Learning in neural networks by eliminating frustrated bonds

J. Phys. A: Math. Gen. 26 (1993) 563–579. Printed in the UK

Learning in neural networks by eliminating frustrated bonds

H J J Jonker† and A C C Coolen‡

† Utrecht Biophysics Research Institute, University of Utrecht, Princetonplein 5, NL 3584 CC Utrecht, The Netherlands
‡ Department of Physics, Theoretical Physics, 1 Keble Road, Oxford OX1 3NP, UK

Received 21 August 1992

Abstract. We study neural network models in which the initial interaction matrix elements are drawn from an arbitrary probability distribution and in which patterns are subsequently stored by eliminating the frustrated bonds, generalizing a proposal by Kinzel. We show that the optimal choice for the *a priori* distribution corresponds to choosing uniform ferromagnetic initial interactions. For the optimal model we study analytically the dynamical behaviour, the equilibrium properties, the sizes of the domains of attraction and a number of information-theoretical performance measures.

1. Introduction

The fundamental principle of storing information in recurrent neural networks is the creation of attractors through the modification of synaptic interactions between the neurons. It was shown by Little [1], Hopfield [2] and Amit *et al* [3] that models of such systems can be studied with statistical-mechanical tools, if one is willing to pay the price of reducing the number of degrees of freedom of individual neurons to a minimum. If the interaction matrix is chosen to be symmetric, then information can be stored in the form of fixed-point attractors; furthermore, equilibrium statistical mechanics will apply. If the interaction matrix is non-symmetric, there will be also the opportunity to have limit-cycle attractors or even more exotic types; equilibrium statistical mechanics, however, will no longer apply. For an overview of the wealth of literature on attractor neural networks we would like to refer to textbooks like [4–6] or the review paper by Abbott [7].

In this paper we study the class of Ising spin attractor neural network models that one obtains by generalizing a proposal by Kinzel [8]: the initial interactions will be drawn independently according to some probability distribution and subsequently patterns are stored by eliminating the frustrated bonds. Selecting individual members of the class corresponds to making a specific choice for the *a priori* distribution; the original Kinzel model [8] is the result of choosing a zero-mean Gaussian (information storage is equivalent to removal of frustrated bonds in a long-range spin glass).

We will show that Kinzel's original choice is not optimal with respect to the maximization of Gardner's stability parameters: the optimal *a priori* distribution we find establishes that, rather than a long-range spin glass, the *tabula rasa* starting point should be a uniform long-range ferromagnet. All models of the generalized Kinzel type have (by construction) two desirable properties: (a) the stability parameters can *never* be negative, whatever the statistical properties of the stored patterns, and (b) the interaction matrices can be constructed by a simple learning rule. The optimal model has the additional advantage of having binary interactions $J_{ij} \in \{0, 1\}$ only.

After the storage process (the elimination of frustrated bonds) the optimal model will be a collection of disconnected ferromagnetic sublattices, therefore we will be able to analyse the system in great detail, both with respect to equilibrium properties (free energy, sizes of domains of attraction) and dynamical behaviour. In addition we will study (in the thermodynamic limit) information-theoretical properties such as the error in the equilibrium state as a function of the error in the initial state and the information gain relative to the maximum information that can be stored in the connection matrix.

2. Generalized Kinzel models

In this section we will generalize a proposal by Kinzel [8] and define dynamics and structure of a class of attractor neural network models in which learning is based on the elimination of frustrated bonds. The network will consist of N binary neurons (or Ising spins) $s_i \in \{-1, 1\}$ which are interconnected via synaptic interactions J_{ij}. The microscopic system state will be denoted by the vector $\boldsymbol{s} \in \{-1, 1\}^N$. The system's dynamics is a stochastic local field alignment, the evolution in time of the microscopic state probability $p_t(\boldsymbol{s})$ is governed by the master equation

$$\frac{\mathrm{d}}{\mathrm{d}t} p_t(\boldsymbol{s}) = \sum_{j=1}^{N} \left[w_j(F_j \boldsymbol{s}) p_t(F_j \boldsymbol{s}) - w_j(\boldsymbol{s}) p_t(\boldsymbol{s}) \right] \tag{1}$$

where F_j denotes the jth spin flip operator, i.e. $F_j \Phi(\boldsymbol{s}) \equiv \Phi(s_1, \ldots, -s_j, \ldots, s_N)$, and the transition rates w_j are defined as usual:

$$w_j(\boldsymbol{s}) \equiv \tfrac{1}{2} \left[1 - \tanh(\beta s_j h_j) \right] .$$

The local alignment fields h_i (or post-synaptic potentials) are given by

$$h_i \equiv \sum_{j \neq i} J_{ij} s_j .$$

The inverse temperature $\beta \equiv T^{-1}$ is a measure of the amount of noise in the system. If the interaction matrix J is symmetric (as will be the case for the models considered in this paper), then the dynamics (1) obeys detailed balance and the (unique) equilibrium probability distribution for the process (1) is the Gibbs distribution

$$p_\infty(\boldsymbol{s}) \sim \mathrm{e}^{-\beta H} \qquad H \equiv -\tfrac{1}{2} \sum_{i \neq j} s_i J_{ij} s_j . \tag{2}$$

In this case equilibrium statistical mechanics applies. At zero temperature the only randomness in the process (1) is in the order of spin updates and consequently the Hamiltonian (2) is a Liapunov function, since for any monotonic function f (with $f' > 0$) of the Hamiltonian the ensemble average according to (1) will decrease as a function of time:

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle f(H) \rangle = \int_0^\infty \mathrm{d}z \, \langle L_z \left[f(H - 2z) - f(H) \right] \rangle \leqslant 0 \qquad L_z(\boldsymbol{s}) \equiv \sum_j \delta \left[z + s_j h_j \right] . \tag{3}$$

In defining the interaction matrix elements $\{J_{ij}\}$ we will generalize a proposal by Kinzel [8] for the storage of a given number p of binary N-bit vectors (or patterns) $\boldsymbol{\xi}^{\mu} \in \{-1,1\}^N$ ($\mu = 1,\ldots,p$). We define the class of generalized Kinzel models to consist of those Ising spin neural network models for which the interaction matrix is generated by the following procedure.

- Take initially all matrix elements J_{ij} equal to K_{ij} which are drawn at random according to a probability distribution $\mathcal{P}(K)$;
- remove all bonds that will be frustrated if the system is in any of the patterns: $J_{ij} \to 0$ if $\exists_{\mu}$ such that $\xi_i^{\mu}\xi_j^{\mu}J_{ij} < 0$.

Individual members of the class are obtained by making a specific choice for the distribution $\mathcal{P}(K)$ (the original Kinzel model [8] corresponds to choosing $\mathcal{P}(K)$ to be a zero-mean Gaussian distribution). The interactions J_{ij} can now be written as

$$J_{ij} = K_{ij}\,\delta_{\boldsymbol{\xi}_i,\mathrm{sgn}(K_{ij})\boldsymbol{\xi}_j}$$

where $\boldsymbol{\xi}_i \equiv (\xi_i^1,\ldots,\xi_i^p)$. By construction the patterns $\boldsymbol{\xi}^{\mu}$ will at least be metastable fixed points (whatever the values of p and N) *for all such models at zero temperature*. This is emphasized by Gardner's stability parameters $\gamma_{i\mu}$ [9]:

$$\gamma_{i\mu} \equiv \frac{\xi_i^{\mu}\sum_{j\neq i} J_{ij}\xi_j^{\mu}}{\sqrt{\sum_{j\neq i} J_{ij}^2}} = \frac{\sum_{j\neq i}|K_{ij}|\delta_{\boldsymbol{\xi}_i,\mathrm{sgn}(K_{ij})\boldsymbol{\xi}_j}}{\sqrt{\sum_{j\neq i}|K_{ij}|^2\delta_{\boldsymbol{\xi}_i,\mathrm{sgn}(K_{ij})\boldsymbol{\xi}_j}}} \geqslant 0\,. \tag{4}$$

3. The optimal *a priori* weight distribution

Our next step consists of proving that, with respect to stabilization of the patterns $\boldsymbol{\xi}^{\mu}$, the optimal *a priori* distribution $\mathcal{P}(K)$ implies binary initial interactions. Upon introduction of the (non-negative) variables ω_{ij} and the average $\langle.\rangle_i$, defined as

$$\omega_{ij} \equiv \frac{\delta_{\boldsymbol{\xi}_i,\mathrm{sgn}(K_{ij})\boldsymbol{\xi}_j}}{\sum_{k\neq i}\delta_{\boldsymbol{\xi}_i,\mathrm{sgn}(K_{ik})\boldsymbol{\xi}_k}} \qquad \langle\Phi\rangle_i \equiv \sum_{j\neq i}\omega_{ij}\Phi_{ij}$$

we can formally write (4) in the form

$$\gamma_{i\mu}^2 = \left\{\sum_{k\neq i}\delta_{\boldsymbol{\xi}_i,\mathrm{sgn}(K_{ik})\boldsymbol{\xi}_k}\right\}\frac{\langle|K|\rangle_i^2}{\langle|K|^2\rangle_i}$$

from which we can deduce:

- $\gamma_{i\mu}^2 \leqslant \sum_{k\neq i}\delta_{\boldsymbol{\xi}_i,\mathrm{sgn}(K_{ik})\boldsymbol{\xi}_k}$ for $i \leqslant N, \mu \leqslant p$.
- The maximum values for the stability parameters are obtained *only* if $|K_{ij}| = |K_{ik}|$ for all $\{j,k\}$ that contribute to the above averages.

Since the variables $\{K_{ij}\}$ are drawn *a priori*, the maximum stabilities are obtained only if the *a priori* probability distribution $\mathcal{P}(K)$ itself obeys $\langle K^2\rangle = \langle|K|\rangle^2$. Apparantly, the optimal choice $\mathcal{P}_{\mathrm{opt}}(K)$ is of the form ($K^* > 0$):

$$\mathcal{P}_{\mathrm{opt}}(K) = \rho\delta[K-K^*] + (1-\rho)\delta[K+K^*] \qquad 0 \leqslant \rho \leqslant 1\,. \tag{5}$$

In this case the stability parameters are given by

$$\gamma_{i\mu} = \gamma_i \equiv \left\{ \sum_{k \neq i} \delta_{\boldsymbol{\xi}_i, \mathrm{sgn}(K_{ik})\boldsymbol{\xi}_k} \right\}^{\frac{1}{2}} .$$

The stability parameters no longer have a pattern index dependence. Equation (5) states that the optimal choice for the *a priori* distribution of interactions implies starting with binary weights (the value of K^* will only set the temperature scale).

Elimination of the remaining freedom in choosing the parameter ρ requires a more detailed specification of the required properties of the distribution of the stability parameters γ_i. We will show that there are at least three sensible measures that are optimized by the choice $\rho = 1$:

$$\Phi_{\mathrm{I}}(\rho) \equiv \left\langle\left\langle \frac{1}{N} \sum_i \gamma_i^2 \right\rangle\right\rangle_{K,\boldsymbol{\xi}} \qquad \Phi_{\mathrm{II}}(\rho) \equiv \left\langle\left\langle \frac{1}{N} \sum_i \gamma_i^4 \right\rangle\right\rangle_{K,\boldsymbol{\xi}}$$

$$\Phi_{\mathrm{III}}(\rho) \equiv -\left\langle\left\langle \frac{1}{N} \sum_i \prod_{j \neq i} \left[1 - \delta_{\boldsymbol{\xi}_i, \mathrm{sgn}(K_{ij})\boldsymbol{\xi}_j}\right] \right\rangle\right\rangle_{K,\boldsymbol{\xi}}$$

where double brackets indicate averaging over both the *a priori* distribution $\mathcal{P}(K)$ of the independent variables $\{K_{ij}\}$ and the distribution $p_{\boldsymbol{\xi}}$ of the independent vectors $\{\boldsymbol{\xi}_i\}$. The physical meaning of Φ_{I} and Φ_{II} is clear. The quantity $-\Phi_{\mathrm{III}}$ represents the expectation value of the fraction of metastable sites (i.e. sites with a zero stability parameter). Of course, presenting these specific measures does not eliminate the possible existence of alternative quantities which are optimal for some $\rho < 1$. If we perform the averages in the above measures and take $\mathcal{P}(K) = \mathcal{P}_{\mathrm{opt}}(K)$ we find

$$\Phi_{\mathrm{I}}(\rho) = (N-1) \sum_{\boldsymbol{\xi}} \left[\tfrac{1}{2}\rho(p_{\boldsymbol{\xi}} - p_{-\boldsymbol{\xi}})^2 + p_{\boldsymbol{\xi}} p_{-\boldsymbol{\xi}}\right]$$

$$\Phi_{\mathrm{II}}(\rho) = \Phi_{\mathrm{I}}(\rho) + (N-1)(N-2) \sum_{\boldsymbol{\xi}} p_{\boldsymbol{\xi}} \left[p_{-\boldsymbol{\xi}}^2 + \rho^2 (p_{\boldsymbol{\xi}} - p_{-\boldsymbol{\xi}})^2\right]$$

$$\Phi_{\mathrm{III}}(\rho) = -\sum_{\boldsymbol{\xi}} p_{\boldsymbol{\xi}} \left[1 - \rho p_{\boldsymbol{\xi}} - (1-\rho) p_{-\boldsymbol{\xi}}\right]^{N-1} .$$

If the pattern distribution obeys $p_{\boldsymbol{\xi}} = p_{-\boldsymbol{\xi}}$ for all $\boldsymbol{\xi} \in \{-1,1\}^p$, then the above quantities do not depend on ρ. In all other cases Φ_{I} and Φ_{II} are maximized only by the choice $\rho = 1$. To find the maximum of Φ_{III} we first define

$$\rho \equiv \tfrac{1}{2}(1+\epsilon) \qquad p_{\boldsymbol{\xi}}^{\pm} \equiv \tfrac{1}{2}\left[p_{\boldsymbol{\xi}} \pm p_{-\boldsymbol{\xi}}\right] .$$

In terms of these new variables we can write Φ_{III} as

$$\Phi_{\mathrm{III}} = -\sum_{n \leqslant N-1,\ \mathrm{even}} |\epsilon|^n \binom{N-1}{n} \sum_{\boldsymbol{\xi}} \left[1 - p_{\boldsymbol{\xi}}^{+}\right]^{N-1-n} p_{\boldsymbol{\xi}}^{+} |p_{\boldsymbol{\xi}}^{-}|^n$$
$$+ \ \mathrm{sgn}(\epsilon) \sum_{n \leqslant N-1,\ \mathrm{odd}} |\epsilon|^n \binom{N-1}{n} \sum_{\boldsymbol{\xi}} \left[1 - p_{\boldsymbol{\xi}}^{+}\right]^{N-1-n} |p_{\boldsymbol{\xi}}^{-}|^{n+1}$$

from which we may conclude that in order to maximize Φ_{III} we must choose $\rho \geqslant \frac{1}{2}$. The exact location of the maximum is not clear in general, but depends both on the choice of N and the pattern distribution. If the distribution of the patterns is such that the moments $x_{\xi} \equiv p_{\xi} N$ are constant in the thermodynamic limit, then for large N the leading contribution to Φ_{III} is

$$\begin{aligned}\Phi_{\mathrm{III}}(\epsilon) &= -\frac{1}{2N} \sum_{\xi} x_{\xi}\, \mathrm{e}^{-\rho x_{\xi}-(1-\rho)x_{-\xi}} + x_{-\xi}\, \mathrm{e}^{-\rho x_{-\xi}-(1-\rho)x_{-\xi}} \\ &= \frac{1}{N} \sum_{\xi} \mathrm{e}^{-x_{\xi}^{+}} \left[|x_{\xi}^{-}| \sinh(\epsilon |x_{\xi}^{-}|) - x_{\xi}^{+} \cosh(\epsilon |x_{\xi}^{-}|) \right]\end{aligned}$$

where $x_{\xi}^{\pm} \equiv [x_{\xi} \pm x_{-\xi}]/2$. Differentiation with respect to ϵ yields

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \Phi_{\mathrm{III}}(\epsilon) = \frac{1}{N} \sum_{\xi} \mathrm{e}^{-x_{\xi}^{+}} |x_{\xi}^{-}| \cosh(\epsilon |x_{\xi}^{-}|) \left[|x_{\xi}^{-}| - x_{\xi}^{+} \tanh(\epsilon |x_{\xi}^{-}|) \right] .$$

Using $\tanh(x) \leqslant x$ $(x \geqslant 0)$ we may conclude that $\Phi(\epsilon)$ is a monotonically increasing function of ϵ provided that $\forall_{\xi} : x_{\xi}^{+} \leqslant 1$ and not all $x_{\xi}^{-} = 0$. Therefore the optimal choice in the scaling regime $p_{\xi}^{+} N \leqslant 1$ $(N \to \infty)$ is choosing $\rho = 1$.

4. Statistical mechanics of the optimal model

We will now study in more detail the optimal model (5) with $\rho = 1$, for which the interaction matrix is $(K^{*} > 0)$

$$J_{ij} = K^{*} \delta_{\xi_i, \xi_j} \tag{6}$$

(note that the above choice (6) for the interaction matrix has in fact also been proposed and briefly discussed in an early paper by van Hemmen and van Enter [10]). Since the matrix (6) is of the form studied in [11], we will make use of the so-called sublattice formalism and introduce a partition of the system into 2^{p} sublattices I_{η}:

$$\{1, \ldots, N\} \equiv \bigcup_{\eta \in \{-1,1\}^{p}} I_{\eta} \qquad I_{\eta} \equiv \{i |\ \xi_i = \eta\} .$$

The number of sites in sublattice I_{η} will be denoted by $|I_{\eta}|$. The sublattice magnetizations m_{η} and the usual order parameters q_{μ} (the so-called overlaps) are given by

$$m_{\eta} \equiv \frac{1}{|I_{\eta}|} \sum_{i \in I_{\eta}} s_i \qquad q_{\mu} \equiv \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu} s_i = \sum_{\eta} \eta_{\mu} \frac{|I_{\eta}|}{N} m_{\eta} .$$

Apart from an irrelevant constant (due to the absence of self-interactions J_{ii}) the Hamiltonian (2) of the model (6) can now be written as

$$H = -\tfrac{1}{2} K^{*} \sum_{\eta} |I_{\eta}|^{2} m_{\eta}^{2}(s) \tag{7}$$

which shows that the sublattices I_η have become independent infinite-range ferromagnets with uniform coupling strength K^*. The stability parameters γ_i of the model (6) therefore depend only on the sizes of the sublattices

$$i \in I_\eta: \qquad \gamma_i = \{|I_\eta| - 1\}^{\frac{1}{2}} .$$

The ground-state energy is obtained if all sublattices are ferromagnetically ordered: $E_0 = \frac{1}{2}K^* \sum_\eta |I_\eta|^2$.

We will now restrict ourselves to the case where p is fixed and the system size N diverges. In order for the Hamiltonian to be extensive the coupling strength K^* must now scale as N^{-1}, since $|I_\eta| = p_\eta N + O(\sqrt{N})$. We will choose

$$K^* \equiv \overline{I}^{-1} \qquad \overline{I} \equiv \frac{1}{N}\sum_\eta |I_\eta|^2 .$$

In the thermodynamic limit $N \to \infty$ the free energy per spin $f[\beta]$ (which is simply the weighted sum of the 2^p individual sublattice free energies) becomes

$$f[\beta] = -\frac{1}{\beta}\log 2 + \sum_\eta \frac{|I_\eta|}{N}\left[\frac{|I_\eta|}{2\overline{I}} m_\eta^2 - \frac{1}{\beta}\ln\cosh\left[\beta \frac{|I_\eta|}{\overline{I}} m_\eta\right]\right] \tag{8}$$

in which the order parameters $\{m_\eta\}$ are the critical points that minimize (8):

$$\forall_\eta : \qquad m_\eta = \tanh\left[\beta\frac{|I_\eta|}{\overline{I}} m_\eta\right] .$$

If p is fixed one can (in the thermodynamic limit) also analyse the dynamical behaviour of the model. Using methods as in [12] or [13], one arrives for $N \to \infty$ at the following deterministic flow equation for the sublattice magnetizations:

$$\frac{\mathrm{d}}{\mathrm{d}t} m_\eta(t) = -m_\eta(t) + \tanh\left[\beta\frac{|I_\eta|}{\overline{I}} m_\eta(t)\right] \tag{9}$$

which leads to the equilibrium solutions

$$m_\eta(\infty) = \mathrm{sgn}[m_\eta(0)] M\left[\beta\frac{|I_\eta|}{\overline{I}}\right] \qquad q(\infty) = \sum_\eta \eta \frac{|I_\eta|}{N}\mathrm{sgn}[m_\eta(0)] M\left[\beta\frac{|I_\eta|}{\overline{I}}\right]$$

where $M[K]$ represents the non-negative solution of the transcendental equation $\tanh(KM) = M$ and where $q \equiv (q_1, \dots, q_p)$. Every sublattice I_η has a critical temperature $T_\eta^c = |I_\eta| \overline{I}^{-1}$ above which $m_\eta(\infty) = 0$.

For the special case $T = 0$ the finite-p flow equations (9) can be solved directly:

$$m_\eta(t) = \mathrm{e}^{-t} m_\eta(0) + (1 - \mathrm{e}^{-t})\mathrm{sgn}[m_\eta(0)]$$

$$q(t) = \mathrm{e}^{-t} q(0) + (1 - \mathrm{e}^{-t})\sum_\eta \eta \frac{|I_\eta|}{N}\mathrm{sgn}[m_\eta(0)] .$$

If p is not finite in the thermodynamic limit but scales with the system size N, the outcome of the $T = 0$ dynamics can still be predicted, using the fact that for $T = 0$

the Hamiltonian (7) is monotonically decreasing to a (local) minimum (see (3)). H is at a (local) minimum if

$$\forall \eta,\ |I_\eta| > 1: \qquad m_\eta(\infty) = \operatorname{sgn}\left[m_\eta(0)\right] . \tag{10}$$

The behaviour of the network at $T = 0$ is quite transparent: from (10) we can conclude that, if the system is prepared in an initial state which is a distorted version of one of the stored patterns, it will reconstruct this pattern completely if the magnetization of each sublattice has the same sign in the initial state as in the pattern to be reconstructed and if there are no sublattices which contain only one spin.

Since the outcome of the zero-temperature dynamics can be expressed in terms of the initial state, we can also study the sizes of the domains of attraction of the stored patterns. There are 2^p sublattices, so if all probabilities p_η are non-zero there will be (for finite p) 2^{2^p} different stable equilibrium states (at the level of sublattice magnetizations). Only $2p$ of these states correspond to stored patterns or their inverses, therefore $2^{2^p} - 2p$ final states are spurious.

There are several ways of measuring the size of the attraction domains. As in [14] one could define the size of the domains of attraction for stored patterns as the fraction f_p of all microstates which will evolve towards one of the stored patterns (or its inverse). If we restrict ourselves to the case where the patterns are drawn at random from an unbiased distribution, then $p_\eta \equiv 2^{-p}$ and the sublattices will in the thermodynamic limit all be roughly of size $N2^{-p}$. For the present model we find for f_p

$$f_p = p2^{1-2^p} .$$

Another measure, due to Cottrell [15], is the Hamming radius of the greatest sphere that can be included in the attraction domain. Calculating this quantity amounts to determining, when starting from a stored pattern, how many spins can be flipped in the worst case, such that the subsequent evolution will still be towards the original stored pattern. Since for the present model a pattern will not be correctly reconstructed if the sign of the magnetization of one or more sublattices has been altered, one deals with the worst situation if the flipped spins *all* belong to the *smallest* sublattice. In this way one arrives at a Hamming radius of

$$\min_\eta \operatorname{int}\left[\frac{|I_\eta| - 1}{2}\right] \tag{11}$$

where int[...] is defined as the integer part. Note that (11) is equal to $2\min_i \gamma_i^2$. If the patterns are drawn at random from an unbiased distribution and if p is finite one finds an expectation value for the Hamming radius of about $\operatorname{int}[N2^{-p-1}]$.

5. Error correction

In this section we will consider the $\rho = 1$ network in the absence of noise ($T = 0$), for large N but arbitrary p. We will study the performance of the network upon choosing an initial state $\boldsymbol{s}(0)$ which is a randomly distorted version of one of the stored patterns, i.e.

$$\operatorname{Prob}[s_i(0)] = (1-r)\delta_{s_i(0),\xi_i^\mu} + r\delta_{s_i(0),-\xi_i^\mu} \tag{12}$$

for some μ. The quantity r is the fractional error in the *initial* state. It is our objective to calculate the expectation value $e(r)$ of the fraction of incorrect spins in the *final* state as a function of r, p, N and the the pattern probability distribution.

Full reconstruction of pattern ξ^μ implies in terms of the overlap order parameters: $q_\rho(\infty) = N^{-1}\sum_\eta |I_\eta|\eta_\rho\eta_\mu$. From the solution of the $T = 0$ dynamics as obtained in the previous section (see (10)) follows the corresponding requirement on the initial state: $\mathrm{sgn}[m_\eta(0)] = \eta_\mu$ for all η. The network will reconstruct pattern ξ^μ without error if and only if the distortion process has not changed the sign of the magnetization of any sublattice. If, on the other hand, the sign of the magnetization of a sublattice I_η has changed, then *all* spins in I_η will be misaligned. Consequently, the total number of incorrect spins in equilibrium equals the number of spins contained in those sublattices for which the magnetization has changed sign due to the mistortion. Let $\overline{E}_\eta$ denote the expectation value of the final number of misaligned spins in sublattice I_η, then the expectation value $\mathrm{e}(r)$ of the fractional error of the system is

$$\mathrm{e}(r) = \frac{1}{N}\sum_\eta \overline{E}_\eta(r)\,. \tag{13}$$

Each term in (13) can, in turn, be written as

$$\overline{E}_\eta(r) = \sum_{Q=0}^{N} Q P_e(Q,r) P_\eta(Q) \tag{14}$$

where $P_\eta(Q)$ is the probability for sublattice I_η to contain Q spins and where $P_e(Q,r)$ is the probability of a magnetization sign change in a sublattice of size Q, due to the distortion (12).

Let us concentrate on $P_e(Q,r)$ first. The probability of R spins changing their state as a result of the process (12) in a lattice of size Q is given by

$$\binom{Q}{R} r^R(1-r)^{Q-R}\,.$$

Since a magnetization sign change occurs if $R > \frac{1}{2}Q$, the probability for this to happen is

$$P_e(Q,r) = \sum_{R=\mathrm{int}[Q/2]+1}^{Q} \binom{Q}{R} r^R(1-r)^{Q-R} + \frac{1}{2}\binom{Q}{Q/2} r^{Q/2}(1-r)^{Q/2}A_Q \tag{15}$$

where $\mathrm{int}[x]$ is the largest integer less than or equal to x and where $A_Q \equiv (1+(-1)^Q)/2$ indicates whether Q is even or odd. In the case where $R = \frac{1}{2}Q$ (Q even), the sublattice magnetization is expected to have a 50% probability of sign change. An exact expression for $P_\eta(Q)$ is harder to find, since the global constraint $\sum_\eta |I_\eta| = N$ is to be satisfied. However, for large N one can take for $P_\eta(Q)$ a binomial distribution (as if the sublattice sizes were independent quantities)

$$P_\eta(Q) = \binom{N}{Q} p_\eta^Q(1-p_\eta)^{N-Q} \tag{16}$$

in which p_η denotes the probability that a randomly chosen spin belongs to sublattice I_η ($p_\eta = 2^{-p}$ for random unbiased patterns). By choosing $P_\eta(Q)$ according to (16) the average value of $\sum_\eta |I_\eta|$ equals N whereas the deviations are of order $\sqrt{N}$ (the hard constraint has been replaced by a soft one). This procedure is thermodynamically equivalent to replacing a canonical ensemble by a grand canonical ensemble. In order to perform the summation in the right-hand side of (14) we introduce some approximations for $P_\eta(Q)$ and $P_e(Q,r)$. We will distinguish two cases: the case where the first moment Np_η of P_η is large (the specific magnitude will be specified later), and the case where Np_η is small.

If the first moment Np_η is large, then $P_\eta(Q)$ can be well approximated by the normal distribution

$$P_\eta(Q) \approx \frac{1}{\sqrt{2\pi}\sigma_\eta} \mathrm{e}^{-(Q-\overline{Q}_\eta)^2/(2\sigma_\eta^2)}$$

where $\overline{Q}_\eta \equiv Np_\eta$ and $\sigma_\eta \equiv \sqrt{Np_\eta(1-p_\eta)}$. A sensible approximation for $P_e(Q,r)$ is somewhat more difficult to obtain. We will express $P_e(Q,r)$ in incomplete Beta-functions, for which good asymptotic approximations exist (see the appendix for details). *If r is not too small ($r > 0.028$) but smaller than $\frac{1}{2}$ then $\overline{E}_\eta$ can be* approximated by

$$\overline{E}_\eta(r) = \frac{1}{2}\sqrt{\frac{\overline{Q}_\eta - \sigma_\eta^2 v(r)}{\pi v(r)}} \cosh\left[\frac{1}{2}v(r)\right] \exp\left[-\left(\overline{Q}_\eta + \frac{1}{18}\right)v(r) + \frac{1}{2}\sigma_\eta^2 v^2(r)\right] \quad (17)$$

where

$$v(r) = \frac{9}{4}\frac{\left[(1-r)^{\frac{1}{3}} - r^{\frac{1}{3}}\right]^2}{(1-r)^{\frac{2}{3}} + r^{\frac{2}{3}}} .$$

If, on the other hand, the first moment $\overline{Q}_\eta \equiv Np_\eta$ is small, we will use a Poisson distribution to approximate $P_\eta(Q)$:

$$P_\eta(Q) = \mathrm{e}^{-\overline{Q}_\eta}\frac{\overline{Q}_\eta^Q}{Q!} .$$

Because $P_\eta(Q)$ vanishes rapidly for Q large with respect to $\overline{Q}_\eta$, $\overline{E}_\eta(r)$ can be approximated by the truncated series

$$\overline{E}_\eta(r) = \mathrm{e}^{-\overline{Q}_\eta}\sum_{k=1}^{20} c_k(r)\overline{Q}_\eta^k \quad (18)$$

where the coefficients c_k are functions of r; namely $c_1(r) = r$, $c_2(r) = r$, $c_3(r) = -r^3 + \frac{3}{2}r^2$, etc.

In figure 1, the Poisson approximation of $\overline{E}_\eta$ has been plotted as a function of $\overline{Q}_\eta$ for different values of r. The shape of this function $\overline{E}_\eta$ illustrates clearly the pattern reconstruction in the ferromagnetic sublattices. If a sublattice contains only a few spins, there is a high probability of its magnetization changing sign due to the distortion. However, the impact of this sublattice on the total performance of the network is small. On the other hand, if a sublattice contains a large number of spins, its impact on the total performance is large, but the probability of a magnetization

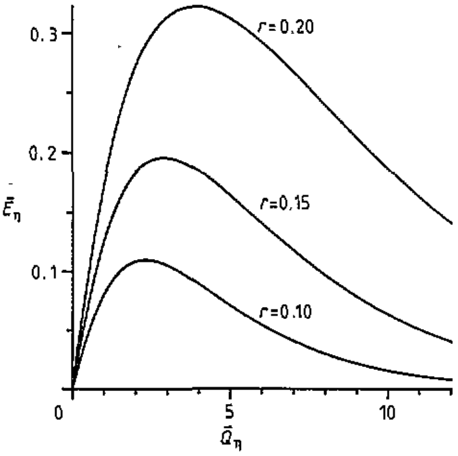

Figure 1. The expectation value of the number of incorrect spins $\overline{E}_{\eta}$ as a function of the average occupation number $\overline{Q}_{\eta}$ for different values of the distortion parameter r.

sign change is small. In both extreme cases (extremely small and extremely large sublattice size) $\overline{E}_{\eta}(r)$ is small; somewhere in between one can distinguish a worst case.

Finally we will now make a choice for the pattern distribution and calculate the expectation value of the fractional error in the reconstructed state explicitly, using the approximations (17) and (18), and compare these predictions with simulation results. We assume that the patterns are randomly drawn from the distribution

$$p_{\xi} = \prod_{\nu=1}^{p} \left[\frac{1+a}{2} \delta_{\xi^{\nu},1} + \frac{1-a}{2} \delta_{\xi^{\nu},-1} \right]$$

which will produce patterns with bias $\lim_{N\to\infty} \frac{1}{N} \sum_i \xi_i^{\mu} = a$ and average mutual overlap $\lim_{N\to\infty} \frac{1}{N} \sum_i \xi_i^{\mu} \xi_i^{\nu} = a^2$ ($\mu \neq \nu$). The probability p_{η} for a randomly drawn spin to belong to sublattice I_{η} can now be written as

$$p_{\eta} = \left(\frac{1+a}{2} \right)^{n} \left(\frac{1-a}{2} \right)^{p-n}$$

where n denotes the number of $+1$-components of the vector η. Since p_{η} depends on the argument η through the value of n only, the same is true for $\overline{E}_{\eta}$ and $\overline{Q}_{\eta}$: $\overline{E}_{\eta} = \overline{E}_{n}$ and $\overline{Q}_{\eta} = \overline{Q}_{n}$. The relative error $e(r)$ can now be calculated by summing over $p+1$ terms:

$$e(r) = \frac{1}{N} \sum_{n=0}^{p} \binom{p}{n} \overline{E}_{n}(r) .$$

In figure 2 the fractional error e (full curve) has been plotted as a function of the pattern bias a for different numbers p of patterns for a network of 4000 spins. The distortion parameter (fractional error in the initial state) was chosen to be $r = 0.2$. For the first moment $\overline{Q}_{n} \leqslant 12$, we used the truncated series (18) for computing $\overline{E}_{n}(r)$. For $\overline{Q}_{n} > 12$ we used (17). These analytical results are in good agreement with the results obtained by performing the actual simulation of the dynamics at

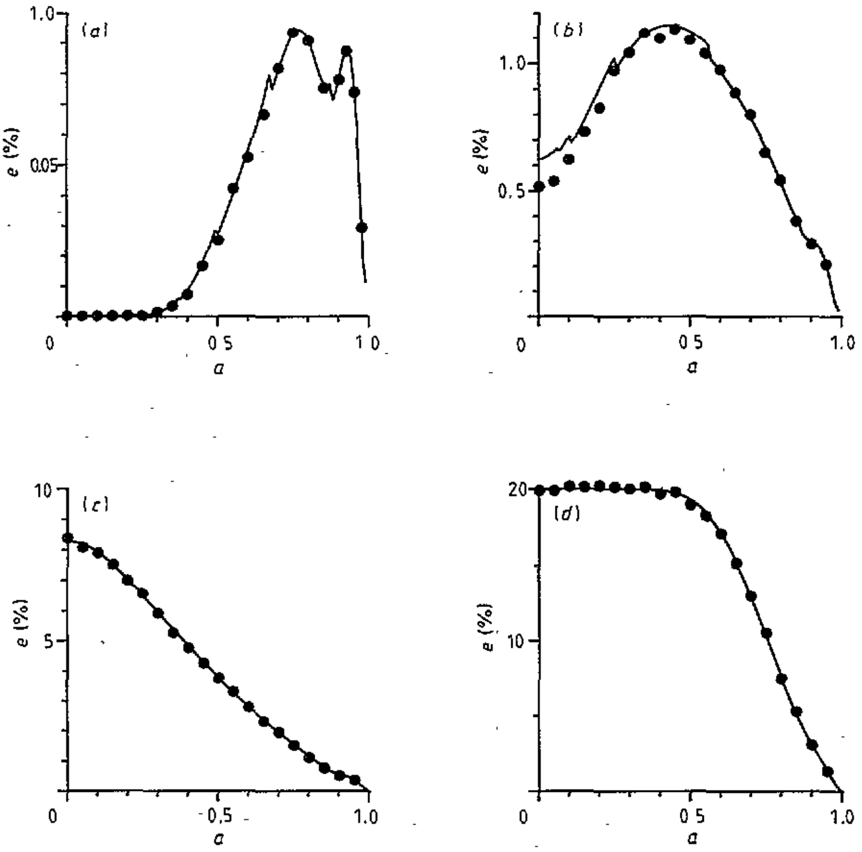

Figure 2. Fractional error ϵ in the final state as a function of the pattern bias a for different values of p; $r = 0.2$, $N = 4000$. Full curves: theoretical predictions; circles: results of 1000 simulations (errors are smaller than marker size) (*a*) $p = 5$, (*b*) $p = 8$, (*c*) $p = 13$, (*d*) $p = 20$.

the spin level (markers). The discontinuities in the full curves are the result of switching from one approximation to the other. In general the Poisson approximation yields better results than the normal approximation (the latter produces a systematic overestimation).

6. Information storage capacity and efficiency

In models of diluted neural networks [16, 17] it is customary to define the storage capacity $\tilde{\alpha}$ as the ratio of the maximal number of patterns to the average connectivity. If the patterns are drawn at random from an unbiased distribution, then in the model of section 4 a neuron is on average connected to a fraction $f \equiv N^{-2}\langle\sum_{ij} J_{ij}/K^*\rangle_\xi \approx 2^{-p}$ of the system. In order to ensure non-vanishing basins of attraction, N must be proportional to 2^p (see (11)): $N = \overline{Q}2^p$ ($\overline{Q}$ is the average number of neurons in a sublattice). For large N, the storage capacity $\tilde{\alpha}$ for the model (6) can now be written as

$$\tilde{\alpha} = -\frac{1}{Q}\,{}^{2}\log f\,.$$

Note the correspondence between this result and a result derived in [18], where, using Gardner calculations, the question was addressed of how to maximize the storage capacity of a network with Ising bonds under the *constraint* that each neuron be coupled to just a fraction f of the system. In the limit $f \to 0$, the storage capacity $\tilde{\alpha}$ turned out also to diverge as $-\,{}^{2}\log f$.

The fact that the storage capacity is not bounded can be understood by realizing that both in the model of [18] and the model of section 4 dilution does not take place in a random way. If connections are eliminated in a non-random way, one actually adds information to the system. In order to arrive at a bounded quantity we will follow an information-theoretic approach.

It has been argued before that not the storage capacity but rather the *information* storage capacity is the relevant quantity of a neural network. For instance, networks storing extremely biased (or sparsely coded) information can store many more patterns than networks storing unbiased (densely coded) information (see e.g. [19–21]). However, the information content of sparsely coded patterns is much less than that of densely coded patterns. Studying the information storage capacity enables one to objectively compare different systems. The additional advantage of having discrete connections is that one can now also calculate the maximal amount of information to be contained in a connection matrix (which provides an upper bound for the information storage capacity). This appeared to be very useful in [21], where, from the outcome of a replica-symmetric calculation exceeding this bound, one could immediately conclude that replica symmetry had to be broken. By studying the information storage capacity relative to the information that can be maximally contained in the connection matrix (the storage ratio) one obtains a measure of the efficiency of the network. However, even this approach does not always give good results. Consider, for instance, the choice $J_{ij} = \delta_{ij}$ with local fields defined as $h_i = \sum_j J_{ij} s_j$. Since now every state is stable, application of the aforementioned method yields a ratio which is obviously much higher than 1. The problem is the lack of a good criterion for the meaning of *storage*. In the previous example the network cannot perform error correction, therefore it seems inappropriate to speak of storage.

A nice general method to cope with such problems is given in [22]. The key idea is not to study the information storage capacity of a network, but rather the network's ability to *gain* information. After all, the task of an attractor neural network is to reconstruct a stored pattern on the basis of a distorted version. Since the initial network state (the distorted version) already contains a certain amount of information, the information a network actually gains is the information present in the final state minus the information present in the initial state. In order to also take into account the efficiency of the network, we will study the average *relative* information gain: the average information gain divided by the information maximally to be contained in the connection matrix

$$i \equiv p\frac{\langle\!\langle \mathrm{Inf(final\ state)} - \mathrm{Inf(initial\ state)}\rangle\!\rangle}{\mathrm{Inf(connection\ matrix)}}\,. \tag{19}$$

In this equation $\mathrm{Inf}(\cdots)$ denotes the function that determines the information associated with its argument, and $\langle\!\langle\cdots\rangle\!\rangle$ denotes averages over distortions and patterns. If one deals with unbiased patterns, the information of the initial state is given by $N(1 - \langle\!\langle S[e_i(r)]\rangle\!\rangle)$, where $e_i(r)$ is the fraction of incorrect spins in the initial state, r is the distortion probability as defined in section 5 and

$$S(f) \equiv -[f \ln f + (1-f)\ln(1-f)]/\ln 2$$

For large N one can apply the central limit theorem and replace $\langle\!\langle S[e_i(r)]\rangle\!\rangle$ by $S(r)$. Taking into account the fractional error $e_f(r)$ in the final state as well, expression (19) reduces to

$$i = pN\frac{S(r) - \langle\!\langle S(e_f(r))\rangle\!\rangle}{\mathrm{Inf(connection\ matrix)}}. \tag{20}$$

The quantity i can be calculated for any network with discrete bonds that stores unbiased patterns (for biased patterns one has to take into account the reduced information content of the patterns, see e.g. [4, 23]). It takes into acount both the information storage capacity and the size of the attraction domains. The trivial model $J_{ij} = \delta_{ij}$ does not provide attraction domains and is therefore not able to gain information, $i = 0$. Storing patterns in a network will in general initially increase i, as p becomes larger; however, the domains of attraction will shrink. This, in turn, has a reducing effect on i since the probability of the initial state being outside the appropriate attraction domain increases. Somewhere in between one may expect a maximum.

In order to calculate i for the network of section 4, we must find the expectation value $\langle\!\langle S(e_f(r))\rangle\!\rangle$ of the information contained in the initial state. Since $e_f(r)$ is a sum over 2^p equally distributed contributions $\overline{E}_\eta(r)$, we can apply the central limit theorem again and replace $\langle\!\langle S[e_f(r)]\rangle\!\rangle$ by $S[e(r)]$, where $e(r)$ is defined by (13). For not too small values of p we can approximate i by

$$i = \frac{2}{1 + 1/(p\ln 2)}\overline{Q}^{-1}\{S(r) - S(\overline{E}(r)/\overline{Q})\} \tag{21}$$

where $\overline{E}(r)$ can be calculated from (17) or (18), depending on the value of $\overline{Q} \equiv N2^{-p}$. In figure 3(*a*), we have plotted the relative information gain i, calculated according to (21), as a function of $\overline{Q}$ for $p = 12$ and different values of r. The markers represent the results of simulations in which i has been determined according to (20). For every r there appears to be a maximum value for i. The absolute maximum can be found at $\overline{Q} = 2.863\ldots$, $r = 0.137\ldots$; in the limit of $p \to \infty$ the corresponding maximal relative information gain i is $0.178\ldots$.

In order to have some kind of reference, we will finally calculate the same quantity for the clipped Hopfield model [16, 24]. Since this network is fully connected, the expression for i now becomes

$$i = 2\alpha\{S(r) - \langle\!\langle S[e_f(r)]\rangle\!\rangle\}$$

where $\alpha \equiv p/N$. Unfortunately it is a very hard problem to determine $\langle\!\langle S[e_f(r)]\rangle\!\rangle$ analytically for the clipped Hopfield model, which is why we have resorted to numerical simulations. The results for $N = 800$ are depicted in figure 3(*b*) for different values of r. The maximum value $i \approx 0.13$ is attained at roughly $\alpha = 0.96\ldots$ for $r = 0.225\ldots$.

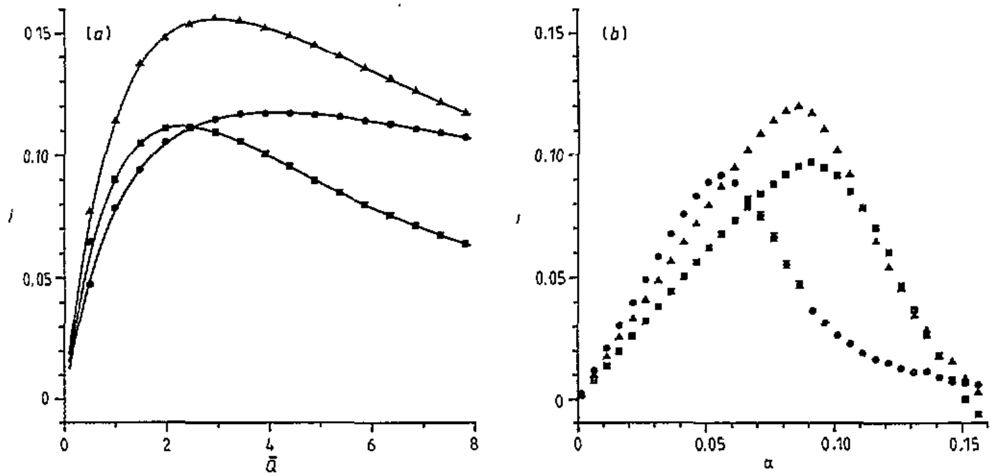

Figure 3. (*a*) Relative information gain i for the model (6) as a function of the average *sublattice size* $\overline{Q} \equiv N2^{-p}$ *for different values of the distortion parameter* r *(12 unbiased* random patterns). Full curves: theoretical predictions according to (21). Markers: results of 1000 simulations (errors are smaller than marker size)—squares, $r = 0.05$; triangles, $r = 0.15$; circles, $r = 0.25$. (*b*) Relative information gain i for the clipped Hopfield model as a function of $\alpha \boxminus p/N$ for different values of r; $N = 800$. Each marker represents an average over 480 simulations: squares, $r = 0.15$; triangles, $r = 0.23$; circles, $r = 0.35$.

7. Discussion

In this paper we have studied the class of Ising spin attractor neural network models one obtains by generalizing a proposal by Kinzel [8]: the initial interactions are drawn independently according to some probability distribution and subsequently patterns are stored by eliminating the frustrated bonds. Selecting individual members of the class corresponds to making a specific choice for the *a priori* distribution; the original Kinzel model [8] is the result of choosing a zero-mean Gaussian. This choice is found not to be optimal with respect to the maximization of Gardner's stability parameters: the optimal *a priori* distribution implies uniform ferromagnetic initial interactions (the strength of which will only set the temperature scale). All models of the generalized Kinzel type have (by construction) two desirable properties: (a) the stability parameters can *never* be negative, whatever the statistical properties of the stored patterns, and (b) the interaction matrices can be constructed by a simple learning rule. The optimal model has the additional advantage of having binary interactions $J_{ij} \in \{0, 1\}$ only.

After the storage process (the elimination of frustrated bonds) the optimal model has become a collection of disconnected ferromagnetic sublattices, therefore it can be analysed in great detail, both with respect to equilibrium properties (free energy, sizes of domains of attraction) and dynamical behaviour. In addition one can calculate (in the thermodynamic limit) information-theoretical quantities such as the error in the equilibrium state as a function of the error in the initial state (for arbitrary p) and the information gain relative to the maximum information that can be stored in the connection matrix.

It is somewhat surprising that a model of such simplicity functions as an associative memory, whatever the choice of patterns. Since only one bit per interaction is needed and since the interaction matrix of the optimal model will in general be very sparse, it is probably easy to implement the network in hardware. However, one should not overestimate the practical value of the model. Although for unbiased random patterns the storage can in principle be efficient (with a larger relative information gain than, for instance, the clipped Hopfield model [16, 24]), the number of neurons should then be proportional to 2^p (which quickly exceeds attainable values). Therefore we believe that the optimal model should be appreciated merely in an academic sense, being in a way complementary to the Willshaw model [19]. Both models employ binary interactions $J_{ij} \in \{0,1\}$; the present model can store efficiently a small number of patterns with a high information content, whereas the Willshaw model can store efficiently a high number of patterns with a low information content. The present model clearly indicates the potential and the restrictions of learning by elimination of frustrated bonds, and because of its tranparency it might serve as a benchmark and as a convenient toy model for testing and illustrating analytical methods.

Appendix. Average alignment in large sublattices

In evaluating $P_e(Q,r)$ (15) we have to distinguish the cases of Q even and Q odd:

$$P_e(2K+1,r) = \sum_{R=K+1}^{2K+1} \binom{2K+1}{R} r^R (1-r)^{2K+1-R}$$

$$P_e(2K,r) = \sum_{R=K+1}^{2K} \binom{2K}{R} r^R (1-r)^{2K-R} + \frac{1}{2}\binom{2K}{K} r^K (1-r)^K .$$

We use the identity [25]

$$\sum_{R=A}^{B} \binom{B}{R} r^R (1-r)^{B-R} = I_r(A, B-A+1)$$

where $I_r(a,b)$ denotes the incomplete beta-function ratio

$$I_r(a,b) \equiv \frac{1}{B(a,b)} \int_0^r t^{a-1}(1-t)^{b-1}\mathrm{d}t \qquad B(a,b) \equiv \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t .$$

The relation [25]

$$I_r(a,b) = \frac{1}{a+b}\left[aI_r(a+1,b) + bI_r(a,b+1)\right]$$

allows us to write $P_e(Q,r)$ as

$$P_e(2K+1,r) = I_r(K+1,K+1) \qquad P_e(2K,r) = I_r(K,K) .$$

For large a and $r < \frac{1}{2}$ we can use the asymptotic form [25] of $I_r(a,a)$:

$$I_r(a,a) \approx \frac{1}{2}\mathrm{erfc}\left[\sqrt{v(r)}\left[\sqrt{2a} - \frac{2}{9}\frac{1}{\sqrt{2a}}\right]\right]$$

where

$$v(r)=\frac{9}{4}\frac{\left[(1-r)^{\frac{1}{3}}-r^{\frac{1}{3}}\right]^2}{(1-r)^{\frac{2}{3}}+r^{\frac{2}{3}}}$$

and the asymptotic expression $\sqrt{\pi}z\exp(z^2)\,\mathrm{erfc}(z)=1+\mathrm{O}(z^{-2})$ to derive

$$P_{\mathrm{e}}(Q,r)\approx\frac{1}{2\sqrt{\pi v(r)Q}}\mathrm{e}^{-v(r)Q}\times\begin{cases}\mathrm{e}^{-\frac{5}{9}v(r)} & \text{if } Q \text{ odd, } Q\to\infty\\ \mathrm{e}^{\frac{4}{9}v(r)} & \text{if } Q \text{ even, } Q\to\infty\,.\end{cases}\qquad\text{(A1)}$$

In order to find an expression for $\overline{E}_\eta$ we substitute for P_η and $P_{\mathrm{e}}(Q,r)$ the normal distribution and the result (A1), respectively, and replace the sum in (14) by an integral:

$$\overline{E}_\eta(r)=\frac{\cosh[\frac{1}{2}v(r)]\mathrm{e}^{-v(r)/18}}{2\pi\sigma_\eta\sqrt{2v(r)}}\int_0^\infty\mathrm{d}Q\sqrt{Q}\mathrm{e}^{-v(r)Q}\exp\left(-\frac{(Q-\overline{Q}_\eta)^2}{2\sigma_\eta^2}\right).\qquad\text{(A2)}$$

In deriving this final result, we have assumed that the integral is dominated by the large-Q contributions. A self-consistency check, which can be performed by determining the point where the integrand of (A2) actually attains its maximum value, yields the condition $v(r)\leqslant 1$. This means that (A2) is only valid if r is larger than approximately 0.028.

Finally one can rewrite (A2) such that the integral depends on one parameter only.

$$\overline{E}_\eta(r)=\frac{1}{2\pi}\sqrt{\frac{\sigma_\eta\sqrt{2}}{v(r)}}\cosh\left[\frac{1}{2}v(r)\right]\mathrm{e}^{-\frac{1}{18}v(r)}\exp\left(-\frac{1}{2\sigma_\eta^2}\overline{Q}_\eta^2\right)g(x)\qquad\text{(A3)}$$

where

$$x\equiv\frac{1}{\sqrt{2}\sigma_\eta}[\overline{Q}_\eta-\sigma_\eta^2v(r)]\qquad g(x)\equiv\mathrm{e}^{x^2}\int_0^\infty\mathrm{d}t\sqrt{t}\mathrm{e}^{-[t-x]^2}\,.$$

The integral can be evaluated numerically, but an upper bound for (A3) can be obtained by replacing the square root in the integral by the tangent $h(t)=(t+t^*)/2\sqrt{t^*}$ that touches at the point $t=t^*$, for any $t^*>0$ (since $\forall\, t\geqslant 0:\ h(t)\geqslant\sqrt{t}$):

$$g(x)\leqslant\frac{1}{4\sqrt{t^*}}[1+\sqrt{\pi}(x+t^*)(1+\mathrm{erf}[x])\mathrm{e}^{x^2}]\,.\qquad\text{(A4)}$$

Minimizing the upper bound (A4) with respect to t^* yields the best value for t^*:

$$t^*=x+[\sqrt{\pi}(1+\mathrm{erf}[x])\mathrm{e}^{x^2}]^{-1}\,.$$

Apparently for large Q a good approximation for (A2) is (the relative error with respect to (A2) being smaller than 4.5×10^{-4} for $Q>12$)

$$\overline{E}_\eta(r)=\frac{1}{2}\sqrt{\frac{\overline{Q}_\eta-\sigma_\eta^2v(r)}{\pi v(r)}}\cosh\left[\frac{1}{2}v(r)\right]\exp\left[-\left(\overline{Q}_\eta+\frac{1}{18}\right)v(r)+\frac{1}{2}\sigma_\eta^2v^2(r)\right].$$

References

[1] Little W A 1974 *Math. Biosci.* **19** 101–20

[2] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554–8

[3] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007–18

[4] Amit D J 1989 *Modelling Brain Function* (Cambridge: Cambridge University Press)

[5] Müller B, Reinhardt J 1990 *Neural Networks, an Introduction* (Berlin: Springer)

[6] Hertz J, Krogh A, Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)

[7] Abbot L F 1990 *Network* **1** 105–22

[8] Kinzel W 1985 *Z. Phys. B: Condens. Matter* **60** 205–13

[9] Gardner E 1988 *J. Phys. A: Math. Gen* **21** 257–70

[10] van Hemmen J L and van Enter A C D 1986 *Phys. Rev.* A **34** 2509–12

[11] van Hemmen J L, Kühn R 1986 *Phys. Rev. Lett.* **57** 913–6

[12] Coolen A C C and Ruijgrok Th W 1988 *Phys. Rev.* A **38** 1007–18

[13] van Hemmen J L and Kühn R 1991 *Models of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer)

[14] Coolen A C C, Jonker H J J and Ruijgrok Th W 1989 *Phys. Rev.* A **40** 5295–8

[15] Cottrell M 1988 *Biol. Cybern.* **58** 129–39

[16] Sompolinsky H 1986 *Phys. Rev.* A **34** 2571–4

[17] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167–73

[18] Bouten M, Kommoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 2605–12

[19] Willshaw D J, Buneman O P and Longuet-Higgins H C 1969 *Nature* **222** 960–2

[20] Tsodyks M V and Feigel'man M V 1988 *Europhys. Lett.* **6** 101–5

[21] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271–84

[22] Palm G 1991 *Concept. in Neurosci.* **2** 97–128

[23] Nadal J P and Toulouse G 1990 *Network* **1** 61–74

[24] van Hemmen J L 1987 *Phys. Rev.* A **36** 1959–62

[25] Abramowitz M and Stegun I A 1970 *Handbook of Mathematical Functions* (New York: Dover)