The space of interactions in neural network models

1988 J. Phys. A: Math. Gen. 21 257

(http://iopscience.iop.org/0305-4470/21/1/030)

J. Phys. A: Math. Gen. **21** (1988) 257-270. Printed in the UK

The space of interactions in neural network models

E Gardner

Department of Physics, Edinburgh University, Mayfield Road, Edinburgh EH9 3JK, UK

Received 13 May 1987, in final form 27 July 1987

Abstract. The typical fraction of the space of interactions between each pair of N Ising spins which solve the problem of storing a given set of p random patterns as N-bit spin configurations is considered. The volume is calculated explicitly as a function of the storage ratio, $\alpha = p/N$, of the value $\kappa(>0)$ of the product of the spin and the magnetic field at each site and of the magnetisation, m. Here m may vary between 0 (no correlation) and 1 (completely correlated). The capacity increases with the correlation between patterns from $\alpha = 2$ for correlated patterns with $\kappa = 0$ and tends to infinity as m tends to 1. The calculations use a saddle-point method and the order parameters at the saddle point are assumed to be replica symmetric. This solution is shown to be locally stable. A local iterative learning algorithm for updating the interactions is given which will converge to a solution of given κ provided such solutions exist.

1. Introduction

There has been a lot of recent interest in McCulloch-Pitts (1943) neural networks (Hebb 1949, Little 1974, Hopfield 1982). Analytic results (Amit *et al* 1985a, b, 1987a, b, Kanter and Sompolinsky 1987, Mézard *et al* 1986, Bruce *et al* 1987, Gardner 1986) have been obtained for thermodynamic and dynamical quantities using particular storage prescriptions for the coupling strengths. The storage capacity for the Hopfield model for random patterns is $p = 0.14N$, while the pseudo-inverse (Kohonen 1984, Personnaz *et al* 1985, Kanter and Sompolinsky 1987) stores N linearly independent patterns. For very correlated patterns, each with magnetisation m, where $1 - m \sim \ln N/N$, there is a prescription (Willshaw *et al* 1969, Willshaw and Longuet-Higgins 1970) which stores of the order of $N^2/(\ln N)^2$ patterns. However, the maximum storage capacity of these networks can be larger. In the random case, the maximum number of patterns is $2N$ (Cover 1965, Venkatesh 1986a, b, Baldi and Venkatesh 1987) and we will show that this increases for correlated patterns.

The network is defined as follows. Ising spins, $S_i = \pm 1$, are defined on each site i, $i = 1, \ldots, N$. They are updated according to the rule

$$S_i(t+1) = \operatorname{sgn}(h_i(t) - T_i) \tag{1}$$

where $S_i(t)$ is the Ising spin at time t and the internal magnetic field $h_i(t)$ at time t and site i is given by

$$h_i(t) = \frac{1}{\sqrt{N}} \sum_{j \neq i} J_{ij} S_j(t) \tag{2a}$$

where J_{ij} is the interaction strength for the bond from site j to site i. The interactions J_{ij} and J_{ji} need not in general be equal. The field T_i is a local threshold at the site i

which is fixed in time and the interactions J_{ij} are defined so that

$$\sum_{j\neq i} J_{ij}^2 = N \tag{2b}$$

at each site i. The configuration $\{S_i\}$ is thus a fixed point of the dynamics (1), provided the quantity

$$R_i = S_i(h_i(\{S_j\}) - T_i) \tag{3}$$

is positive on each site i.

This paper follows a recent letter (Gardner 1987a) and will be concerned with the problem of choosing interaction strengths J_{ij} such that $p = \alpha N$ prescribed N-bit spin configurations or patterns,

$$\xi_i^\mu = \pm 1 \qquad \mu = 1, \ldots, p;\, i = 1, \ldots, N$$

will be stored as fixed points of the dynamics defined in (1). It will turn out, however, that the requirement that each pattern is a fixed point is not sufficient to guarantee a finite basin of attraction and the stronger condition

$$\xi_i^\mu(h_i(\{\xi_j^\mu\}) - T_i) > \kappa \tag{4}$$

where κ is a positive constant, will be imposed at each site i and for each pattern μ. Larger values of κ should imply larger basins of attraction.

The quantity of interest will be the density of states or the typical fractional volume of the space of solutions for the couplings $\{J_{ij}\}$ to (2b) and (4) and this will first be calculated. The volume vanishes above a value α_c of α which depends on the stability κ and this determines the maximum storage capacity of the network. Secondly, a local iterative algorithm will be given which will converge to a solution of given κ provided such solutions exist.

In § 2, the volume will be calculated for uncorrelated patterns, where the thresholds T_i are set equal to zero. For $\kappa = 0$, the volume vanishes as α increases towards 2 and this determines the maximum storage capacity in agreement with the known results (Cover 1965, Venkatesh 1986a, b). The upper storage capacity $\alpha_c(\kappa)$ is calculated and decreases with κ. In § 3, the calculation is repeated for patterns with a fixed magnetisation m and it is shown that the storage capacity increases with the correlation m^2 between the patterns and, in particular, that α_c tends to infinity as m tends to 1 (for $\kappa = 0$). The network, therefore, can store more patterns if the constraints (4) are correlated. However, correlated patterns contain less information than random patterns and the information capacity of the network will turn out to decrease slightly with m.

The calculation of the typical fractional volume of the space of interactions $\{J_{ij}\}$ which solve (4) is done by introducing replicas in this space while the prescribed patterns remain quenched. This is the inverse of what is done in the spin-glass problem (Edwards and Anderson 1975, Sherrington and Kirkpatrick 1975) where the interactions are quenched and the spins are allowed to vary. Since all pairs of spins are connected, the fractional volume can be obtained exactly using a saddle-point method. The integration is over variables,

$$M_i^\alpha = \frac{1}{\sqrt{N}} \sum_{j\neq i} J_{ij}^\alpha$$

$$q_i^{\alpha\beta} = \frac{1}{N} \sum_{j\neq i} J_{ij}^\alpha J_{ij}^\beta \qquad \alpha \neq \beta$$

and the replica-symmetric ansatz

$$M_i^\alpha = M \qquad q_i^{\alpha\beta} = q$$

is assumed at the saddle point. The physical interpretation of the order parameter q is similar to that of the Edwards-Anderson order parameter in spin glasses and characterises the typical overlap between pairs of solutions for the couplings. As α increases, different solutions to (4) become more correlated and q increases. In particular, the fractional volume vanishes as q tends to its maximum value which is 1 (by equation (2b)) and the condition $q = 1$ therefore determines α_c. The local stability of the replica-symmetric solution is proved in the appendix.

Since explicit solutions for the optimal J_{ij} are not known, it is necessary to have an algorithm for constructing solutions. In § 4 a local iterative learning algorithm will be defined which is a generalisation of perceptron learning (Rosenblatt 1962, Minsky and Papert 1969) to many threshold functions and to non-zero values of κ necessary in order to obtain finite basins of attraction. The advantage of this kind of algorithm is that a convergence theorem exists. Provided solutions to the problem of storing the patterns with fixed $\kappa > 0$ to equation (4) exist, the algorithms are guaranteed to converge to one such solution.

2. Calculation of the fractional volume of interactions for uncorrelated patterns with zero local threshold

In this section, the threshold T_i in equation (1) will be set equal to zero and the ξ_i^μ will be taken to be random patterns. Since the quantity

$$\prod_{\mu,i} \theta(\xi_i^\mu h_i(\{\xi_j^\mu\}) - \kappa) \tag{5}$$

is one if the patterns can be stored and zero otherwise, the fraction of phase space V_T which satisfies (2b) and (4) is given by

$$V_T = \frac{\int \Pi_{i\neq j}\, \mathrm{d}J_{ij}\, \Pi_{\mu,i}\, \theta(\xi_i^\mu h_i(\{\xi_j^\mu\}) - \kappa)\delta(\Sigma_{j\neq i} J_{ij}^2 - N)}{\int \Pi_{i\neq j}\, \mathrm{d}J_{ij}\, \Pi_i\, \delta(\Sigma_{j\neq i} J_{ij}^2 - N)} \tag{6}$$

for a given realisation of the random patterns $\{\xi_i^\mu\}$. The fractional volume V_T may be written

$$V_T = \prod_{i=1}^{N} V_i$$

where V_i is the fractional volume in the space of interactions $\{J_{ij}\}$ for fixed i. In the thermodynamic limit, we therefore have to study

$$\lim_{N\to\infty} \frac{1}{N} \ln V_T = \frac{1}{N} \sum_i \ln V_i. \tag{7}$$

We now assume that this quantity is self-averaging and it is necessary only to calculate $\langle \ln V\rangle$, the average of $\ln V_i$ over the quenched distribution of the patterns $\{\xi_i^\mu;\ \mu = 1, \ldots, p\}$. This is done using the replica method,

$$\langle \ln V\rangle = \lim_{n\to 0} \frac{\langle V^n\rangle - 1}{n}. \tag{8}$$

The method assumes the validity of the analytic continuation from positive integer to zero values of n. The expectation $\langle V^n \rangle$ is given by

$$\langle V^n \rangle = \left\langle \prod_{\alpha=1}^{n} \int \prod_{j\neq i} \mathrm{d}J_{ij}^{\alpha} \prod_{\mu} \theta\left(\xi_i^{\mu} \sum_{j\neq i} \frac{J_{ij}^{\alpha}}{\sqrt{N}} \xi_j^{\mu} - \kappa\right) \delta\left(\sum_{j\neq i} (J_{ij}^{\alpha})^2 - N\right)\right\rangle \times \left[\prod_{\alpha=1}^{n} \int \prod_{j\neq i} \mathrm{d}J_{ij}^{\alpha}\, \delta\left(\sum_{j\neq i} (J_{ij}^{\alpha})^2 - N\right)\right]^{-1} \tag{9}$$

where $\alpha = 1, \ldots, n$ is the replica index and J_{ij}^{α} is the realisation of the J_{ij} for replica α.

The mean-field calculation of (9) is done by introducing integral representations of the θ functions for each pattern μ and each replica α,

$$\theta\left(\xi_i^{\mu} \sum_{j\neq i} \frac{J_{ij}^{\alpha}}{\sqrt{N}} \xi_j^{\mu} - \kappa\right) = \int_{\kappa}^{\infty} \frac{\mathrm{d}\lambda_{\mu}^{\alpha}}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}x_{\mu}^{\alpha} \exp\left[\mathrm{i}x_{\mu}^{\alpha}\left(\lambda_{\mu}^{\alpha} - \xi_i^{\mu} \sum_{j\neq i} J_{ij}^{\alpha} \xi_j^{\mu}/N^{1/2}\right)\right]. \tag{10}$$

The average over the random patterns ξ_j^{μ} in (9) at sites $j \neq i$ gives

$$\exp\left[\sum_{\mu, j\neq i} \ln\cos\left(\sum_{\alpha} x_{\mu}^{\alpha} J_{ij}^{\alpha}/N^{1/2}\right)\right]. \tag{11}$$

Neglecting terms which are of order $1/N$ relative to the leading term, equation (11) becomes

$$\exp\left[-\tfrac{1}{2}\sum_{\mu} \sum_{\alpha,\beta} x_{\mu}^{\alpha} x_{\mu}^{\beta} \left(\sum_{j\neq i} J_{ij}^{\alpha} J_{ij}^{\beta}/N\right)\right]. \tag{12}$$

The calculation of (9) can be done by introducing a variable $q^{\alpha\beta}$,

$$q^{\alpha\beta} = \frac{1}{N} \sum_{j\neq i} J_{ij}^{\alpha} J_{ij}^{\beta} \qquad \alpha < \beta \tag{13}$$

and a momentum $F^{\alpha\beta}$ conjugate to $q^{\alpha\beta}$, in order to impose the constraint (13). The variable E^{α} will also be introduced for each α in order to impose the constraint (5). $\langle V^n \rangle$ can then be written

$$\begin{aligned}\langle V^n \rangle = \int \prod_{\alpha=1}^{n} \mathrm{d}E_{\alpha} \int \prod_{\alpha<\beta} \frac{\mathrm{d}q_{\alpha\beta}\, \mathrm{d}F_{\alpha\beta}}{(2\pi/N)} & \\ \times \exp\left[N\left(\alpha G_1(q_{\alpha\beta}) + G_2(F_{\alpha\beta}, E_{\alpha}) - \sum_{\alpha<\beta} F_{\alpha\beta} q_{\alpha\beta} + \sum_{\alpha} \tfrac{1}{2} E_{\alpha}\right)\right] & \\ \times \left(\int \prod_{\alpha=1}^{n} \mathrm{d}E_{\alpha} \exp[N(G_2(0, E_{\alpha}) + \tfrac{1}{2} E_{\alpha})]\right)^{-1} &\end{aligned} \tag{14}$$

where

$$G_1(q_{\alpha\beta}) = \ln \prod_{\alpha=1}^{n} \int_{-\infty}^{\infty} \mathrm{d}x_{\alpha} \int_{\kappa}^{\infty} \frac{\mathrm{d}\lambda_{\alpha}}{2\pi} \exp\left(\mathrm{i}\sum_{\alpha} x_{\alpha}\lambda_{\alpha} - \tfrac{1}{2}\sum_{\alpha} x_{\alpha}^2 - \sum_{\alpha<\beta} q_{\alpha\beta} x_{\alpha} x_{\beta}\right) \tag{15}$$

$$G_2(F_{\alpha\beta}, E_{\alpha}) = \ln \prod_{\alpha=1}^{n} \int \mathrm{d}J^{\alpha} \exp\left(-\tfrac{1}{2}\sum_{\alpha} E_{\alpha} J_{\alpha}^2 + \sum_{\alpha<\beta} F_{\alpha\beta} J^{\alpha} J^{\beta}\right) \tag{16}$$

because the integrals over x and λ factorise over the patterns μ and the integrals over the J factorise over the sites j. In the large-N limit $\langle V^n \rangle$ is given by taking the saddle point over the variables $F_{\alpha\beta}$, $q_{\alpha\beta}$ and E_{α} of the function

$$G(q_{\alpha\beta}, F_{\alpha\beta}, E_{\alpha}) = \alpha G_1(q_{\alpha\beta}) + G_2(F_{\alpha\beta}, E_{\alpha}) - \sum_{\alpha<\beta} q_{\alpha\beta} F_{\alpha\beta} + \tfrac{1}{2}\sum_{\alpha} E_{\alpha}. \tag{17}$$

In order to find this saddle point, the replica-symmetric ansatz

$$\begin{aligned} q^{\alpha\beta} &= q && \alpha<\beta \\ F^{\alpha\beta} &= F && \alpha<\beta \\ E^{\alpha} &= E && \text{for all } \alpha \end{aligned} \tag{18}$$

will be assumed. This assumption is reasonable because the space of the solutions to (4) is connected; any solution of (4) can be continuously deformed into any other solution. In the appendix it will be shown that this solution is locally stable.

The saddle-point equations for F and E are algebraic and so these variables can be eliminated and, as n tends to zero, $\langle V^n\rangle$ is given by

$$\langle V^n\rangle = \exp[Nn(\min_q G(q) + \mathrm{O}(1/N))] \tag{19}$$

in the large-N limit where

$$G(q) = \alpha \int \mathrm{D}t \ln H\left(\frac{\sqrt{q}t+\kappa}{(1-q)^{1/2}}\right) + \tfrac{1}{2}\ln(1-q) + \tfrac{1}{2}q/(1-q) \tag{20}$$

$$\mathrm{D}t = \frac{\exp(-\frac{1}{2}t^2)}{(2\pi)^{1/2}}\,\mathrm{d}t \tag{21}$$

$$H(x) = \int_x^\infty \mathrm{D}z. \tag{22}$$

The maximum of G over the variable q is given by the saddle-point equation

$$q = (1-q)\frac{\alpha}{2\pi}\int \mathrm{D}t\left[H\left(\frac{\sqrt{q}t+\kappa}{(1-q)^{1/2}}\right)\right]^{-2}\exp\left(-\frac{(\sqrt{q}t+\kappa)^2}{(1-q)}\right). \tag{23}$$

The physical interpretation of the variable q at the replica-symmetric saddle point can be found by differentiating with respect to F,

$$q = \frac{1}{N}\sum_{j\neq i} J_{ij}^{\alpha} J_{ij}^{\beta} \qquad \alpha<\beta. \tag{24}$$

q is therefore the typical overlap between pairs of solutions to (4) and is similar to the Edwards–Anderson order parameter of spin glasses. As $\alpha \to 0$, $q \to 0$ from equation (23); for $\alpha = 0$ all J_{ij} solve (4) and the typical overlap is equal to the most probable overlap between random pairs of configurations in the space of interactions. As α increases, solutions become more correlated and q increases. As $q \to 1$ the number of solutions tends to zero and the typical volume tends to zero. The upper storage capacity of the network is therefore given by taking the limit $q \to 1$ in equation (23).

As $q \to 1$, the integral in (23) is dominated by values of $t > -\kappa$ and the maximum value of α is given by

$$\alpha_c = \left(\int_{-\kappa}^{\infty} \mathrm{D}t(t+\kappa)^2\right)^{-1}. \tag{25}$$

Taking the limit $\kappa \to 0$ in equation (25) gives $\alpha_c = 2$ in agreement with the known results (Cover 1965, Venkatesh 1986a, b, Baldi and Venkatesh 1987). As the stability κ increases, the constraints (4) become stronger and the optimal value α_c of α decreases. $\alpha_c(\kappa)$ is plotted in figure 1.

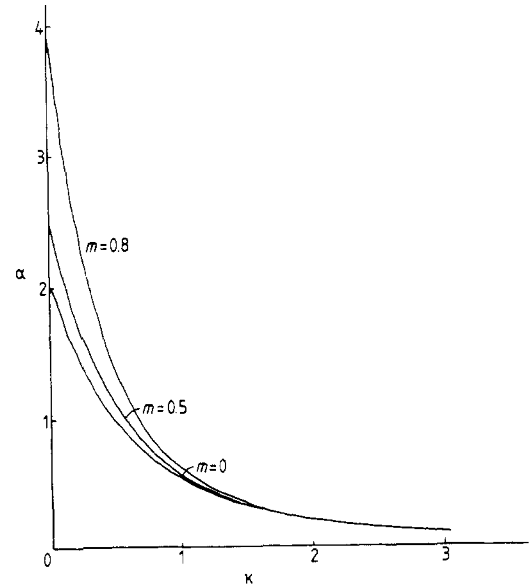

Figure 1. The critical storage ratio, α_c, as a function of κ for values of $m=0$, 0.5 and 0.8.

3. Correlated patterns

In this section, the calculation of § 2 will be repeated for correlated patterns including the local threshold term, T_i. A simple way of imposing a correlation between the patterns is to choose all of them to have the same magnetisation m. The ξ_i^μ are independent random variables with distribution

$$P(\xi_i^\mu)=\tfrac{1}{2}(1+m)\delta(\xi_i^\mu-1)+\tfrac{1}{2}(1-m)\delta(\xi_i^\mu+1). \tag{26}$$

The expectation $\langle V^n\rangle$ of equation (9) can be found by averaging over the distribution (26) and using the integral representation for the θ functions in equation (10). This gives a term

$$\exp\left\{\sum_{\mu,j\neq i}\ln\left[\tfrac{1}{2}(1+m)\exp\left(-\mathrm{i}\sum_{\alpha}\frac{J_{ij}^\alpha}{\sqrt{N}}x_\mu^\alpha\xi_i^\mu\right)+\tfrac{1}{2}(1-m)\exp\left(\mathrm{i}\sum_{\alpha}\frac{J_{ij}^\alpha}{\sqrt{N}}x_\mu^\alpha\xi_i^\mu\right)\right]\right\}. \tag{27}$$

Expansion of the logarithm up to second order in $\Sigma_\alpha J_{ij}^\alpha x_\mu^\alpha/\sqrt{N}$ gives

$$\exp\left[-\mathrm{i}\sum_{\mu\alpha}(mM_\alpha-T)x_\mu^\alpha\xi_i^\mu-\tfrac{1}{2}(1-m^2)\left(\sum_{\mu,\alpha}(x_\mu^\alpha)^2+2\sum_{\alpha<\beta}q_{\alpha\beta}x_\mu^\alpha x_\mu^\beta\right)\right] \tag{28}$$

where $q^{\alpha\beta}$ is given by equation (13) and

$$M_\alpha=\frac{1}{\sqrt{N}}\sum_{j\neq i}J_{ij}^\alpha. \tag{29}$$

Higher-order terms in the expansion vanish as $N\rightarrow\infty$. The constraints (13), (29) and (5) are imposed by introducing order parameters $F^{\alpha\beta}$, K^α, E^α, respectively. In the large-N limit, however, the effect of K^α is of order $1/N$ relative to the other terms.

In this limit, $\langle V^n\rangle$ can be written

$$\lim_{N\to\infty}\frac{1}{N}\ln\langle V^n\rangle$$

$$=\lim_{N\to\infty}\frac{1}{N}\ln\Bigg(\Bigg\{\int\prod_{\alpha=1}^{n}\mathrm{d}M_\alpha\,\mathrm{d}E_\alpha\prod_{\alpha<\beta}\mathrm{d}q_{\alpha\beta}\,\mathrm{d}F_{\alpha\beta}$$

$$\times\exp\bigg[N\bigg(\alpha G_1^1(q_{\alpha\beta},M_\alpha)+G_2(F_{\alpha\beta},E_\alpha)-\sum_{\alpha<\beta}q_{\alpha\beta}F_{\alpha\beta}+\tfrac{1}{2}\sum_\alpha E_\alpha\bigg)\bigg]\Bigg\}$$

$$\times\bigg(\int\prod_{\alpha=1}^{n}\mathrm{d}E_\alpha\,\exp[N(G_2(0,E_\alpha)+\tfrac{1}{2}E_\alpha)]\bigg)^{-1}\Bigg)\tag{30}$$

where G_2 is again given by equation (16) and

$$G_1^1=\ln\bigg\langle\prod_{\alpha=1}^{n}\int_{-\infty}^{\infty}\mathrm{d}x^\alpha\int_\kappa^\infty\frac{\mathrm{d}\lambda^\alpha}{2\pi}\exp\bigg(\mathrm{i}\sum_\alpha x_\alpha[\lambda_\alpha-(mM_\alpha-T)\xi]$$

$$-\tfrac{1}{2}(1-m^2)\sum_\alpha x_\alpha^2-(1-m^2)\sum_{\alpha<\beta}q_{\alpha\beta}x_\alpha x_\beta\bigg)\bigg\rangle\tag{31}$$

where $\langle\ \rangle$ means an average over the variable ξ with the distribution (26).

In the large-N limit, $(1/N)\ln\langle V^n\rangle$ is given by taking the saddle point over the variables $F_{\alpha\beta}$, $q_{\alpha\beta}$, E_α and M_α of the function

$$G(q_{\alpha\beta},M_\alpha,F_{\alpha\beta},E_\alpha)=\alpha G_1^1(q_{\alpha\beta},M_\alpha)+G_2(F_{\alpha\beta},E_\alpha)-\sum_{\alpha<\beta}q_{\alpha\beta}F_{\alpha\beta}+\tfrac{1}{2}\sum_\alpha E_\alpha\tag{32}$$

and the replica-symmetric ansatz (18), together with the condition

$$M_\alpha=M\tag{33}$$

will be assumed in order to find a saddle point. The local stability of the solution is checked in the appendix. Elimination of the variables F and E as in the previous section gives, for the limits $n\to 0$, $N\to\infty$,

$$\langle V^n\rangle=\exp\Big[Nn\Big(\operatorname*{ext}_{M,q}G(q,M,T)+\mathrm{O}(1/N)\Big)\Big]\tag{34}$$

where the extremum ext means a maximum with respect to the variable M and a minimum with respect to the variables q and where

$$G(q,M,T)=G(q,v)$$

$$=\alpha\bigg\{\tfrac{1}{2}(1+m)\int\mathrm{D}t\ln H\bigg[\bigg(\frac{-mv+\kappa}{(1-m^2)^{1/2}}+\sqrt{q}t\bigg)(1-q)^{-1/2}\bigg]$$

$$+\tfrac{1}{2}(1-m)\int\mathrm{D}t\ln H\bigg[\bigg(\frac{mv+\kappa}{(1-m^2)^{1/2}}+\sqrt{q}t\bigg)(1-q)^{-1/2}\bigg]$$

$$+\tfrac{1}{2}\ln(1-q)+\tfrac{1}{2}q/(1-q)\bigg\}\tag{35}$$

and where

$$v=M-T/m.\tag{36}$$

The threshold T can therefore be eliminated. Any local external field can be compensated for by variation of the order parameter M. The physical interpretation of M at the replica-symmetric saddle point is obtained from equation (29) and is the typical ferromagnetic bias in the couplings.

In order to find the storage capacity as a function of α and m, the limit $q \to 1$ is taken in equations (36) and (37). The equation for $\alpha_c(m, \kappa)$ is

$$1 = \alpha_c(m, \kappa)\left[\tfrac{1}{2}(1+m)\int_{(vm-\kappa)/(1-m^2)^{1/2}}^{\infty} \mathrm{D}t\left(\frac{\kappa - vm}{(1-m^2)^{1/2}} + t\right)^2 \right.$$
$$\left. + \tfrac{1}{2}(1-m)\int_{(-vm-\kappa)/(1-m^2)^{1/2}}^{\infty} \mathrm{D}t\left(\frac{\kappa + vm}{(1-m^2)^{1/2}} + t\right)^2\right] \tag{37}$$

where v is given by

$$\tfrac{1}{2}(1+m)\int_{(vm-\kappa)/(1-m^2)^{1/2}}^{\infty} \mathrm{D}t\left(\frac{\kappa - vm}{(1-m^2)^{1/2}} + t\right)$$
$$= \tfrac{1}{2}(1-m)\int_{(-vm-\kappa)/(1-m^2)^{1/2}}^{\infty} \mathrm{D}t\left(\frac{\kappa + vm}{(1-m^2)^{1/2}} + t\right). \tag{38}$$

The storage capacity increases with correlation m, as one would expect, since the constraints in equation (4) become more correlated. In particular, for $\kappa = 0$ and small values of m, (37) and (38) give

$$\alpha_c = 2(1 + 2m^2/\pi + \mathrm{O}(m^4)) \tag{39}$$

and as m tends to 1, α_c diverges as

$$\alpha_c = -\frac{1}{(1-m)\ln(1-m)} \tag{40}$$

for $\kappa = 0$.

For general values of κ and m, equations (37) and (38) can be solved numerically. In figure 1, $\alpha_c(\kappa)$ is plotted for $m = 0$, 0.5, 0.8 and in figure 2, $\alpha_c(m)$ and $v(m)$ are plotted for $\kappa = 0$. It is interesting to compare these optimal results with those of specific

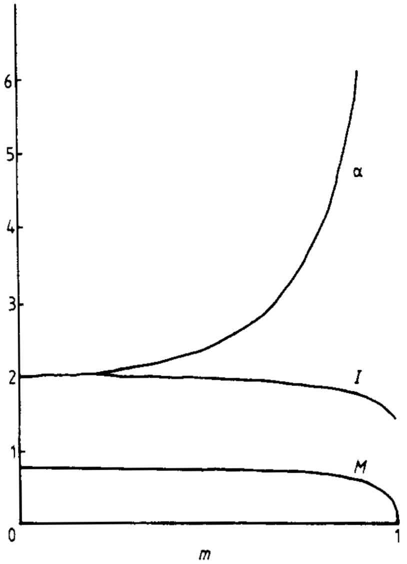

Figure 2. The critical storage ratio α_c, the typical ferromagnetic bias M (for zero thresholds T_i) and the information capacity I as functions of m for $\kappa = 0$.

storage prescriptions for the interactions. The divergence as m tends to 1 in equation (40) is obtained for a model of patterns which are very correlated (Willshaw *et al* 1969, Willshaw and Longuet-Higgins 1970). The number of different spins in the patterns is of order ln N, implying $1-m \sim \ln N/N$ and the storage capacity is of order $N^2/(\ln N)^2$ instead of order N. This relation between m and α agrees with equation (4), although the largest value of the coefficient of $N^2/(\ln N)^2$ is a factor $2(\ln 2)^2 \sim 0.96$ smaller than the optimal result (40) with $\kappa = 0$.

Although the storage capacity increases with the correlation between patterns, the amount of information per pattern decreases. The total information capacity is the total number of bits stored in the patterns

$$I = \frac{N^2}{\ln 2}\,\alpha_c(m)\{\tfrac{1}{2}(1-m)\ln[\tfrac{1}{2}(1-m)] + \tfrac{1}{2}(1+m)\ln[\tfrac{1}{2}(1+m)]\}. \tag{41}$$

For random patterns ($m = 0$) we have

$$I = 2N^2 \tag{42}$$

or twice the number of bonds.

The information capacity I, however, decreases slightly with m. For small m,

$$I = 2N^2[1-(2/\pi - 1/2\ln 2)m^2] = 2N^2(1-0.084m^2) \tag{43}$$

and, as m tends to 1,

$$I = N^2/2\ln 2 = 0.721\mathrm{N}^2. \tag{44}$$

In figure 2, I is plotted as a function of m for $\kappa = 0$.

4. Local iterative learning algorithm

In this section, a local learning algorithm for updating the couplings which, provided solutions to (4) of given κ exist, is guaranteed to converge to one such solution, will be given. The algorithm is a gradient descent and its convergence follows from a generalisation of the perceptron convergence theorem (Rosenblatt 1962, Minsky and Papert 1969). It is a generalisation of the algorithm for $\kappa = 0$ (Wallace 1985, 1986, Bruce *et al* 1986).

The algorithm is defined as follows (for zero thresholds T_i). Let $\{J_{ij}\}$ be any set of couplings with the diagonal term J_{ii} set equal to zero. A mask ε_i^μ is defined at each site i and for each pattern μ.

$$\varepsilon_i^\mu = \theta\left[\kappa\left(\sum_{j\neq i} J_{ij}^2\right)^{1/2} - \sum_{j\neq i} J_{ij}\xi_i^\mu\xi_j^\mu\right] \tag{45}$$

and the couplings are updated according to the rule

$$\Delta J_{ij} = \varepsilon_i^\mu \xi_i^\mu \xi_j^\mu. \tag{46}$$

The algorithm must be done in series over the patterns but can be done either in series or in parallel for the sites and is iterated until ε_i^μ vanishes for each site i and pattern μ. Equation (46) is similar to the Hebb rule (Hebb 1949) except for the presence of the ε_i^μ; changes are made to enhance the recall of pattern μ only at sites which are in error according to condition (4).

The convergence theorem is stated as follows. Suppose a solution J_{ij}^* exists such that

$$\xi_i^\mu \sum_{j\neq i} J_{ij}^*\xi_j^\mu > (\kappa+\delta)\left(\sum_{j\neq i}(J_{ij}^*)^2\right)^{1/2} \tag{47}$$

where δ is some positive number for each pattern μ and each site i. Then the algorithm of (45) and (46) will terminate in a finite number of steps. Before proving the theorem, some notation will be introduced.

The scalar product of a pair of interaction matrices J and U at the site i is defined by

$$(J \cdot U)_i = \sum_{j \neq i} J_{ij} U_{ij} \tag{48}$$

and the norm of J at the site i by

$$\|J\|_i = ((J \cdot J)_i)^{1/2}. \tag{49}$$

Let $\{J_{ij}^{(n)}\}$ be the set of interactions after n applications of (46) at the site i and let $X_i^{(n)}$ be defined by

$$X_i^{(n)} = \frac{(J^{(n)} \cdot J^*)_i}{\|J^{(n)}\|_i \|J^*\|_i}. \tag{50}$$

The theorem will be proved by assuming that the algorithm does not terminate after n steps, and that this requires $X_i^{(n)}$ to become greater than 1 if n is sufficiently large. Since $X_i^{(n)}$ is bounded above by 1, by the Schwarz inequality this is impossible and the algorithm must terminate. At time step n, the numerator of (50) changes to

$$\begin{aligned}\Delta(J^{(n)} \cdot J^*)_i &= \varepsilon_i^\mu \sum_{j \neq i} \xi_i^\mu \xi_j^\mu J_{ij}^* \\ &> \left(\sum_{j \neq i} (J_{ij}^*)^2\right)^{1/2} (\kappa + \delta)\end{aligned} \tag{51}$$

because of equation (47) and, therefore, at time step n the numerator of (50) is bounded below

$$(J^{(n)} \cdot J^*)_i > \|J^*\|_i (\kappa + \delta) n + (J^{(0)} \cdot J^*)_i. \tag{52}$$

The change in the denominator comes from the change in the norm of $J^{(n)}$,

$$\begin{aligned}\Delta(J^{(n)} \cdot J^{(n)})_i &= 2\varepsilon_i^\mu \sum_{j \neq i} J_{ij} \xi_i^\mu \xi_j^\mu + N\varepsilon_i^\mu \\ &< \varepsilon_i^\mu (2\kappa \|J^{(n)}\|_i + N)\end{aligned} \tag{53}$$

since only wrong bits have $\varepsilon = 1$ by equation (45) and so

$$\Delta \|J^{(n)}\|_i < \kappa + N/2\|J^{(n)}\|_i \tag{54}$$

for $\varepsilon = 1$.

Suppose the algorithm has been iterated n times (i.e. $\varepsilon_i^\mu \neq 0$ has occurred n times) and has not terminated. The $X_i^{(n)}$ must be less than one at each step. Therefore, by (52),

$$\|J^{(n)}\|_i > m(\kappa + \delta) + \frac{(J^{(0)} \cdot J^*)}{\|J^*\|_i} \tag{55}$$

for each $m < n$ and so, by (54),

$$\|J^{(n)}\|_i < \kappa n + \frac{N}{2(\kappa + \delta)} \sum_{m=1}^{n-1} \left[m \left(1 + \frac{(J^{(0)} \cdot J^*)_i}{(\kappa + \delta)\|J^*\|_i m} \right) \right]^{-1} + \|J^{(0)}\|_i \tag{56}$$

and so

$$X_i^{(n)} > (\kappa+\delta+\mathrm{O}(1/n))\left(\kappa+\frac{\ln n}{n}\frac{N}{2(\kappa+\delta)}+\mathrm{O}(1/n)\right)^{-1}. \tag{57}$$

Therefore $X_i^{(n)}$ becomes larger than one for sufficiently large n, contradicting the hypothesis that the algorithm does not terminate.

The algorithm (45) and (46) can be generalised to include learning of the local threshold term T_i by defining a new site $i=N+1$ which has spin $\xi_{N+1}^{\mu}=+1$ for all values of μ and letting $J_{i\,N+1}=T_i$.

Another generalisation is to the construction of a symmetric J_{ij}. The change in J_{ij}, equation (46), is replaced by

$$\Delta J_{ij}=(\varepsilon_i^{\mu}+\varepsilon_j^{\mu})\xi_i^{\mu}\xi_j^{\mu}. \tag{58}$$

In this case, the convergence theorem can be proved only if the algorithm is done in parallel in the sites. The proof is similar to that of the asymmetric algorithm except that the scalar product at site i (48) is replaced by

$$J\cdot U=\sum_{\substack{i,j\\ i\neq j}} J_{ij}U_{ij}. \tag{59}$$

5. Conclusions

In this paper, a calculational method has been introduced which allows the maximum storage capacity of neural networks to be determined. In particular, if the patterns are correlated in the sense that they all have an equal magnetisation m, the capacity increases with the correlation between the patterns from $\alpha=2$ for random patterns and diverges as m tends to one.

This increase in capacity allows for the possibility that neural networks can be more efficient than comparison algorithms. If α is restricted to be less than one, as in the Hopfield model or in the pseudo-inverse, the recognition can be done more efficiently by simply comparing the noisy initial vector with each input pattern in order pN steps, whereas one step of parallel iteration in a neural network involves multiplying the $N\times N$ interaction matrix J_{ij} by a vector and involves N^2 steps. In this sense, provided the number of iterations to stability is not too large, neural networks can be more efficient if α is sufficiently larger than one. This relative efficiency therefore increases with the correlation m. Basins of attraction, however, are likely to be smaller in the neural network compared with recognition with nearly 100% noise for comparison algorithms.

Since no explicit expressions for the optimal couplings exist, it is necessary to have a method for constructing them. The algorithms of § 4 are proved to converge to a solution of given κ provided such solutions exist. Other algorithms with convergence theorems similar to those of perceptrons also exist. For example (Gardner *et al* 1987) training with noisy initial vectors can also lead to finite basins of attraction. There are also algorithms like those of § 4 (Krauth and Mézard 1987) and algorithms which are similar but exclude the scaling of κ by the norm of J at the site i (Diederich and Opper 1987). The algorithms are also similar to the back propagation algorithms of Rumelhart *et al* (1985) used in hidden unit models, although in this case no convergence theorem exists.

The methods used in §§ 2 and 3 can be generalised to many other situations. If, for example, one is interested in the storage of patterns allowing for a fraction of the bits to be in error, the upper capacity can be increased (Gardner and Derrida 1988). This can be thought of as an optimisation problem with cost function equal to the total number of wrong bits,

$$f=\frac{1}{N^2}\sum_{i,\mu}\varepsilon_i^\mu \tag{60}$$

where ε_i^μ is defined in equation (45). For $\alpha<2$, the minimum cost function is zero, while for $\alpha>2$ this value increases. It is also possible to generalise to different distributions of the interactions; for example, $J_{ij}=\pm1$ (Gardner and Derrida 1988).

Associative memory and other properties of the learned models can also be determined using similar methods. In particular, the content-addressability as a function of κ has been calculated for a diluted version of the model (Gardner 1987b). In this model finite values of κ do lead to finite basins of attraction whose size increases with the parameter κ. Numerical evidence (Forrest 1988) using the algorithms of § 4 for the fully connected model also suggests that finite values of κ lead to finite content addressability.

There are many other possible generalisations. In particular, the above calculations have been done with asymmetric J_{ij} and it would be interesting to understand the effect of imposing the symmetry $J_{ij}=J_{ji}$ on the interactions. It would also be interesting to generalise the calculations to other properties of typical solutions, to cycles of patterns, (Kanter and Sompolinsky 1986) and to models with hidden units (Rumelhart *et al* 1985).

Acknowledgments

I would like to thank B Derrida, H Gutfreund, D J Willshaw and the Groupe de Physique Theorique at Ecole Normale for useful discussions. I also thank the University of Edinburgh for the award of a Dewar Fellowship.

Appendix

In this appendix we will show that the replica-symmetric solution of §§ 2 and 3 is locally stable. The stability is determined from the signs of the eigenvalues of the matrix of quadratic fluctuations in the $n(n+1)$ variables M^α, $Q^{\alpha\beta}$, E^α and $F^{\alpha\beta}$ at the replica-symmetric saddle point (23), (36) and (37) of equations (17) and (32). Because the solutions are unique in the replica-symmetric subspace, it should be necessary only to consider transverse fluctuations to this space. The eigenfunctions of G_1^1 and G_2 whose eigenvalues are not degenerate with the longitudinal eigenvalues span an $n(n-3)$-dimensional subspace of the full $n(n+1)$-dimensional space and their structure is the same as for the spin-glass problem (de Almeida *et al* 1978). If $\mu_{\alpha\beta}$ and $t_{\alpha\beta}$ are the fluctuations in $q^{\alpha\beta}$ and $F^{\alpha\beta}$, respectively, for $\alpha<\beta$ these eigenfunctions of G_1^1 and G_2 are parallel and are of the form

$$\mu_{\alpha\beta}=\begin{cases}c_1\\ c_2\\ c_3\end{cases}\qquad t_{\alpha\beta}=\begin{cases}d_1 & \alpha=\alpha_0,\beta=\beta_0\\ d_2 & \alpha\text{ or }\beta=\alpha_0\text{ or }\beta_0\\ d_3 & \alpha,\beta\neq\alpha_0,\beta_0\end{cases} \tag{A1}$$

while the fluctuations in M^α and E^α vanish. The values of c_i, d_i, $i=1,2,3$, are chosen so that these eigenfunctions are orthogonal to the degenerate scalar and the vector eigenfunctions and span an $n(n-3)$-dimensional space.

There are therefore two $\frac{1}{2}n(n-3)$-fold degenerate eigenvalues of $\partial^2 G/\partial^2(q, F)$ (equations (17) and (32)) which are eigenvalues of the matrix

$$\begin{pmatrix} P'-2Q'+R' & 1 \\ 1 & P-2Q+R \end{pmatrix} \tag{A2}$$

where

$$\begin{aligned} P &= \frac{\partial^2 G_1^1}{\partial q_{\alpha\beta}\,\partial q_{\alpha\beta}} & P' &= \frac{\partial^2 G_2}{\partial F_{\alpha\beta}\,\partial F_{\alpha\beta}} & & \\ Q &= \frac{\partial^2 G_1^1}{\partial q_{\alpha\beta}\,\partial q_{\alpha\gamma}} & Q' &= \frac{\partial^2 G_2}{\partial F_{\alpha\beta}\,\partial F_{\alpha\gamma}} & & \beta \neq \gamma \\ R &= \frac{\partial^2 G_1^1}{\partial q_{\alpha\beta}\,\partial q_{\gamma\delta}} & R' &= \frac{\partial^2 G_2}{\partial F_{\alpha\beta}\partial_{\gamma\delta}} & & \alpha \neq \gamma, \beta \neq \delta. \end{aligned} \tag{A3}$$

At $\alpha = 0$, the solution to the mean-field equations (23) and (37) is $q=0$, $P'-2Q'+R'=0$ and so the product of the eigenvalues of (A2) is -1. The solution is stable in this limit because it is simply an integral over the phase space of couplings. The sign -1 is due to change of variable $F \to \mathrm{i}F$ in equation (14) from its introduction as the variable conjugate to q. In this limit $\alpha \to \alpha_c$, $q \to 1$ and

$$\begin{aligned} &P-2Q+R \to \frac{\alpha_c}{(1-q)^2}\left(\tfrac{1}{2}(1+m)\int_{(-\kappa+vm)/(1-m^2)^{1/2}}^{\infty} \mathrm{D}t + \tfrac{1}{2}(1-m)\int_{(-\kappa-vm)/(1-m^2)^{1/2}}^{\infty} \mathrm{D}t\right) \\ &P'-2Q'+R' \to (1-q)^2 \end{aligned} \tag{A4}$$

and using equations (37) and (38), the product of eigenvalues is

$$\begin{aligned} &\left[-\kappa(1+m)\int_{(-\kappa+vm)/(1-m^2)^{1/2}}^{\infty}\left(\frac{\kappa-vm}{(1-m^2)^{1/2}}-t\right)\mathrm{D}t\right] \\ &\qquad \times\left[\tfrac{1}{2}(1+m)\int_{(-\kappa+vm)/(1-m^2)^{1/2}}^{\infty}\left(t+\frac{\kappa-vm}{(1-m^2)^{1/2}}\right)^2 \mathrm{D}t\right. \\ &\qquad \left.+\tfrac{1}{2}(1-m)\int_{(-\kappa-vm)/(1-m^2)^{1/2}}^{\infty}\left(t+\frac{\kappa+vm}{(1-m^2)^{1/2}}\right)^2 \mathrm{D}t\right]^{-1} \end{aligned} \tag{A5}$$

which is negative provided κ is positive.

The sign of the product of eigenvalues therefore does not change as q increases from zero to one and α increases from zero to α_c, and the replica-symmetric solution is therefore stable. A vanishing eigenvalue occurs only in the limit $\alpha \to \alpha_c$ and $\kappa \to 0$.

References

de Almeida J R and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983

Amit D J, Gutfreund H and Sompolinsky H 1985a *Phys. Rev. Lett.* **55** 1530

—— 1985*b* *Phys. Rev.* A **32** 1007

—— 1987a *Ann. Phys., NY* **173** 30

—— 1987b *Phys. Rev.* A in press

Baldi P and Venkatesh S 1987 *Phys. Rev. Lett.* **58** 913

Bruce A D, Canning A, Forrest B, Gardner E and Wallace D J 1986 *Proc. Conf. on Neural Networks for Computing, Snowbird, UT (AIP Conf. Proc. 151)* ed J S Denker (New York: AIP) p 65

Bruce A D, Gardner E and Wallace D J 1987 *J. Phys. A: Math. Gen.* **20** 2909

Cover T M 1965 *IEEE Trans. Electron. Comput.* **EC-14** 326

Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949

Edwards S F and Anderson P W 1975 *J. Phys. F: Met. Phys.* **5** 965

Forrest B 1988 *J. Phys. A: Math. Gen.* **21** 245

Gardner E 1986 *J. Phys. A: Math. Gen.* **19** L1047

—— 1987a *Europhys. Lett.* **4** 481

—— 1987b in preparation

Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271

Gardner E, Stroud N and Wallace D J 1987 *Edinburgh preprint 87/394*

Hebb D O 1949 *The Organisation of Behaviour* (New York: Wiley).

Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554

Kanter I and Sompolinsky H 1986 *Phys. Rev. Lett.* **57** 2861

—— 1987 *Phys. Rev.* A **35** 380

Kohonen T 1984 *Self Organisation and Associative Memory* (Berlin: Springer)

Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745

Little W A 1974 *Math. Biosci.* **19** 101

McCulloch W S and Pitts W A 1943 *Bull. Math. Biophys.* **5** 115

Mézard M, Nadal J P and Toulouse G 1986 *J. Physique* **47** 1457

Minsky M L and Papert S 1969 *Perceptrons* (Cambridge, MA: MIT Press)

Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359

Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan Books)

Rumelhart D E, Hinton G E and Williams R J 1985 *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* vol 1, ed D E Rumelhart and J L McClelland (Cambridge, MA: MIT Press)

Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **32** 1792

Venkatesh S 1986a *Proc. Conf. on Neural Networks for Computing, Snowbird, UT (AIP Conf. Proc. 151)* ed J S Denker (New York: AIP)

—— 1986b *PhD thesis* California Institute of Technology

Wallace D J 1985 *Advances in Lattice Gauge Theory* D W Duke and J F Owens (Singapore: World Scientific) p 326

—— 1986 *Lattice Gauge Theory: A Challenge to Large Scale Computing* ed B Bunk and K H Mutter (New York: Plenum) p 313

Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960

Willshaw D J and Longuet-Higgins H C 1970 *Machine Intelligence* **5** 351