Increased storage capacity for hierarchically structured information in a neural network of Ising type

1989 J. Phys. A: Math. Gen. 22 L1037

(http://iopscience.iop.org/0305-4470/22/21/012)

J. Phys. A: Math. Gen. **22** (1989) L1037-L1041. Printed in the UK

LETTER TO THE EDITOR

Increased storage capacity for hierarchically structured information in a neural network of Ising type

L B Ioffe†, R Kühn and J L van Hemmen

Sonderforschungsbereich 123 an der Universität Heidelberg, D-6900 Heidelberg 1, Federal Republic of Germany

Received 10 July 1989

Abstract. Modelling formal neurons by Ising spins, we describe a simple two-neuron interaction which allows optimal storage capacity for hierarchically structured information. This takes care both of the low-activity limit in simple, Hopfield-type, networks and of the correlations which occur inside the classes of information hierarchies.

The basic idea [1] underlying the extensive recent modelling of neural networks is to focus on collective behaviour through the introduction of an energy function, or Hamiltonian H_N, with suitable symmetric couplings $J_{ij} = J_{ji}$, to simplify the neurons by taking them as formal two-state elements, and to let the system perform a downhill motion in the (free-) energy landscape associated with H_N. As usual, N denotes the size of the system. A pattern can be retrieved if it is near to or coincident with a (local) minimum of H_N. In this way the network can function as an auto-associative (content-addressable) memory.

The patterns themselves are specific random configurations $\{\xi_i^\mu, 1 \leqslant i \leqslant N\}$, which we label by $1 \leqslant \mu \leqslant K$ where K is the total number of stored patterns. (For hierarchies, μ is a multi-index.) In Hopfield-type models, the dynamical variables are Ising spins S_i, $1 \leqslant i \leqslant N$, and the ξ_i^μ are independent random variables which take the values $+1$ and -1 with probability $1-p$ and p, respectively. In the original Hopfield model [1-3], $p = 0.5$ and the fraction of stored patterns $\alpha = K/N$ is bounded by $\alpha_c \simeq 0.14$. In general, $\alpha_c = \alpha_c(p)$ depends on p but is of the same order of magnitude [2]. If $\alpha \leqslant \alpha_c(p)$, then a pattern can be retrieved with relatively small error (a few per cent) whereas for $\alpha > \alpha_c(p)$ no pattern can be retrieved.

For $p = 0.5$, and apart from the small errors which occur in the retrieval, each stored pattern contains N bits and the maximal amount of information per synapse (storage capacity) is α_c bits. In the low-activity limit $p \to 0$, which is what we are interested in here, the information content is reduced to $\alpha_c(p)p|\log_2 p|$. If we want this to be non-zero as $p \to 0$, then $\alpha_c(p)$ should diverge. However, both the Hebbian prescription [1-4] and the non-local quasi-inverse rule [5] give an $\alpha_c(p)$ of the order of one. So in the low-activity limit the storage capacity is bound to vanish.

The same type of problem occurs in any network that has to store and retrieve hierarchically structured information [6-9]. The reason is simply that *inside* each class the individual patterns are strongly correlated.

† Permanent address; Landau Institute for Theoretical Physics, USSR Academy of Sciences, 117940 Moscow.

To accommodate a finite amount of information per synapse, it has been suggested [10-12] that a different representation of the neural activity is necessary. Instead of Ising spins $S_i = \pm 1$ one could, for instance, take occupation numbers V_i which are 0 or 1 (V-representation). If one does so, it can indeed be shown [10, 11] that $\alpha_c(p) \sim 1/p|\ln p|$ and that the theoretical upper bound $\alpha_c^{\max} = (2p|\ln p|)^{-1}$ of Gardner [13] can be saturated. What remains, however, is the intriguing question as to what the interaction looks like for Ising spins (Gardner's method did not allow an explicit representation) and how in the more general case of hierarchically structured information the reduction of internal noise is taken care of.

Using Ising spins, we present a simple two-neuron interaction that gives the appropriate scaling $\alpha_c(p) \sim (p|\ln p|)^{-1}$ as $p \to 0$. Furthermore, we extend our formalism so as to allow hierarchically structured information [6-9]. Since the simplest non-trivial hierarchy is a set of patterns at low activity, we study this case in some detail. We first present a signal-to-noise ratio analysis, then proceed to the calculation of the free energy, and finally generalise our results to a hierarchy with several levels.

Patterns with low activity can be obtained [9] as the second generation in a hierarchy. Let $\bar{\xi} = 1 - 2p$ be the mean value of ξ_i^μ, with $p = \mathrm{Prob}\{\xi_i^\mu = -1\}$. In general, averages over the disorder are denoted by angular brackets. We have $\langle(\xi - \bar{\xi})^2\rangle = \langle\xi(\xi - \bar{\xi})\rangle = 1 - \bar{\xi}^2 = 4p(1-p) \equiv 4\Delta_0$. We now take a 'progenitor' $\xi_0 = \bar{\xi}$ and consider the patterns ξ_i^μ as descendants in the sense of Parga and Virasoro [6]: $\mathrm{Prob}\{\xi_i^\mu | \xi_0\} = \frac{1}{2}(1 + \xi_0/\xi_i^\mu)$. Here $\mathrm{Prob}\{\xi_i^\mu | \xi_0\}$ is the conditional probability of getting ξ_i^μ *given* ξ_0. Then the coupling constants are

$$J_{ij} = N^{-1}\left(1 + \frac{1}{4\Delta}\sum_\mu \delta_i^\mu \delta_j^\mu\right) \tag{1}$$

where $\delta_i^\mu = \xi_i^\mu - \bar{\xi}$ and Δ is at our disposal. For $\Delta = \Delta_0$, we recover [7-9] a model whose retrieval behaviour is just that of the Hopfield model. By varying Δ one can slightly improve [7] $\alpha_c(p)$, though never beyond 0.20, so that in the limit $p \to 0$ the information content per synapse vanishes. To see why this is so physically, we perform a signal-to-noise ratio analysis.

The local field generated by pattern ν at site i is

$$h_i^{(\nu)} = \xi_i^\nu \frac{\Delta_0}{\Delta} + \bar{\xi}\left(1 - \frac{\Delta_0}{\Delta}\right) + \frac{1}{4N\Delta}\sum_{\mu(\neq\nu)} \delta_i^\mu \sum_{j(\neq i)} \delta_j^\mu(\bar{\xi} + \delta_j^\nu). \tag{2}$$

Here we have put $\xi_j^\nu = \bar{\xi} + \delta_j^\nu$. The sum in (2) consists of two parts, to be denoted by $\bar{\xi} f_i^{(1)}$ and $f_i^{(2)}$, and represents the noise produced by the other patterns. It has mean zero and is taken to be Gaussian. The two terms preceding the sum are deterministic and the behaviour of the model would be rather poor as $p \to 0$. One might therefore wish to improve this by tuning $\Delta \approx \Delta_0$. (Note, however, that $\Delta_0 \sim p$.) We have $\langle(f_i^{(1)})^2\rangle \sim p^2\alpha/\Delta^2$, $\langle(f_i^{(2)})^2\rangle \sim p^3\alpha/\Delta^2$, while $\langle f_i^{(1)} f_i^{(2)}\rangle$ vanishes as $N \to \infty$. The tuning $\Delta \approx \Delta_0$ would leave us with an $\alpha_c(p)$ of order one—which is of no help.

The way out is to eliminate the most dangerous noise term, namely $\bar{\xi} f_i^{(1)}$. To this end, we add a (non-local) interaction

$$\Delta J_{ij} = -\frac{1}{4N\Delta}\sum_\mu \left(N^{-1}\sum_k \delta_k^\mu\right)(\delta_i^\mu + \delta_j^\mu). \tag{3}$$

Its effect is (i) to cancel $\bar{\xi} f_i^{(1)}$ and (ii) to produce an extra $-2\alpha(\Delta_0/\Delta)\bar{\xi}$ so that in the limit $p \to 0$ the deterministic part of the local field adds to $\xi_i^\nu(\Delta_0/\Delta) + \bar{\xi}(1 - 2\alpha\Delta_0/\Delta)$.

This suggests that our policy should be to fix $\gamma = \alpha\Delta_0/\Delta$ in such a way that $2\gamma \approx 1$. We will see shortly that this is indeed correct.

We start by introducing the numbers $x^\mu = N^{-1}\Sigma_i\, \delta_i^\mu$ and the order parameters

$$m = N^{-1}\sum_{i=1}^{N} S_i \qquad m_\mu = (4\Delta_0 N)^{-1}\sum_{i=1}^{N} \delta_i^\mu S_i. \tag{4}$$

If the system is in the state $S_i = \xi_i^\nu$, then $m = \bar{\xi}$ and $m_\nu = 1$ as $N \to \infty$ while $m_\mu = 0$ for $\mu \neq \nu$. Inserting (1), (2) and (3) into the Hamiltonian $H_N = -\frac{1}{2}\Sigma_{i,j}\, J_{ij}S_iS_j$ and using (4), we obtain

$$H_N = -\frac{N}{2}\left(m^2 + \frac{1}{4\Delta}\sum_\nu [(4\Delta_0 m_\nu - mx^\nu)^2 - (mx^\nu)^2]\right). \tag{5}$$

We now select finitely many patterns μ whose behaviour we are interested in and treat the remaining ones in a replica-symmetric context [2, 14]. Since the x^ν couple to the ν-patterns only, we may replace $-(4\Delta)^{-1}\Sigma_\nu\,(mx^\nu)^2$ by $-\gamma m^2$. The free energy $f(\beta)$ then turns out to be

$$\begin{aligned}
-\beta f(\beta) = {} & -\frac{\beta}{2}[1-\gamma(1+A^{-1})]m^2 - 2\beta\Delta_0\left(\frac{\Delta_0}{\Delta}\right)\sum_\mu m_\mu^2 \\
& -\frac{\alpha}{2}\left[\ln(A) - \beta\left(\frac{\Delta_0}{\Delta}\right)(q-m^2)A^{-1}\right] - \tfrac{1}{2}\beta^2(q-m^2)r(1-q) \\
& + \left\langle \int \frac{\mathrm{d}z}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) \ln \left[\!\left[2\cosh\left\{\beta\left[\Gamma m + \left(\frac{\Delta_0}{\Delta}\right)\sum_\mu m_\mu \delta^\mu \right.\right.\right.\right. \\
& \left.\left.\left.\left. + \sqrt{(q-m^2)r}\, z\right]\right\}\right]\!\right]\right\rangle
\end{aligned} \tag{6}$$

where $A = [1-\beta(\Delta_0/\Delta)(1-q)]$, $\Gamma = 1-\gamma(1+A^{-1})$ and $r = \gamma(\Delta_0/\Delta)A^{-2}$.

The m, m_μ, and q are determined by the following set of fixed-point equations (see (7) below) and have to be chosen in such a way that they maximise the right-hand side of (6). We have

$$\begin{aligned}
m &= \left\langle\!\left\langle \tanh\left\{\beta\left[\Gamma m + \left(\frac{\Delta_0}{\Delta}\right)\sum_\mu m_\mu\delta^\mu + z\sqrt{(q-m^2)r}\right]\right\}\right\rangle\!\right\rangle \\
m_\mu &= \left\langle\!\left\langle \frac{\delta^\mu}{4\Delta_0} \tanh\left\{\beta\left[\Gamma m + \left(\frac{\Delta_0}{\Delta}\right)\sum_\mu m_\mu\delta^\mu + z\sqrt{(q-m^2)r}\right]\right\}\right\rangle\!\right\rangle
\end{aligned} \tag{7}$$

while $q = \langle\langle \tanh^2\{\beta[\cdots]\}\rangle\rangle$ with the same argument as in (7) describes the noise produced by the extensively many other patterns. The double angular brackets indicate an average over the finitely many ξ^μ and the Gaussian z. Scrutinising the argument of the hyperbolic tangents in (7), one immediately observes the beneficial effects of (3). The 'strength' of the noise generated by the other patterns, i.e. q, is *reduced* to $q - m^2$.

At high temperatures, all order parameters vanish. If one insists on m being $1-2p$, one simply adds an extra external field. The low-T behaviour is more interesting. We focus our attention on a *single* pattern μ. Let $h_+ = \Gamma m + 2p(\Delta_0/\Delta)m_\mu$ and $h_- = h_+ - 2(\Delta_0/\Delta)m_\mu$. Moreover, let $a_\pm = h_\pm/[2(1-m^2)r]^{1/2}$. In the limit $\beta \to \infty$ one finds $q = 1$ and

$$m = (1-p)\,\mathrm{erf}(a_+) + p\,\mathrm{erf}(a_-) \qquad m_\mu = \tfrac{1}{2}[\mathrm{erf}(a_+) - \mathrm{erf}(a_-)]. \tag{8}$$

Good retrieval is guaranteed if and only if $a_+ \gg 0$ and $a_- \ll 0$. Under these conditions, $m_\mu \approx 1$ and $m \approx \bar{\xi}$—as they should. However, this will never happen if Γ stays away from zero since then h_+ and h_- have the same sign ($\Delta_0 \sim p$). One might therefore think that $\Gamma = 0$ would do through tuning γ. It does not, however. The appropriate tuning of γ is obtained by noticing that h_+ and h_- should be of the same order as their difference, namely $2(\Delta_0/\Delta)m_\mu$. Hence we put $\Gamma = 2(1-\varepsilon)(\Delta_0/\Delta)$ with $0 < \varepsilon < 1$ at our disposal. Through this ansatz it can be shown that, as $p \to 0$, a satisfying retrieval remains possible as long as $\alpha \leqslant \alpha_c(p) \sim (2p|\ln p|)^{-1}$. Even at criticality, $A \sim 1$ and the tuning mentioned in the introduction is correct: $2\gamma \approx 1$.

By their very construction, patterns inside a class of a *hierarchy* [6] are also correlated. This correlation is taken care of by a method analogous to the one presented above. It is illustrated most coveniently by a two-level hierarchy but it is by no means restricted to that case.

The first generation consists of ξ_i^λ which are independent and ±1 with equal probability. Given λ, the next generation consists of $\xi_i^{\lambda\mu}$ which are $+\xi_i^\lambda$ with probability $1-p$ and $-\xi_i^\lambda$ with probability p. (This is not a martingale, but all the ξ are equal to ± 1.) An ansatz *à la* equation (1) which allows one to store about $0.2N$ hierarchically correlated patterns is [7, 9]

$$J_{ij} = N^{-1} \sum_\lambda \left(\xi_i^\lambda \xi_j^\lambda + (4\Delta)^{-1} \sum_\mu \delta_i^{\lambda\mu} \delta_j^{\lambda\mu} \right).$$

Here $\delta_i^{\lambda\mu} = \xi_i^{\lambda\mu} - \bar{\xi}_i^{\lambda\mu}$ with $\bar{\xi}_i^{\lambda\mu} = (1-2p)\xi_i^\lambda$ as the conditional expectation of $\xi_i^{\lambda\mu}$ *given* ξ_i^λ. We first sum over $1 \leqslant \mu \leqslant K_2$ and then over $1 \leqslant \lambda \leqslant K_1$, so that $K_{\text{tot}} = K_1 K_2 + K_1$ is the total number of stored patterns and $\alpha = K_{\text{tot}}/N$.

What are the modified coupling constants that take care of the noise produced by the built-in correlations? Though we have developed a systematic procedure to derive suitable ΔJ_{ij}, it is simpler, and perhaps more physical, to derive them by analogy to (5).

So what is the gist of the extra terms in (5)? Suppose the system is in the state $S_i = \xi_i^\mu$. Then the $-mx^\nu$ in (5) reduce the noise generated by the *other* terms with $\nu \neq \mu$, i.e. their influence on the energy surface near μ. To wit, $4\Delta_0 m_\nu - mx^\nu = N^{-1} \sum_i \delta_i^\nu (\xi_i^\mu - m)$ and $m \approx 1-2p$ with $p \to 0$. So only at pN sites do the terms of this sum differ appreciately from zero. Since they are randomly fluctuating, their contribution is of order $\sqrt{pN}$. This gives an extra p which roughly allows for the divergence of $\alpha_c(p)$ as p^{-1}. Note also that we could have dropped the $-(mx^\nu)^2$.

For the hierarchy we define the numbers $x^{\lambda\mu\nu} = N^{-1} \Sigma_i\, \delta_i^{\lambda\mu} \xi_i^\nu$. The order parameters m_λ and $m_{\lambda\mu}$ are obtained directly from (4) through the substitution $1 \to \xi_i^\lambda$ and $\delta_i^\mu \to \delta_i^{\lambda\mu}$. One easily verifies that $4\Delta_0 m_{\lambda\mu} - \sum_\nu m_\nu x^{\lambda\mu\nu}$ does the reduction, provided $K_1 \ll N$. This we henceforth assume. The full-blown Hamiltonian is

$$H_N = -\tfrac{1}{2}N \left[\sum_\lambda m_\lambda^2 + \frac{1}{4\Delta} \sum_{\lambda,\mu} \left(4\Delta_0 m_{\lambda\mu} - \sum_\nu m_\nu x^{\lambda\mu\nu} \right)^2 \right] \tag{9}$$

from which the J_{ij} follow. Sticking to a single 'progenitor' $\xi_0 = 1$, one recovers (3)—except for the $-(mx^\nu)^2$.

In the case of finitely many condensed patterns, to be denoted by μ and (μ, ν), the fixed-point equations corresponding to (7) are $m_\mu = \langle\!\langle \xi^\mu \tanh(\beta B) \rangle\!\rangle$, $m_{\mu\nu} = \langle\!\langle (\delta^{\mu\nu}/4\Delta_0) \tanh(\beta B) \rangle\!\rangle$, and $q = \langle\!\langle \tanh^2(\beta B) \rangle\!\rangle$, where

$$B = (1-\gamma A^{-1}) \sum_\mu m_\mu \xi^\mu + \left(\frac{\Delta_0}{\Delta}\right) \sum_{\mu,\nu} m_{\mu\nu} \delta^{\mu\nu} + z \left[\left(q - \sum_\mu m_\mu^2 \right) r \right]^{1/2}. \tag{10}$$

Also here the highly correlated case ($p \to 0$) is the most interesting one. One may repeat the previous analysis nearly word for word.

In summary, we have shown how strong correlations in hierarchically structured information can be taken care of. The neurons are represented by *Ising* spins which all belong to the *same* network. The optimal storage capacity $\alpha_c(p)$ turns out to be $(2p|\ln p|)^{-1}$.

LBI thanks the members of the neural networks group of the University of Heidelberg for the hospitality extended to him during his stay in Heidelberg, where this work was done. All three authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft.

References

[1] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554
[2] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev. Lett.* **32** 1007; 1987 *Ann. Phys., NY* **173** 30; 1987 *Phys. Rev.* A **35** 2293
[3] van Hemmen J L and Kühn R 1986 *Phys. Rev. Lett.* **57** 913
van Hemmen J L 1987 *Phys. Rev.* A **36** 1959
[4] Hebb D O 1949 *The Organization of Behavior* (New York: Wiley) p 62
[5] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
[6] Parga N and Virasoro M A 1986 *J. Physique* **47** 1857
[7] Feigel'man M V and Ioffe L B 1987 *Int. J. Mod. Phys.* **1** 51
[8] Krogh A and Hertz J A 1988 *J. Phys. A: Math. Gen.* **21** 2211
[9] Bös S, Kühn R and van Hemmen J L 1988 *Z. Phys.* B **71** 261
Bös S 1988 *Diploma Thesis* (Universität Heidelberg)
[10] Tsodyks M V and Fiegel'man M V 1988 *Europhys. Lett.* **6** 101
[11] Buhmann J, Divko R and Schulten K 1989 *Phys. Rev.* A **39** 2689
[12] Horner H 1989 *Z. Phys.* B **75** 133
[13] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[14] van Hemmen J L and Zagrebnov V A 1987 *J. Phys. A: Math. Gen.* **20** 3989