer $s^{\mu 1} = -\,\xi^{\mu}$.

It is interesting to notice that the dominant term in (2.6) has a form similar to the one used by Hopfield [13].

From (2.6) we see that

$$\frac{1}{N}\,\xi^{\mu} \sum_{j=1}^{N} J_j\,\sigma_j^{\mu} = 2\,\eta m_{\text{in}} + O\left(\frac{J^0}{N}\right)$$

this shows that $M \sim \sqrt{N}$ and so $H(aM) \sim \mathrm{e}^{-\text{const.}\,N}$.

When $\hat{\eta} \sim \dfrac{1}{\sqrt{N}}$ we can distinguish two different situations according to the relative sizes of P and N. When $P \ll N$ we observe what was called « memorization » behaviour in [1], that is a quick learning of the training patterns, but a high probability of errors for the test set. On the other hand, for $P \sim N$ the generalization regime sets in. The probability of error for a test set of P new patterns is small. In fact we observe that for a sufficiently small η, M increases slowly with P in such a way that $H(aM)$, equation (1.17), is of order $\dfrac{1}{P}$. Therefore for $P \to \infty$ the behaviour of the perceptron with respect to the test set and the training set is qualitatively similar.

We now turn to the case $m_{\text{out}} < 1$. The main results are : 1) During training the magnetization m_s^{tr} stabilizes itself at the value m_{out}, apart from small fluctuations, for all values of α. The error shows a dependence upon α ; it can be written as :

$$E = \frac{1 - \langle \xi^{\mu\nu}\,\xi^{\mu}\rangle\,\langle s^{\mu\nu}\,\xi^{\mu}\rangle}{2} - \frac{r}{2} \tag{2.7}$$

where

$$r = \langle \xi^{\mu\nu}\,s^{\mu\nu}\rangle - \langle \xi^{\mu\nu}\,\xi^{\mu}\rangle\,\langle s^{\mu\nu}\,\xi^{\mu}\rangle \tag{2.8}$$

is the correlation between the ξ's and the s's ; this means that in the case $m_s^{\text{tr}} = m_{\text{out}}$, the error is

$$E = \frac{1 - m_{\text{out}}^2}{2} = \frac{r}{2} \tag{2.9}$$

(the angular brackets in (2.7) and (2.8) stand for a average over the indices μ and ν). r becomes negligible for values of α sufficiently greater than α_c (see Fig. 4). For the test patterns, the magnetization m_s^{test} obtained is less then m_{out}, the difference being higher, the higher is r. r decreases with α (see Fig. 3), and becomes eventually zero for $\alpha \gg 1$ within statistical errors. This means that the correlation between the ξ's and the s's and, accordingly, the one between the J's and the σ's, vanish. For $r = 0$ the error is $E = \dfrac{1 - m_{\text{out}}^2}{2}$. It is interesting to notice that the minimum error predicted by the thermodynamical analysis in

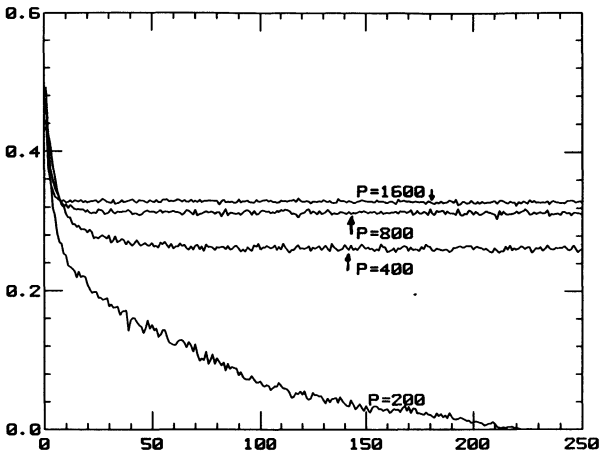

Fig. 3. — Error produced by the perceptron during training *versus* learning time for various dimensions of the training patterns set. $N = 100$, $N' = 1$, $m_{in} = 0.02$, $m_{out} = 0.6$.

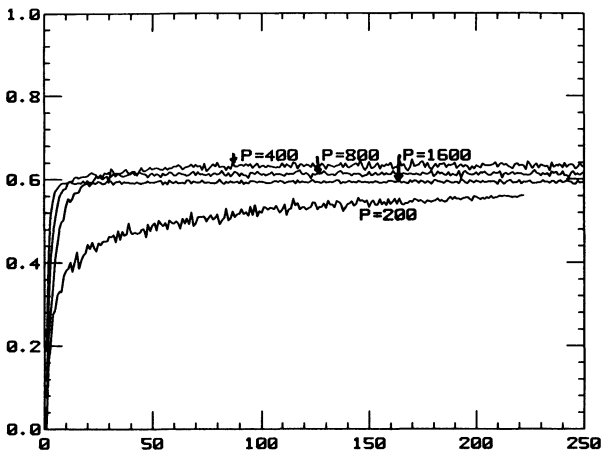

Fig. 4. — Average overlap between the pattern produced by the perceptron and the corresponding prototype *versus* learning time for various dimensions of the training patterns set. $N = 100$, $N' = 1$, $m_{in} = 0.02$, $m_{out} = 0.6$.

Sect. 1 in the limit $\alpha \to \infty$ is $\dfrac{1 - m_{out}}{2}$, which is definitely smaller than the one obtained. For a sufficiently small η, and assuming that the correlation between the J's and the σ's are negligible an argument can be given that accounts for the fact that the δ-rule drives the system to a stationary regime for which $m_s^{test} = m_{out}$.

At each learning step the variation of J_j is

$$\delta J_j = \eta \, (\xi^{\mu\nu} - s^{\mu\nu}) \, \sigma_j^{\mu\nu} \tag{2.10}$$

where

$$s^{\mu\nu} = \text{sgn}\left(\sum_{j=1}^{N} J_j \sigma_j^{\mu\nu}\right)$$

we can average the quantity δJ_j over j

$$\overline{\delta J} = \eta\,(\xi^{\mu\nu} - s^{\mu\nu})\, m_{\text{in}} \tag{2.11}$$

using the fact that the variable $\xi^{\mu\nu}$ and $s^{\mu\nu}$ are independent, due to the independence of J_j and $\sigma_j^{\mu\nu}$ we can write the probability distribution of $\overline{\delta J}$ as follows

$$\overline{\delta J} = \begin{cases} 0\,, & \text{with prob.} \quad \dfrac{1+m_{\text{out}}}{2}\,(1-H(aM)) + \dfrac{1-m_{\text{out}}}{2}\,H(aM) \\ 2\,\eta m_{\text{in}}\,, & \text{—} \quad \dfrac{1+m_{\text{out}}}{2}\,H(aM) \\ -2\,\eta m_{\text{in}}\,, & \text{—} \quad \dfrac{1-m_{\text{out}}}{2}\,(1-H(aM)) \end{cases} \tag{2.12}$$

so the first two moments of $\overline{\delta J}$ are

$$\langle \overline{\delta J} \rangle = 2\,\eta m_{\text{in}} \left(H(aM) - \frac{1-m_{\text{out}}}{2} \right) \tag{2.13a}$$

and

$$\langle (\overline{\delta J})^2 \rangle = 4(\eta m_{\text{in}})^2 \left(\frac{1-m_{\text{out}}}{2} + m_{\text{out}}\, H(aM) \right). \tag{2.13b}$$

The (2.13a) can be viewed as the equation of motion for M ; since H is a monotonically decreasing function of its argument, the system will evolve to a stationary state for which

$$\langle \overline{\delta J} \rangle = 0 \quad \text{i.e. } H(aM) = \frac{1-m_{\text{out}}}{2}\,;$$

this implies $m_s^{\text{test}} = m_{\text{out}}$.

Obviously the mean error during training is given by the probability $P(s^{\mu\nu} = -\xi^{\mu\nu})$ which is

$$P(s^{\mu\nu} = -\xi^{\mu\nu}) = \frac{1+m_{\text{out}}}{2}\,(1-H(aM)) + \frac{1-m_{\text{out}}}{2}\,H(aM)$$

that in the stationary regime gives :

$$P(s^{\mu\nu} = -\xi^{\mu\nu}) = \frac{1-m_{\text{out}}^2}{2}$$

consistently with numerical results.

Conclusions.

The problem of storing categories in Hopfield-like neural networks deserved the attention of several authors [8, 9, 10, 11, 12] ; they searched for modifications of the Hebb rule in such a way as to allow the memorization of hierarchically organized patterns. In these works (with

the possible exception of [11]) categorization does not emerge spontaneously, but comes out as something put by hand into the model.

Here we have analysed a perceptron « under stress » when it is trying to learn too many patterns.

Analytical calculations show that in such a situation there is an optimal strategy : to group the patterns in a finite number of classes each of which is characterized by a m_{out} that increases with the number of individuals inside. One could say then that the perceptron « forgets the details » and replaces the pattern by an emergent prototype.

Numerical calculations show that the δ-rule follows a slightly different strategy : the prototype does not emerge, the category is correctly recognised but details distinguishing the individuals are confused rather than neglected. As a consequence the error as defined in equation (2.2) is larger than the optimal one. We have checked that « back propagation » (minimizing the error), as it should improves on this aspect, but we refrain from drawing conclusions about this fact before an analysis of the local minima in the error surface.

A description of a network closely similar to the one analysed in this paper is discussed in [6]. In this example the ratio between the number of patterns and the number of neurons is such that this perceptron is also working above the limit of capacity and therefore our analysis in section 1 applies to it. The learning rule however is different from the δ-rule because the trace of a pattern in the J_{ij} matrix is assumed to decay with time. The authors then show that the prototype is learnt from the exemplars. This is to be compared with the different behaviour we found in section 2.

Finally we would like to point out two aspects that deserve further consideration. The possibility of replica symmetry breaking in subsections 1b and its relation with the existence of local minima in the error surface.

The second point which seems much more difficult is to generalize the results of this paper to an infinite number of classes or to an ultrametric tree of patterns with more layers of branchings.

It is a pleasure to thank M. Mezard for several helpful conservations. One of us (MAV) acknowledges a fellowship by the John Simon Guggenheim Memorial Fundation.

References

[1] PATARNELLO, S., CARNEVALI, P., Learning networks of neurons with boolean logic, *Europhys. Letters* **4** (1987) 503 ;
CARNEVALI, P., PATARNELLO, S., Exhaustive thermodynamical analysis of boolean learning network, *Europhys. Lett.* **4** (1987) 1199.

[2] GARDNER, E., The space of interactions in neural network models, *J. Phys. A* **21** (1988) 257 ; Maximum storage capacity in neural networks, *Europhys. Lett.* **4** (1987) 481.

[3] ROSEMBLATT, F., Principles of neurodynamics, Spartan Books, 1962, N.Y.

[4] MINSKY, M., PAPPERT, S., Perceptrons, MIT Press, Cambridge, Ma., 1969.

[5] RUMELHART, D. E., MCCLELLAND, J. L., Parallel distributed processing : explorations in the microstructure of cognition, Bredford Books, Cambridge, Ma., 1986, vol. 1, section 8.

[6] RUMELHART, D. E., MCCLELLAND, J. L., Parallel distributed processing : explorations in the microstructure of cognition, Bredford Books, Cambridge, Ma., 1986, vol. 2, section 17.

[7] In a slightly different context a similar calculation has been performed in : W. Krauth, M. Mezard, J. P. Nadal Basins of attraction in a perceptron-like neural network, preprint LPTENS 88/8.

[8] PARGA, N., VIRASORO, M. A., The ultrametric organization of memories in a neural network, *J. Phys. France* **47** (1986) 1857.

[9] DOTSENKO, V., *Physica* **140A** (1986) 410.
[10] GUTFREUND, H., *Phys. Rev.* **A 37** (1988) 570.
[11] TOULOUSE, G., DEHAENE, S., CHANGEAUX, J. P., Spin glass model of learning by selection, *Proc. Natl. Acad. Sci.* USA 1986, vol. 83, p. 1695.
[12] FEIGELMAN, M. V., IOFFE, L. B., The augmented models of associative memory, asymmetric interactions and hierarchy of pattern, *Int. J. of Mod. Phys.* **B 1** (1987) 51.
[13] DERRIDA, B., GARDNER, E., Optimal storage properties of neural network models, *J. Phys.* **A 21** (1988) 271.
[14] HOPFIELD, J. J., Neural network and physical system with emergent collective computational abilities, *Proc. Nat. Acad. Sci.* USA 1982, vol. 79, p. 2554.