IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Mapping correlated Gaussian patterns in a perceptron

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1989 J. Phys. A: Math. Gen. 22 L803

(http://iopscience.iop.org/0305-4470/22/16/007)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 129.252.86.83
The article was downloaded on 01/06/2010 at 06:58

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. 22 (1989) L803-L808. Printed in the UK

LETTER TO THE EDITOR

Mapping correlated Gaussian patterns in a perceptron

J F Fontanari† and R Meir

Division of Chemistry, Mail Code 164-30 CH, California Institute of Technology, Pasadena, CA 91125, USA

Received 10 May 1989, in final form 19 June 1989

Abstract. We study the performance of a single-layer perceptron in realising a binary mapping of Gaussian input patterns. By introducing non-trivial correlations among the patterns, we generate a family of mappings including easier ones where similar inputs are mapped into the same output, and more difficult ones where similar inputs are mapped into different classes. The difficulty of the problem is gauged by the storage capacity of the network, which is higher for the easier problems.

The use of statistical mechanics techniques in the analysis of feedback neural networks has led to a deep understanding of the equilibrium properties of these systems (Amit *et al* 1987). Feedback neural networks, e.g. the Hopfield–Little model (Hopfield 1982, Little 1974), have a non-trivial dynamics which possesses a huge number of attractors. Part of these attractors can be imprinted in the network through a learning procedure which specifies the strengths of the couplings between the neurons, thus allowing the network to be used as an associative memory (Hopfield 1982). On the other hand, single-layer feedforward neural networks, e.g. Rosenblatt's perceptron (Rosenblatt 1962), have a rather dull dynamics but a very rich learning process which had been fully studied in the 1960s through rigorous mathematical analysis, simulations on digital computers and by constructing an actual machine (Block 1962, Minsky and Papert 1969).

Recently Gardner (1988) and Gardner and Derrida (1988) have successfully rederived some of the results concerning the maximum storage capacity of the perceptron in the framework of the equilibrium statistical mechanics (see also Opper (1988, 1989) for a more recent contribution). The architecture of the perceptron considered in those studies is shown in figure 1. The input layer consists of N neurons $\{\xi_i = \pm 1, i = 1, \ldots, N\}$, each one connected to the output neuron $S = \pm 1$ through the couplings J_i.

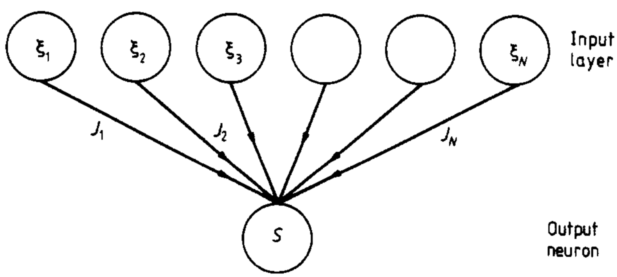

Figure 1. The architecture of a single-layer perceptron consisting of N input neurons ξ_i, each one connected to the output neuron S through the couplings J_i.

† On leave of absence from Instituto de Fisica e Quimica de Sao Carlos, Universidade de Sao Paulo, 13560 Sao Carlos SP, Brazil.

0305-4470/89/160803+06$02.50 © 1989 IOP Publishing Ltd

Given the states of the neurons in the input layer and the coupling strengths, the state of the output neuron is given by

$$S = \mathrm{sgn}\left(\sum_{i=1}^{N} J_i \xi_i\right). \tag{1}$$

The perceptron's task is to learn the mapping between p input patterns $\{\xi_i^\mu, i = 1, \ldots, N; \mu = 1, \ldots, p\}$ and p output states $\{S^\mu, \mu = 1, \ldots, p\}$. To achieve this, there must exist a vector $\boldsymbol{J} = (J_1, J_2, \ldots, J_N)$ such that the p equations

$$S^\mu = \mathrm{sgn}\left(\sum_{i=1}^{N} J_i \xi_i^\mu\right) \qquad \mu = 1, \ldots, p \tag{2}$$

are simultaneously satisfied. If such a vector *does* exist, then the perceptron learning algorithm (Rosenblatt 1962) is guaranteed to converge.

The feasibility of a mapping by a perceptron depends strongly on the statistical properties of the patterns (Minsky and Papert 1969) as well as on the number of patterns (Cover 1965). For random input patterns and large N, Gardner (1988) has found that the maximum number of patterns that can be correctly mapped into their respective outputs is $2N$, a result first derived by Cover (1965). Correlations among the patterns were introduced by considering statistically independent biased patterns

$$\langle S^\mu \rangle = \langle \xi_i^\mu \rangle = m \tag{3}$$

$$\langle \xi_i^\mu \xi_j^\nu \rangle = m^2 + (1 - m^2)\delta_{\mu\nu}\delta_{ij} \tag{4}$$

where $m \in [0, 1]$. These correlations increase the maximum storage capacity of the perceptron (α_c) defined as the ratio between the maximum number of patterns correctly mapped and the number of input neurons N (Gardner 1988). However, this formulation includes only mappings which associate similar input patterns to the same output. The more interesting mappings where similar inputs can be associated with different outputs cannot be studied in the context of biased patterns.

In order to study a more general mapping, we notice that the terms which contain the information about the statistical properties of the patterns in (2) have the form $\zeta_i^\mu \equiv S^\mu \xi_i^\mu$. In this letter we consider mappings where $\boldsymbol{\zeta}_i = (\zeta_i^1, \zeta_i^2, \ldots, \zeta_i^p)$ are distributed according to

$$P(\boldsymbol{\zeta}_i) = \frac{1}{\sqrt{2\pi \det \phi}} \exp(-\tfrac{1}{2}\boldsymbol{\zeta}_i^{\mathrm{T}} \phi^{-1} \boldsymbol{\zeta}_i) \qquad \forall i \tag{5}$$

where the correlation matrix ϕ is given by

$$\phi_{\mu\nu} = \langle \zeta_i^\mu \zeta_i^\nu \rangle = [\tilde{c} + (1 - \tilde{c})\delta_{\mu\nu}]. \tag{6}$$

With this choice of ϕ one can show that $P(\boldsymbol{\zeta}_i)$ exists only if $1 + (p-1)\tilde{c} > 0$, which guarantees that the determinant in (5) is positive. We assume that different sites are uncorrelated; the only correlations we consider are those between different patterns. Moreover, we also assume that the outputs $S^\mu = \pm 1$ are chosen at random. Actually we are making a Gaussian or mean-spherical approximation, $\langle (\xi_i^\mu)^2 \rangle = 1$, of the Ising spins of the original model. Nevertheless, since most of the real-world applications of the perceptrons involve mappings of continuous variables into one of the two classes represented by the output neuron, this Gaussian approximation is attractive by itself. The parameter $\tilde{c}$ which appears in (6) is defined as

$$\tilde{c} = \langle S^\mu S^\nu \xi_i^\mu \xi_i^\nu \rangle \qquad \mu \neq \nu. \tag{7}$$

Hence $\tilde{c}>0$ ($\tilde{c}<0$) corresponds to mappings which lead similar inputs to equal (different) outputs. For $\tilde{c}=0$ one recovers the random mapping. It is well known that the $\tilde{c}<0$ mappings are difficult problems for the perceptrons.

Next we follow Gardner (1988) in calculating the fraction of the phase space of the vectors $\boldsymbol{J}$ which satisfy (2),

$$V=\mathcal{N}^{-1}\int_{-\infty}^{\infty}\prod_i \mathrm{d}J_i \prod_\mu \Theta\left(N^{-1/2}\sum_i J_i\zeta_i^\mu - k\right)\delta\left(\sum_i J_i^2 - N\right) \tag{8}$$

where $\Theta(x)=1$ for $x>0$ and 0 otherwise. The parameter $k\geqslant 0$ ensures that noisy versions of the input patterns are mapped into the same class as the non-corrupted patterns. The spherical constraint $\boldsymbol{J}^2=N$ defines the norm of the vectors $\boldsymbol{J}$ and

$$\mathcal{N}=\int_{-\infty}^{\infty}\prod_i \mathrm{d}J_i\delta\left(\sum_i J_i^2 - N\right) \tag{9}$$

is the volume of the $\boldsymbol{J}$ phase space.

In the thermodynamic limit the sensible physical quantity is $(1/N)\langle\ln V\rangle$ which, as usual, is calculated through the often claimed unreliable, but nevertheless popular, replica trick

$$\langle\ln V\rangle=\lim_{n\to 0}\frac{\langle V^n\rangle - 1}{n}. \tag{10}$$

In the following we assume that the number of input patterns (p) is proportional to N, $p=\alpha N$. Introducing the integral representation of the theta function

$$\Theta\left(N^{-1/2}\sum_i J_i^\rho\zeta_i^\mu - k\right)=\int_k^\infty \frac{\mathrm{d}\lambda_\mu^\rho}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}x_\mu^\rho \exp\left[\mathrm{i}x_\mu^\rho\left(\lambda_\mu^\rho - N^{-1/2}\sum_i J_i\zeta_i^\mu\right)\right] \tag{11}$$

for each pattern μ and each replica ρ, and performing the averages over the *Gaussian* ζ_i^μ

$$\left\langle \exp\left[\mathrm{i}N^{-1/2}\sum_i^N\sum_\mu^p\sum_\rho^n x_\mu^\rho J_i^\rho\zeta_i^\mu\right]\right\rangle=\exp\left[-\frac{1}{2N}\sum_{ij}\sum_{\mu\nu}\sum_{\rho\sigma}x_\mu^\rho x_\nu^\sigma J_i^\rho J_j^\sigma\langle\zeta_i^\mu\zeta_j^\nu\rangle\right] \tag{12}$$

one can write $\langle V^n\rangle$ as

$$\langle V^n\rangle_{\mathrm{av}}=\mathcal{N}^{-n}\int\prod_\rho \mathrm{d}E_\rho\int\prod_{\rho<\sigma}\frac{\mathrm{d}q_{\rho\sigma}\,\mathrm{d}F_{\rho\sigma}}{2\pi/N}\int\prod_\rho\frac{\mathrm{d}r_\rho\,\mathrm{d}R_\rho}{2\pi/p}\exp[N(G_0+\alpha G_1+G_2)] \tag{13}$$

where

$$G_0=-\sum_{\rho<\sigma}q_{\rho\sigma}(F_{\rho\sigma}+c\alpha R_\rho R_\sigma)-\alpha\sum_\rho r_\rho(R_\rho+\tfrac{1}{2}cr_\rho)+\tfrac{1}{2}\sum_\rho E_\rho \tag{14}$$

$$G_1=\ln\left\{\prod_\rho\int_{-\infty}^{\infty}\mathrm{d}x_\rho\int_k^\infty\frac{\lambda_\rho}{2\pi}\exp\left[\sum_\rho x_\rho(\mathrm{i}\lambda_\rho+R_\rho-\tfrac{1}{2}x_\rho)-\sum_{\rho<\sigma}x_\rho x_\sigma q_{\rho\sigma}\right]\right\} \tag{15}$$

$$G_2=\ln\left\{\prod_\rho\int_{-\infty}^{\infty}\mathrm{d}J_\rho\exp\left[-\tfrac{1}{2}\sum_\rho E_\rho J_\rho^2+\sum_{\rho<\sigma}F_{\rho\sigma}J_\rho J_\sigma\right]\right\} \tag{16}$$

with $q_{\rho\sigma}$ and r_ρ defined by

$$q_{\rho\sigma}=\frac{1}{N}\sum_{i=1}^N J_i^\rho J_i^\sigma \tag{17}$$

$$r_\rho=\frac{1}{p}\sum_{\mu=1}^p x_\mu^\rho \tag{18}$$

and, in order to obtain a sensible thermodynamic limit, we have made the rescaling $\tilde{c} = c/p$. In the limit $N \to \infty$ the integrals in (13) may be readily calculated by steepest-descent integration. Assuming that the saddle point is replica symmetric,

$$q_{\rho\sigma} = q \qquad F_{\rho\sigma} = F \qquad \rho < \sigma \tag{19}$$

$$r_\rho = r \qquad R_\rho = R \qquad E_\rho = E \tag{20}$$

the integrations in (15) and (16) are easily performed, so that we finally obtain

$$\frac{1}{N}\langle \ln V \rangle = \alpha \int_{-\infty}^{\infty} \mathrm{D}t \ln\left(\frac{1}{2}\operatorname{erfc}(\Lambda)\right) - \frac{\alpha R^2}{2c(1-q)} + \frac{1}{2}\ln(1-q) + \frac{1}{2}\frac{q}{1-q} \tag{21}$$

where

$$\Lambda = \frac{k - R - q^{1/2}t}{[2(1-q)]^{1/2}} \tag{22}$$

$$\mathrm{D}t = \frac{\mathrm{d}t}{(2\pi)^{1/2}} \exp\left(-\frac{t^2}{2}\right) \tag{23}$$

$$\operatorname{erfc}(x) = 2\int_{\sqrt{2}x}^{\infty} \mathrm{D}t \tag{24}$$

and we have eliminated the saddle-point parameters E, F and r since they are trivially related to q and R, which are then obtained by solving the saddle-point equations

$$R = 2c\left(\frac{1-q}{2\pi}\right)^{1/2} \int_{-\infty}^{\infty} \mathrm{D}t \frac{\exp(-\Lambda^2)}{\operatorname{erfc}(\Lambda)} \tag{25}$$

$$q = 2\alpha\left(\frac{1-q}{2\pi}\right)^{1/2} \int_{-\infty}^{\infty} \mathrm{D}t \frac{\exp(-\Lambda^2)}{\operatorname{erfc}(\Lambda)} (k - R - q^{-1/2}t). \tag{26}$$

Since q, given by (17), is the inner product of two different vectors which solve (2), we expect that at $\alpha = \alpha_c$ (where $V \to 0$) q tends to 1 (Gardner 1988). Hence taking the limit $q \to 1$ in (25) and (26), we obtain

$$\hat{R} + k = c\left[\frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{\hat{R}^2}{2}\right) - \frac{\hat{R}}{2}\operatorname{erfc}\left(\frac{\hat{R}}{\sqrt{2}}\right)\right] \tag{27}$$

$$\alpha_c^{-1} = \frac{1+\hat{R}^2}{2}\operatorname{erfc}\left(\frac{\hat{R}}{\sqrt{2}}\right) - \frac{\hat{R}}{(2\pi)^{1/2}}\exp\left(-\frac{\hat{R}^2}{2}\right) \tag{28}$$

with $\hat{R} \equiv R - k$. Fixing c and k, we can solve (27) for $\hat{R}$ and then use this result in (28) to obtain α_c. Figure 2 shows α_c as a function of c for several values of k. For large, positive c one finds $\alpha_c \simeq c/2$ in agreement with the fact that the perceptron performs well on mappings leading similar inputs to similar outputs. For small c one finds

$$\alpha_c \simeq \begin{cases} 2 - \dfrac{8k}{\sqrt{2\pi}} + \dfrac{4c}{\pi} & k \ll 1 \\[2ex] \dfrac{1+2c}{k^2} & k \gg 1 \end{cases}$$

recovering Gardner's results for $c = 0$. As $c \to -1$, $\hat{R}$ diverges to $-\infty$ and for $c < -1$ there is no real value of $\hat{R}$ which satisfies (27), since the determinant of the correlation matrix appearing in (5) becomes negative.

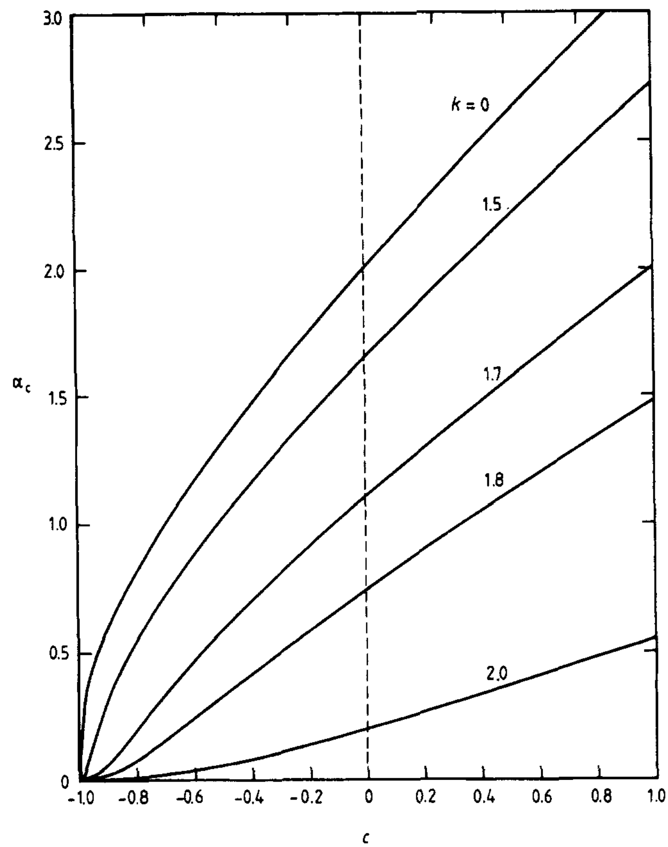

Figure 2. The maximum storage capacity α_c as a function of the correlation parameter c for $k = 0, 1.5, 1.7, 1.8, 2$. Larger values of k increase the robustness of the perceptron to noisy effects in the input patterns.

The mapping with the statistical properties we described can be easily obtained by first generating the ζ_i according to the probability distribution given in (5). The outputs $\{S^\mu = \pm 1\}$ are chosen randomly with equal probability. Knowing ζ_i^μ and S^μ, the input patterns are given by $\xi_i^\mu = \zeta_i^\mu S^\mu$ for every μ and i. This procedure implies that the input patterns are not independent of the output,which is perhaps a more realistic assumption than the commonly assumed independence. Finally we remark that the replica symmetric ansatz for the saddle-point parameters proved to be reliable in the study of the random mapping, $c = 0$ (Gardner 1988), and we believe this result also holds for non-zero c.

We have also extended Gardner's calculations (for random mappings) to the case where all the connections are positive (or negative). We find in this case that the storage capacity decreases from $2N$ to N. This result may be of interest for hardware implementations of the perceptron. Another result we have obtained is that the maximum number of patterns that can be stored in a perceptron with interactions of order x is $p_{\max} = \alpha_c N^x / x!$ where $\alpha_c = 2$ $(x \geqslant 2)$ independent of x, in contrast to the results of Gardner (1987) for the Hopfield model.

Summarising, we have studied the performance of a single-layer perceptron in realising the mapping of random, Gaussian distributed, input patterns into one of the two classes represented by the state of the output neuron. By introducing non-trivial correlations among the patterns we were able to study a family of mappings including easier ones where similar inputs are mapped in the same class and more difficult ones where similar inputs are mapped into different classes. The difficulty of the problem is gauged by the storage capacity of the network, which is higher for the easier problems.

We thank the referee for pointing out a flaw in the original version of the manuscript. The research at Caltech was supported by contract N00014-87-K-0377 from the Office of Naval Research. JFF is partly supported by a CNPq (Brazil) fellowship and RM is supported by a Weizmann fellowship.

References

Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30
Block H D 1962 *Rev. Mod. Phys.* **34** 123
Cover T M 1965 *IEEE Trans. Electron. Comput.* **EC-14** 326
Gardner E 1987 *J. Phys. A: Math. Gen.* **20** 3453
—— 1988 *J. Phys. A: Math. Gen.* **21** 257
Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
Hopfield J J 1982 *Proc. Natl Acad. Sci., USA* **79** 2554
Little W A 1974 *Math. Biosci.* **19** 101
Minsky M L and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)
Opper M 1988 *Phys. Rev.* A **38** 3824
—— 1989 *Europhys. Lett.* **8** 389
Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan Books)