Quenched versus annealed dilution in neural networks

1990 J. Phys. A: Math. Gen. 23 4643

(http://iopscience.iop.org/0305-4470/23/20/025)

J. Phys. A: Math. Gen. 23 (1990) 4643-4657. Printed in the UK

Quenched versus annealed dilution in neural networks

M Bouten†, A Engel‡, A Komoda† and R Serneels†

† Limburgs Universitair Centrum, Universitaire Campus, 3610 Diepenbeek, Belgium
‡ Sektion Physik der Humboldt Universität, Bereich 04, Invalidenstr. 42, Berlin, 1040, Federal Republic of Germany (formerly German Democratic Republic)

Received 14 May 1990

Abstract. The capacity for storing random patterns in a diluted neural network is determined following the method of Gardner. Two types of dilution are considered. In the quenched case, the broken couplings are chosen at random and are independent of the stored patterns. By contrast, in the annealed case, the disconnected couplings are selected in order to optimize the storage of the patterns. By the same token, the vanishing couplings are strongly correlated with the stored patterns. We also determine the distribution of the synaptic strengths. This distribution illustrates the difference between quenched and annealed dilution most clearly.

1. Introduction

The statistical physics of neural networks deals with systems of N formal neurons denoted by spin variables $S_i = \pm 1$, $i = 1, \ldots, N$, interacting via synaptic couplings J_{ij}. In the thermodynamic limit $N \to \infty$, these networks show emergent properties for information processing, in particular they can function as associative memories. This means that, given a set of p N-bit words $\{\xi_i^\mu\}$, $\xi_i^\mu = \pm 1$, $i = 1, \ldots, N$, $\mu = 1, \ldots, p$, one can choose a synaptic matrix J_{ij} in such a way that the system relaxes to one of its patterns when an incomplete or noisy version of that pattern is presented to it as an initial condition. Many interesting static and dynamic properties of neural network models have been studied during recent years. An impressive account of the current activity in this field is provided by a recent special issue of this journal (*J. Phys. A: Math. Gen.* **22** (1989) number 12).

The methods of statistical mechanics have been applied mainly to fully connected neural networks, in which each neuron interacts with every other neuron, i.e. $J_{ij} \neq 0$ for all $i \neq j$. The full connectivity allows the determination of the partition function by the saddle-point method. In the present paper, we study diluted models in which each neuron is connected to only a fraction of the other neurons. There are several reasons why the investigation of diluted networks should be interesting. First, the connectivity of biological networks, though high, is far from complete. The human brain, for example, consists of 10^{10} to 10^{11} neurons, each one connected to about 10^4 others. Diluted networks have been used to study such locally connected architectures (Canning and Gardner 1988, Noest 1989). Secondly, dilution is an attractive tool to study the robustness of the cooperative behaviour of neural networks against malfunctioning of some of the elements. For biological networks as well as for hardware realizations, it is desirable that the system function does not break down if only a few elements deteriorate in their performance. As a matter of fact, this kind of robustness is often advanced to advocate the superiority of neural network architectures over more traditional computing systems. Finally, the dynamics of fully connected network models is characterized by strong internal feedback and by complicated correlation loops (Gardner *et al* 1987). This makes an exact analytical treatment of the dynamical behaviour impossible. In order to determine, e.g., the basin of attraction, which is one of the most interesting quantities characterizing an associative memory, one has to rely on numerical methods. Dilution reduces the internal feedback and, in the extreme case where only o(ln N) synapses remain per neuron, it allows a complete analytical description of the dynamics (Derrida *et al* 1987) including the determination of the basin of attraction (Gardner 1989).

There exist different types of diluted network models depending on the procedure used for disconnecting couplings in the fully connected network. The different types can be grouped in two classes which will be called quenched and annealed† dilution. In the case of quenched dilution, a certain fraction of all synapses is cut at random. This means, more specifically, that the cutting of connections is totally independent of the patterns that are stored or have to be stored in the network. Previous calculations for diluted networks have mainly considered quenched dilution (Derrida *et al* 1987, Gardner 1989, Sompolinsky 1986, Van Hemmen 1987). In the case of annealed dilution, the choice of severed connections is not random at all. The zero couplings are selected in a way which takes account of the patterns to be stored and they help to achieve good storage of these patterns. The chopper model (Kinzel 1985, Van Hemmen and Van Enter 1986) and the three-synapses model of Sompolinsky (1987) and of Van Hemmen (1987) are examples of annealed dilution. In a recent paper, Bouten *et al* (1990) determined the storage capacity of the diluted network with Ising couplings in the case of annealed dilution.

The present paper is organized as follows. In Section 2, we use the techniques of Gardner (1988) to determine the storage capacity of a diluted network in the case of annealed dilution. The case of quenched dilution is studied in section 3. Here we consider two types of dilution, depending on whether the random cutting of connections is done before or after the learning process. In section 4, we determine the distribution of the coupling coefficients both in the case of annealed and of quenched dilution. The results are discussed in the last section.

2. Storage capacity of diluted networks: annealed case

In this section we study the storage properties of a neural network under the constraint that the fraction $(1-f)$ of all connections into every neuron should be cut. The choice of vanishing coupling coefficients should be made selectively with the aim of optimizing the storage capacity of the network.

The calculation of the storage capacity in the annealed case will be done following the original method of Gardner (1988). To simplify our notation, we consider a network with $N+1$ neurons and focus our attention on the neuron $i=0$. We use the shorthand notation

$$J_{0j}=J_j \qquad \xi_0^\mu \xi_j^\mu = \eta_j^\mu . \tag{1}$$

The coupling coefficients J_j have to satisfy three conditions. To express these conditions,

† We are indebted to a referee of a previous paper (Bouten *et al* 1990) for suggesting this terminology.

it is convenient to write them as a product of two factors

$$J_j = c_j T_j \tag{2}$$

with $c_j = 0$ or 1 and T_j a real number. The first condition

$$\sum_{j=1}^{N} c_j = fN \tag{3}$$

fixes the degree of dilution ($f \in [0, 1]$). The second condition

$$\sum_{j=1}^{N} J_j^2 = \sum_{j=1}^{N} c_j T_j^2 = fN \tag{4}$$

sets the scale for the remaining non-zero bonds and ensures that J_j is of order 1 on average. Finally, the third condition

$$\frac{1}{\sqrt{Nf}} \sum_{j=1}^{N} J_j \eta_j^\mu > K \qquad \mu = 1, \ldots, p \tag{5}$$

expresses the fact that the coupling coefficients J_j must store the p patterns $\{\xi_i^\mu\}$ as fixed-point attractors of the neural network dynamics. These stability conditions contain the stability parameter $K\,(>0)$ which controls the size of the basins of attraction (Krauth *et al* 1988, Kepler and Abbott 1988, Forrest 1988).

Following Gardner's method, one has to calculate the volume in the space of interactions where the three conditions (3)-(5) are satisfied. Using the form (2), this necessitates summing over the two possible values of each c_j and integrating over all values of every T_j. The normalization condition (4) restricts the domain of integration of those T_j for which the corresponding c_j is equal to 1. Since the other T_j (with corresponding $c_j = 0$) do not occur in the conditions (3)-(5), they are completely unrestrained and will yield divergent integrals. We must therefore add an extra constraint in order to make all integrals convergent. A convenient choice could be

$$\sum_{j=1}^{N} (1 - c_j) T_j^2 = (1 - f)N \tag{6}$$

which is the analogue of (4). An alternative choice consists in introducing a simple convergence factor $\exp[-\frac{1}{2}\sum_j (1 - c_j) T_j^2]$ inside the integrals. We will use this latter approach, but it is easy to verify that both procedures are equivalent.

We now follow Gardner (1988) in calculating the fractional volume in the space of interactions where the conditions (3)-(5) are satisfied. The conditions (4) and (5) are the same as in Gardner's paper except for a trivial factor f. The new condition (3) is easily incorporated by appending a Kronecker delta to Gardner's expression for the fractional volume (Bouten *et al* 1990). The fractional volume V where conditions (3), (4) and (5) are satisfied is given by

$$V = \frac{\begin{array}{l}\Sigma_{\{c_i\}} \int (\Pi_j \,\mathrm{d}T_j) \exp[-\frac{1}{2}\Sigma_j (1 - c_j) T_j^2] \delta_{\mathrm{Kr}}[\Sigma\, c_j, fN] \\ \qquad\qquad \times\, \delta[\Sigma_j\, c_j T_j^2 - fN]\, \Pi_\mu\, \theta[\Sigma_j\, c_j T_j \eta_j^\mu - K\sqrt{fN}\,]\end{array}}{\Sigma_{\{c_i\}} \int (\Pi_j \,\mathrm{d}T_j) \exp[-\frac{1}{2}\Sigma_j (1 - c_j) T_j^2] \delta_{\mathrm{Kr}}[\Sigma\, c_j, fN] \delta[\Sigma_j\, c_j T_j^2 - fN]}. \tag{7}$$

Like Gardner, we want to calculate the entropy $\langle \ln V \rangle$ where the angle brackets mean the average over the patterns $\{\xi_i^\mu\}$. The average of $\ln V$ can be determined using the replica method

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

The average V^n is easily calculated if we replace the step functions and the delta functions by their Fourier representation. Using the same techniques and the same notation as Gardner, we obtain

$$\langle V^n \rangle = \int \prod_a \frac{\mathrm{d}E_a}{2\pi} \int \prod_{a<b} \frac{\mathrm{d}q_{ab}\,\mathrm{d}F_{ab}}{(2\pi/N)} \int \prod_a \frac{\mathrm{d}\psi_a}{2\pi}$$
$$\times \exp\left(N\left(\alpha G_1(q_{ab}) + G_2(F_{ab}, E_a, \psi_a) - f \sum_{a<b} F_{ab}q_{ab} + \frac{f}{2}\sum_a E_a + \frac{f}{2}\sum_a \psi_a \right)\right)$$
$$\times \left\{ \int \prod_a \frac{\mathrm{d}E_a}{2\pi} \int \prod_a \frac{\mathrm{d}\psi_a}{2\pi} \exp\left[N\left(G_2(0, E_a, \psi_a) + \frac{f}{2}\sum_a E_a + \frac{f}{2}\sum_a \psi_a \right)\right]\right\}^{-1}. \tag{9}$$

The integration variables q_{ab} have the same meaning as in Gardner's case. They differ only in the prefactor:

$$q_{ab} = \frac{1}{fN} \sum_j J_j^a J_j^b . \tag{10}$$

The variables E_a and F_{ab} are the same as in Gardner's paper: F_{ab} is the variable conjugate to q_{ab}, while E_a is introduced to impose the constraint (4). The new variables ψ_a have been introduced to express condition (3). The functions G_1 and G_2 are given by the following expressions:

$$G_1(q_{ab}) = \ln \prod_a \int_K^\infty \frac{\mathrm{d}\lambda_a}{2\pi} \int_{-\infty}^\infty \mathrm{d}x_a \exp\left(\mathrm{i}\sum_a x_a\lambda_a - \tfrac{1}{2}\sum_a x_a^2 - \sum_{a<b} q_{ab}x_a x_b \right) \tag{11}$$

$$G_2(F_{ab}E_a\psi_a) = \ln \sum_{\{c_a\}} \prod_a \int \mathrm{d}T_a$$
$$\times \exp\left(-\tfrac{1}{2}\sum_a (1-c_a)T_a^2 - \tfrac{1}{2}\sum_a E_a c_a T_a^2 + \sum_{a<b} F_{ab}c_a T_a c_b T_b - \tfrac{1}{2}\sum_a c_a\psi_a \right). \tag{12}$$

The function G_1 is the same as Gardner's. The function G_2 differs from Gardner's by two factors. The first term in the exponential function is the convergence factor discussed above, while the last term originates from the new condition (3).

In the large-N limit, one can use steepest-descent methods to evaluate the integral (9). This yields a set of equations for the saddle point. In order to be able to solve these equations, we look for a replica-symmetric solution

$$\begin{array}{lll} q_{ab} = q & F_{ab} = F & 1 \leqslant a < b \leqslant n \\ E_a = E & \psi_a = \psi & 1 \leqslant a \leqslant n. \end{array} \tag{13}$$

The four saddle-point equations for the replica-symmetric solution are given in appendix 1. For general values of α and f they can only be solved numerically.

The storage capacity α_c of the neural network is obtained when the value of q in the solution of the saddle-point equations tends to 1. This is the saturation limit of the network. When $q \rightarrow 1$, it becomes possible to replace the integrals in the equations by their asymptotic expressions. This is also done in appendix 1. The resulting equations become much simpler. The fraction f of non-zero couplings and the storage capacity α_c are both expressed as functions of a single parameter u which ranges between zero and infinity:

$$f = 1 - \mathrm{Erf}\, u \qquad \alpha_c \int_{-K}^\infty \mathrm{D}z(z+K)^2 = f + \frac{2}{\sqrt{\pi}}\, u\, \mathrm{e}^{-u^2} \tag{14}$$

where Erf is the standard error function while Dz is the gaussian measure

$$\mathrm{D}z = \frac{\exp(-z^2/2)}{\sqrt{2\pi}}\,\mathrm{d}z. \tag{15}$$

The two equations (14) give a parametric representation of the storage capacity α_c as a function of f. When $u \to 0$, we obtain the fully connected network $f = 1$ and α_c tends to Gardner's result for the storage capacity

$$\alpha_c = \left(\int_{-K}^{\infty} \mathrm{D}z(z+K)^2\right)^{-1}. \tag{16}$$

When $u \to \infty$, the network becomes extremely diluted $f \to 0$ and α_c tends to zero as

$$\alpha_c \to -2f \ln f\left(\int_{-K}^{\infty} \mathrm{D}z(z+K)^2\right)^{-1}. \tag{17}$$

The dependence of the storage capacity α_c on the fraction of non-zero couplings f is shown in figure 1 for the case $K = 0$ (curve a). At $f = 1$, the function $\alpha_c(f)$ is tangent to the horizontal line $\alpha_c = 2$. A weak dilution reduces the storage capacity by a very small amount only. As f decreases further, the function $\alpha_c(f)$ stays well above the linearly decaying storage capacity $\alpha_c = 2f$ (curve b). As an example, for $f = 0.75$ when 25% of all couplings are disconnected, the value of α_c is still 1.98 which means a 1% decrease only in storage capacity. When f becomes very small and tends to zero, α_c obviously must also tend to zero, but it does so very slowly. This is more clearly seen when we consider the storage capacity per synapse α_c/f as is usual for extremely diluted networks (Derrida *et al* 1987). From (17) we obtain, for the case $K = 0$,

$$\frac{\alpha_c}{f} = -4 \ln f \tag{18}$$

showing a logarithmic divergence when f tends to zero.

The dependence of α_c on the stability parameter K is identical for all values of f. This dependence has been studied by Gardner (1988) who gives a plot of α_c as a function of K.

Since the above results have been obtained within the assumption of replica symmetry, we have checked the local stability of the replica-symmetric saddle point. While these stability conditions are barely satisfied for the fully connected network $f = 1$, they become better satisfied as f decreases. This is reasonable because one expects a lesser degree of frustration in diluted networks.

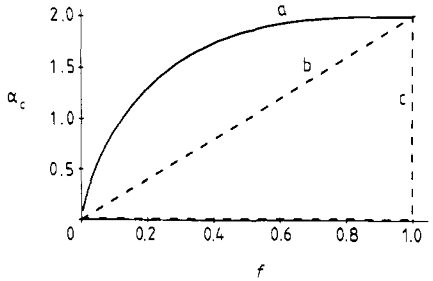

Figure 1. The storage capacity $\alpha_c(f)$ as a function of the fraction f of non-zero couplings for $K = 0$ in the case of annealed dilution (solid curve a) and of quenched dilution (broken line b: dilution before learning; broken line c: dilution after learning).

3. Storage capacity of diluted networks: quenched case

We now turn to the study of the storage capacity of a neural network in which the fraction $(1-f)$ of all synapses into neuron $i=0$ are cut randomly. The choice of vanishing coupling coefficients is independent of the learning process and, once it has been made, it is final. We will consider two different cases, depending on the time order of the dilution process and the learning process.

We first consider a network in which the random cutting is carried out before the learning process starts. The fN remaining connections are to be determined to guarantee storage of the given set of patterns. This again leads to the stability conditions (5). There is, however, an important difference from the previous section. In the annealed case, the choice of the coupling coefficients which will be zero is made during the learning process in such a way that they contribute to the storage properties of the network. This means that, in the sum (5), one is free to put the $(1-f)N$ zeros at the most favourable positions. In the present quenched case, on the other hand, the vanishing coefficients have been chosen at random, prior to the learning, and no freedom remains to shift the zeros around. The storage problem for the fN remaining coupling coefficients, conditions (4) and (5), is identical to the original Gardner problem for the fully connected network with fN neurons. The maximum number of patterns $p_{\max}$ that can be stored is therefore given by (Gardner 1989)

$$\frac{p_{\max}}{fN}=\left(\int_{-K}^{\infty} \mathrm{D}z(z+K)^2\right)^{-1}. \tag{19}$$

From this, we obtain the storage capacity per neuron

$$\alpha_c=\frac{p_{\max}}{N}=f\left(\int_{-K}^{\infty} \mathrm{D}z(z+K)^2\right)^{-1}. \tag{20}$$

The storage capacity depends linearly on the parameter f. It is represented by the broken line b in figure 1 for the case $K=0$.

We now turn to the more interesting case in which the random cutting is carried out after the fully-connected network has completed the learning process. The question now is: how well do the memorized patterns remain stored in the network after $(1-f)N$ connections are broken in a random way? This is the important problem of the robustness of memories against the cutting of bounds.

Imagine we have a fully connected network which has memorized the p random patterns $\{\xi_j^\mu\}$. The coupling coefficients J_j then satisfy the two conditions

$$\sum_{j=1}^{N} J_j^2=N \tag{21}$$

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_j\eta_j^\mu>K. \tag{22}$$

Suppose we now choose at random $(1-f)N$ of the coefficients J_j and put them equal to zero. Only fN terms will remain in the sums (21) and (22) and many of the inequalities (22) will become violated. However, for a pattern ξ^μ to be a fixed point of the retrieval dynamics, it is not necessary that (22) be satisfied with $K>0$. The pattern ξ^μ will remain a fixed point as long as the remaining sums in (22) are positive. It is therefore interesting to determine the distribution of the possible values of the remaining sums for different choices of the parameter f. Without lack of generality, we may assume that the $(1-f)N$ vanishing coefficients J_j are those with $j>fN$. So we must calculate the distribution of the truncated sums

$$\gamma_\mu = \frac{1}{\sqrt{N}} \sum_{j=1}^{fN} J_j \eta_j^\mu . \tag{23}$$

For $f=1$, this distribution has been obtained by Kepler and Abbott (1988). Their calculation can be generalized for $f<1$. Using the same techniques, one obtains after a long calculation

$$P_{K,f}(\gamma) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{D}t \frac{\exp[-\frac{1}{2}(f\sqrt{q}t+\gamma)^2/f(1-fq)]}{\sqrt{f(1-fq)}} \times \frac{H[\sqrt{(1-fq)/(1-q)(1-f)}\{K+[(1-f)\sqrt{q}t-\gamma(1-q)]/(1-fq)\}]}{H[(K+\sqrt{q}t)/\sqrt{1-q}]} . \tag{24}$$

The function $H(x)$ is related to the error function:

$$H(x) = \int_x^\infty \frac{\mathrm{d}t}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) = \frac{1}{2}\left[1-\mathrm{Erf}\left(\frac{x}{\sqrt{2}}\right)\right] . \tag{25}$$

The parameter q in (24) is the usual order parameter of Gardner (1988) and is related to the storage ratio $\alpha = p/N$ by her saddle-point equation (equation (22) of Gardner (1988)). Near the saturation limit $q \to 1$, the expression (24) can be simplified

$$P_{K,f}^{\mathrm{sat}}(\gamma) = \frac{1}{\sqrt{2\pi f}} \exp\left(-\frac{\gamma^2}{2f}\right) H\left(\frac{K-\gamma}{\sqrt{1-f}}\right) + \frac{1}{\sqrt{2\pi f(1-f)}} \exp\left(-\frac{1}{2}\frac{(\gamma-fK)^2}{f(1-f)}\right) H(-K). \tag{26}$$

When f tends to 1, the two distributions (24) and (26) reduce to the corresponding expressions of Kepler and Abbott (1988).

Using (24) and (26), we can now study the robustness of the memories in the fully connected network for different values of the parameters K and α (or q). As noted above, the patterns will remain fixed points of the retrieval dynamics as long as all γ are positive. Consider, as a first example, a network which has stored patterns using the stability parameter $K=0$. When $p=2N$, the network is at the saturation limit $q=1$ and we can use the simpler expression (26). A 1% cut of the connections (i.e. $f=0.99$) yields the distribution shown in figure 2. More than 26% of the γ are negative and as many fixed-point conditions are violated. Decreasing the number of cuts does not improve the situation. Using $H(0)=\frac{1}{2}$, one easily sees from the last term in (26) that any number of cuts, however small, will make at least 25% of the γ negative. The saturated network, in the case $K=0$, has absolutely no robustness. This, of course, is not surprising since the storage possibilities of the network are stretched to breaking point when $p=2N$. For smaller values of p, but keeping $K=0$, the situation does not improve drastically even though the network is far from saturated. In a network which stores only $N/2$ patterns, cutting 1% of the connections still produces 4% negative γ.

The robustness of the memories can be improved substantially by using a larger stability parameter K during the learning process. Figure 3 shows the distribution $P(\gamma)$ obtained when respectively 5% or 10% of the connections are cut randomly in a network which has stored memories up to saturation, but using the stability parameter

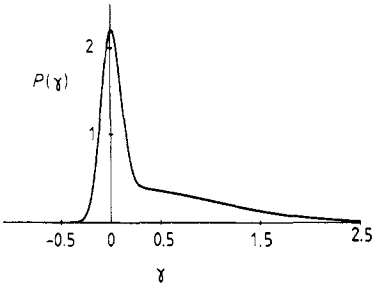

Figure 2. Probability distribution of the γ (see text) for a saturated network in the case $K=0$ for $f=0.99$.

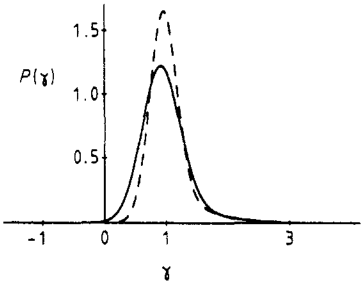

Figure 3. Probability distribution of the γ for a saturated network in the case $K=1$ for $f=0.95$ (broken curve) and $f=0.90$ (solid curve).

$K=1$. When 10% connections are cut, only 0.1% of the γ become negative. This number falls to less than 10^{-5} when only 5% of connections are cut. The number of patterns stored at saturation when $K=1$ is about $N/2$. Comparison with the results, quoted above, for $K=0$ indicates the advantage of storing the given set of patterns with the largest possible value of the stability parameter K. This increases both the size of the basins of attraction and the robustness of the stored memories.

In the preceding discussion, we have required that the fixed-point conditions $\gamma_\mu > 0$ be strictly satisfied for all patterns. When only very few γ become negative and are small in absolute value, it is reasonable to assume that one of the neighbouring system states will become a fixed point. The stored patterns can then still be recalled but with a small error. This effect, which is difficult to evaluate in the present context, will further increase the robustness of the stored memories.

4. Distribution of the coupling coefficients in diluted networks

When a fraction $1-f$ of all connections are cut in a network, the distribution of the coupling coefficients acquires the obvious form

$$P(J)=(1-f)\delta(J)+P_r(J). \tag{27}$$

The first term represents the $(1-f)N$ vanishing coefficients while the second term gives the distribution of the remaining connection strengths. The distribution $P_r(J)$ depends on the parameters α and f and on the way the dilution has been achieved.

In the case of annealed dilution, the coefficients J_j satisfy the conditions (3), (4) and (5). The distribution of the non-vanishing coefficients can then be expressed as

$$P_r(J)=\left\langle\frac{\Sigma_{\{c_i\}}(\Pi_j\int \mathrm{d}T_j)\Phi(\{c_i\},\{T_i\})\delta_{\mathrm{Kr}}(c_1,1)\delta(T_1-J)}{\Sigma_{\{c_i\}}(\Pi_j\int \mathrm{d}T_j)\Phi(\{c_i\},\{T_i\})}\right\rangle_{\{\eta\mu\}} \tag{28}$$

where the function Φ is non-zero only where the conditions (3)-(5) are satisfied:

$$\begin{aligned}\Phi(\{c_i\},\{T_i\})=\delta_{\mathrm{Kr}}[\textstyle\sum_j c_j,fN]\delta[\sum_j c_jT_j^2-fN]\\ \times\prod_\mu \theta[\textstyle\sum_j J_j\eta_j^\mu-K\sqrt{fN}]\exp[-\frac{1}{2}\sum_j(1-c_j)T_j^2].\end{aligned} \tag{29}$$

Using replicas to lift the denominator to the numerator and noting that the dependence on the parameter J is solely contained in the integration variable of the first neuron in the first replica, one obtains

$$P_r(J)=\lim_{n\to 0}\sum_{\{c_\alpha\}}\left(\prod_\alpha\int \mathrm{d}T_\alpha\right)\exp\left(-\frac{\psi}{2}\sum c_\alpha-\frac{E}{2}\sum c_\alpha T_\alpha^2\right.$$
$$\left.-\tfrac{1}{2}\sum(1-c_\alpha)T_\alpha^2+F\sum_{\alpha<\beta}c_\alpha T_\alpha c_\beta T_\beta\right)\delta_{\mathrm{Kr}}[c_1,1]\delta(T_1-J). \qquad (30)$$

Here we have again assumed replica symmetry to be valid. The parameters ψ, E and F have their values from the solution of the saddle-point equations (A1.1)-(A1.4). As the expression (30) does not depend on q, any dependence on K must come via the saddle-point equations. The expression (30) can be transformed using standard techniques into

$$P_r(J)=\frac{1}{\sqrt{2\pi}}\int \mathrm{d}z\frac{\mathrm{e}^{-\psi/2}\exp[-\frac{1}{2}(E+F)J^2+z\sqrt{F}J]}{1+\mathrm{e}^{-\psi/2}\sqrt{1/(E+F)}\exp\{[F/2(E+F)]z^2\}}. \qquad (31)$$

By integrating over J, one easily verifies that the distribution $P_r(J)$ is normalized to f as it should be to make the whole distribution (27) normalized to 1.

Let us first consider the case of full connnectivity $f=1$. From (A1.4), it is seen that this obtains when the parameter $\mathrm{e}^{-\psi/2}/\sqrt{E+F}$ tends to infinity. Expression (31) then simplifies to

$$P_r(J)=\frac{\sqrt{E+F}}{2\pi}\exp[-\tfrac{1}{2}(E+F)J^2]\int_{-\infty}^{\infty}\mathrm{d}z\exp\left[-\frac{1}{2}\left(\frac{E+2F}{E+F}\right)z^2+z\sqrt{F}J\right]. \qquad (32)$$

Since for $f=1$, the saddle-point equation (A1.3) becomes

$$1=\frac{E+2F}{(E+F)^2} \qquad (33)$$

we obtain a Gaussian distribution with mean 0 and variance 1:

$$P_r(J)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{J^2}{2}\right). \qquad (34)$$

This distribution is independent of the number of stored patterns p and of the stability parameter K. The result (34) shows that many coupling coefficients have very small values in the fully connected network.

For general values of $f<1$, it is useful to rewrite (31) as

$$P_r(J)=\frac{1}{\sqrt{2\pi}}\frac{E+F}{\sqrt{E+2F}}\exp\left(-\frac{1}{2}\frac{(E+F)^2}{E+2F}J^2\right)$$
$$\times\int_{-\infty}^{\infty}\frac{\mathrm{d}t}{\sqrt{2\pi}}\frac{\exp[-\frac{1}{2}(t-\sqrt{F(E+F)/(E+2F)}J)^2]}{1+\exp\{-\frac{1}{2}[(F/(E+2F))t^2-2\sigma]\}} \qquad (35)$$

where we have introduced the notation

$$\mathrm{e}^{-\sigma}=\frac{\mathrm{e}^{-\psi/2}}{\sqrt{E+F}}. \qquad (36)$$

For given values of α, f and K, one has to solve the saddle-point equation for E, F and ψ and plug these values in the expression (35). For general values of α, this can only be done numerically. However, near the saturation limit when $\alpha\to\alpha_c(K,f)$, the saddle-point equations simplify and it becomes possible to evaluate the integral (35) analytically. This is done in appendix 2. The final result is

$$P_r(J)=\frac{1}{\sqrt{2\pi}}\frac{E+F}{\sqrt{E+2F}}\exp\{-\tfrac{1}{2}[(E+F)^2/(E+2F)]J^2\}\theta[|J|-\sqrt{2\sigma/(E+F)}]. \tag{37}$$

This is a gaussian distribution with mean zero and variance $(E+2F)/(E+F)^2$ from which the middle section has been cut out. The width of the gap depends on the parameter f but not on K. The gap increases gradually with growing dilution and approaches its maximum value 1 when f tends to zero. Concurrently with the broadening gap, the width of the Gaussian decreases as is necessary to keep the value of J^2 unchanged and equal to 1, as is required by condition (4). These results are illustrated in figure 4, which shows the distribution $P_r(J)$ near the saturation limit for four values of the parameter f.

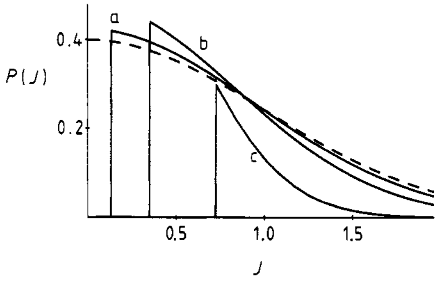

Figure 4. Probability distribution of the positive-valued coupling coefficients in the case of annealed dilution near saturation for $f=1$ (broken curve) and $f=0.92$, 0.67 and 0.18 (solid curves a, b and c, respectively).

Let us now turn to the case of quenched dilution. Here the distribution $P_r(J)$ is the standard Gaussian normalized to f

$$P_r(J)=\frac{f}{\sqrt{2\pi}}\exp(-J^2/2). \tag{38}$$

This is true in both cases of quenched dilution considered in section 3. If the random cutting is done prior to the learning process, we have seen that the conditions on the coupling coefficients J_j are identical with those for a fully connected network with fN neurons. The result (38) then follows immediately from the distribution (34) for a fully connected network. If, on the other hand, the cutting is done after the learning process is finished, we start from the Gaussian distribution (34). Random cutting of connections will then preserve the form of the distribution and solely change its normalization.

5. Discussion

The primary result of this paper is the determination of the maximal storage capacity α_c for a neural network model designed by the optimal Gardner prescription and diluted in three different ways. Figure 1 summarizes our findings for the case of a stability parameter $K=0$.

Curve c of figure 1 refers to the case of quenched dilution after learning. Here one starts with a fully connected model storing $p=2N$ patterns and cuts synapses at random. No correlation exists between the cuts and the patterns. As a result, for any degree of dilution, many stability conditions are violated and all patterns are likely to become destabilized. This lack of robustness is due to the extreme specialization of the synaptic matrix. To reach the ultimate limit $\alpha_c=2$ in the fully connected network, every coupling coefficient has to take on a sharply defined value. Any deviation from these values results in drastic changes of the cooperative behaviour. In our case, the storage capacity jumps abruptly from its maximal value $\alpha_c=2$ at $f=1$ to $\alpha_c=0$ for all $f<1$.

For values of α smaller than 2, the αN patterns can be stored in the network using different learning rules. The greatest robustness against quenched dilution is obtained when the stability parameter K is made as large as posible. This can be done using the minimum-overlap algorithm of Krauth and Mézard (1987). In this case, a much smaller percentage of unstable bits may occur, e.g. for $K=1$ corresponding to $\alpha\approx\frac{1}{2}$, a 10% dilution produces only 0.1% negative stabilities. Moreover, these negative stabilities are rather small in their absolute value. Hence, it is reasonable that configurations near to the patterns will act as attractors allowing retrieval with a small error (Amit *et al* 1990).

Curve b shows a simple linear dependence of α_c on f and corresponds to quenched dilution before learning. Here, a fraction $(1-f)$ of the synapses is set equal to zero. After that, the remaining synapses are chosen to stabilize as many patterns as possible. Again, there is no correlation between the dilution and the patterns to be stored. The linear decrease of α_c with increasing dilution results from the fact that the information about the patterns has to be carried by a restricted set of synapses. In this sense, the diluted system is equivalent to a fully connected system of smaller size.

The most interesting result is given by curve a and concerns the storage capacity for annealed dilution. Again, a fraction $(1-f)$ of the synapses is set equal to zero, but this time both the position of the zero synapses and the values of the remaining non-zero ones are determined in order to optimize the storage capacity. Now the dilution process is clearly correlated with the patterns under consideration and hence the zero bonds also carry information about the patterns. Due to the many partly contradictory requirements on the coupling coefficients, there is a large degree of frustration in fully connected models. This gives rise to many synapses with optimal value near to zero. A properly placed zero bond now carries almost as much information as such an optimal one. Therefore α_c decreases very slowly from the maximal value $\alpha_c=2$ at $f=1$. If 25% of the synapses are cut we find $\alpha_c=1.98$, i.e. a decrease in storage capacity of just 1%.

If f tends to zero, so does, of course, the storage capacity. Our explicit result $\alpha_c\approx-4f\ln f$ indicates, however, that the slope at $f=0$ is infinite. Hence, in extremely diluted networks, every non-zero synapse, if optimally placed, can carry a lot of information. This is well known from the storage of patterns with low level of activity (Wilshaw *et al* (1969, Nadal and Toulouse 1990) and is demonstrated here for the first time for patterns with symmetric statistics.

It is interesting to compare our findings with the results for a diluted Hopfield model (Sompolinsky 1986, Van Hemmen 1987). Although the dependence of α_c on f found there looks qualitatively similar to our curve a, the reasons are rather different. The Hebb rule used in the Hopfield model is non-optimal. As a result, near the saturation limit $\alpha_c\sim0.14$ many coupling coefficients in the fully connected network do not even have the correct sign. Quenched dilution now also eliminates part of these unfavourable bonds so that the curve $\alpha_c(f)$ stays somewhat above the linear law $f\alpha_c(1)$. It is reasonable that this kind of robustness requires some redundancy of the fully connected net. In the case of extreme dilution $f \to 0$, one finds for the Hopfield model $\alpha_c/f = 2/\pi$ (Derrida *et al* 1987) which is much smaller than the optimal value, as expected from the quenched character of the dilution.

Our study of robustness to random dilution displays a great similarity to the work of Virasoro (1988) on the effects of the destruction of synapses in neural networks. Whereas we have concentrated on the robustness of a network in which randomly chosen patterns are stored and on the role played by the stability parameter K, Virasoro (1988) has gone a step further, studying the consequences of random dilution in a network which stores ultrametric patterns. His calculation yields an interesting model for explaining the syndrome prosopagnosia in neurology.

In addition to the storage capacities, we have studied the distribution $P(J)$ of the optimal values of the synapses. The results show the differences between quenched and annealed dilution most clearly. For quenched dilution the distribution of the non-zero couplings is Gaussian with zero mean and a variance determined by the proper normalization constraint. Hence, despite the large number of zero bonds, there are many synapses with rather small absolute value. In the case of annealed dilution near saturation, $P(J)$ is a Gaussian from which the middle section has been removed (figure 4). This means that all non-zero synapses have absolute values larger than a given threshold which depends on the degree of dilution. If, for a synapse, the optimal value would lie inside the removed interval, it would be replaced by one of the broken bonds. In this way, the system optimizes its structure both with respect to the value of the non-zero bonds and with respect to the position of the zero bonds. This result clearly indicates that, for the case of annealed dilution, learning and dilution are intimately connected. We started by fixing the degree of dilution $(1-f)$ and then designed the network in the optimal way. Alternatively, one could start with an optimized fully connected model and then cut synapses in the order of their absolute value until the desired degree of dilution is reached. A similar procedure has already been used for the Hebb rule (Sompolinsky 1987, Van Hemmen 1987). Our result for the distribution $P(J)$ suggests that both procedures may be equivalent.

Acknowledgments

This work has been supported by the Belgian programme on Inter University Attraction Poles (IUAP) and by the Inter-University Institute for Nuclear Sciences (IIKW). One of us (AK) is indebted to the latter Institute for financial support.

Appendix 1

The four equations for the replica-symmetric saddle point are

$$fF = \frac{\alpha}{2\pi(1-q)} \int \mathrm{D}z \, \frac{\exp[-(K+\sqrt{q}\, z)^2/(1-q)]}{[H((K+\sqrt{q}\, z)/\sqrt{1-q})]^2} \tag{A1.1}$$

$$fq = \int \mathrm{D}z \left(\frac{\exp(-\psi/2)\sqrt{1/(E+F)}\, \exp\{\frac{1}{2}[F/(E+F)]z^2\}}{1+\exp(-\psi/2)\sqrt{1/(E+F)}\, \exp\{\frac{1}{2}[F/(E+F)]z^2\}} \right) \frac{E+F-Ez^2}{(E+F)^2} \tag{A1.2}$$

$$f=\int \mathrm{D}z\left(\frac{\exp(-\psi/2)\sqrt{1/(E+F)}\exp\{\frac{1}{2}[F/(E+F)]z^2\}}{1+\exp(-\psi/2)\sqrt{1/(E+F)}\exp\{\frac{1}{2}[F/(E+F)]z^2\}}\right)\frac{E+F+Fz^2}{(E+F)^2} \tag{A1.3}$$

$$f=\int \mathrm{D}z\left(\frac{\exp(-\psi/2)\sqrt{1/(E+F)}\exp\{\frac{1}{2}[F/(E+F)]z^2\}}{1+\exp(-\psi/2)\sqrt{1/(E+F)}\exp\{\frac{1}{2}[F/(E+F)]z^2\}}\right). \tag{A1.4}$$

Instead of (A1.2) and (A1.3) we can use the simpler derived equations

$$f(1-q)=\frac{1}{E+F}\int \mathrm{D}z\left(\frac{\exp(-\psi/2)\sqrt{1/(E+F)}\exp\{\frac{1}{2}[F/(E+F)]z^2\}}{1+\exp(-\psi/2)\sqrt{1/(E+F)}\exp\{\frac{1}{2}[F/(E+F)]z^2\}}\right)z^2 \tag{A1.5}$$

$$Fq+E=1. \tag{A1.6}$$

The equations for the fully connected network $f=1$ are recovered when we let the parameter $\mathrm{e}^{-\psi/2}/\sqrt{E+F}$ tend to infinity. The equations (A1.2), (A1.3) and (A1.5) become simple algebraic equations as is known in the calculations of Gardner (1988). For general f, the four equation (A1.1), (A1.4), (A1.5) and (A1.6) have to be solved to obtain q, F, E and ψ for given values of f, α and K. This can only be done numerically.

The storage capacity α_c for given values of f and K is obtained when q tends to 1. In this limit, it is possible to replace the saddle-point equations by their asymptotic expressions. For equation (A1.1), one easily obtains

$$\lim_{q\to 1} fF(1-q)^2=\alpha_c\int_{-K}^{\infty}\mathrm{D}z(z+K)^2. \tag{A1.7}$$

This shows that F tends to infinity like $(1-q)^{-2}$. Equation (A1.5) shows that $E+F$ also tends to infinity when $q\to 1$ but more slowly, like $(1-q)^{-1}$. Thus the ratio $F/(E+F)$ also tends to infinity. This ratio occurs as parameter in the integrals (A1.4) and (A1.5). The limit of these integrals can easily be evaluated using the identity

$$\lim_{a\to\infty}\frac{1}{1+\exp[(a/2)(z^2-b)]}=\theta(b-z^2) \tag{A1.8}$$

where b can be any constant. To simplify the notation, we introduce two new symbols a and σ defined by

$$a=\frac{F}{E+F}\qquad \mathrm{e}^{-\sigma}=\frac{\mathrm{e}^{-\psi/2}}{\sqrt{E+F}}. \tag{A1.9}$$

This allows us to rewrite the term between brackets in the integrals as

$$\frac{\exp\{[(a/2)(z^2-(2\sigma/a))]\}}{1+\exp\{[(a/2)(z^2-(2\sigma/a))]\}}. \tag{A1.10}$$

When $q\to 1$, we know that $a\to\infty$. Let us also choose $\sigma\to\infty$ but so that the ratio σ/a remains fixed and equal to a positive number u^2. Equations (A1.4) and (A1.5) then become

$$f=1-\mathrm{Erf}\,u \tag{A1.11}$$

$$\lim_{q\to 1} f(1-q)(E+F)=1-\mathrm{Erf}\,u+\frac{2}{\sqrt{\pi}}u\,\mathrm{e}^{-u^2}. \tag{A1.12}$$

Finally, from equation (A1.6), one gets

$$\lim_{q\to 1} F(1-q)^2=\lim_{q\to 1}(1-q)(E+F). \tag{A1.13}$$

From (A1.11), we see that any value of f between 0 and 1 can be obtained by allowing the constant u to vary between 0 and infinity. From (A1.7), (A1.11) and (A1.12), we obtain the corresponding storage capacity

$$\alpha_c \int_{-K}^{\infty} \mathrm{D}z(z+K)^2 = 1 - \mathrm{Erf}\, u + \frac{2}{\sqrt{\pi}}\, u\, \mathrm{e}^{-u^2}. \tag{A1.14}$$

Appendix 2

From appendix 1, we know that the saturation limit $q \to 1$ can be obtained for all values of f by letting both parameters a and σ tend to infinity while keeping the ratio $\sigma/a = u^2$ fixed and positive.

The expression for the integral in (35)

$$I(J) = \int_{-\infty}^{\infty} \frac{\mathrm{d}t}{\sqrt{2\pi}} \frac{\exp[-\frac{1}{2}(t - \sqrt{F(E+F)/(E+2F)}\, J)^2]}{1 + \exp[\![-\frac{1}{2}\{[F/(E+F)]t^2 - 2\sigma\}]\!]} \tag{A2.1}$$

can be simplified when $q \to 1$ in different steps. We first use

$$\lim_{q\to 1} \frac{F}{E+2F} = 1 \tag{A2.2}$$

which follows from appendix 1. Changing to a new integration variable

$$w = t - \sqrt{E+F}\, J \tag{A2.3}$$

transforms $I(J)$ into

$$I(J) = \lim_{q\to 1} \int_{-\infty}^{\infty} \mathrm{D}w \frac{1}{1 + \exp[-\frac{1}{2}[(w + \sqrt{E+F}\, J)^2 - 2\sigma]]}. \tag{A2.4}$$

As $I(J)$ is an even function of J, we can restrict the following discussion to positive values of J. Then both $\sqrt{E+F}\, J$ and 2σ tend to ∞ when $q \to 1$. We now use the following relation which holds asymptotically for any two large positive numbers A and B:

$$\frac{1}{1 + \exp[-\frac{1}{2}((w+A)^2 - B^2)]} = \theta[w - (B - A)] + \theta[-w - (B + A)]. \tag{A2.5}$$

Making use of this relation in (A2.3), we obtain

$$I(J) = 1 - \int_R^S \mathrm{D}w \tag{A2.6}$$

where

$$R = -\sqrt{E+F}\, \{[2\sigma/(E+F)]^{1/2} + J\}$$
$$S = \sqrt{E+F}\, \{[2\sigma/(E+F)]^{1/2} - J\}.$$

When $E + F$ tends to infinity, the lower limit of the integral always tends to $-\infty$ while the upper limit tends to ∞ or $-\infty$ depending on the sign of $\sqrt{2\sigma/(E+F)} - J$. So we obtain finally

$$I(J) = \theta[|J| - \sqrt{2\sigma/(E+F)}]. \tag{A2.7}$$

We remark that, when $q \to 1$, the ratio $2\sigma/(E+F)$ tends to a finite limit which depends only on the parameter f. From the definition of a, we have

$$\frac{2\sigma}{E+F} = 2\frac{\sigma}{a}\frac{F}{(E+F)^2}. \tag{A2.8}$$

Using (A1.7), (A1.11) and (A1.13), this becomes

$$\frac{2\sigma}{E+F} = \frac{2u^2 f}{1 - \operatorname{Erf} u + (2/\sqrt{\pi})u\,\mathrm{e}^{-u^2}}. \tag{A2.9}$$

This result depends only on the parameter u which determines f by (A1.11). The value increases monotonically between 0 and 1 when u varies between 0 and ∞.

References

Amit D J, Evans M R, Horner H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361
Bouten M, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 2605
Canning A and Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 3275
Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
—— 1989 *J. Phys. A: Math. Gen.* **22** 1969
Gardner E, Derrida B and Mottishaw E 1987 *J. Physique* **48** 741
Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
Kinzel W 1985 *Z. Phys.* B **60** 205
Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** 745
Krauth W, Nadal J P and Mézard M 1988 *J. Phys. A: Math. Gen.* **21** 2995
Nadal J P and Toulouse G 1990 *Network* **1** 61
Noest A J 1989 *Phys. Rev. Lett.* **63** 1739
Sompolinsky H 1986 *Phys. Rev.* A **34** 2571
—— 1987 *Heidelberg Colloq. on Glassy Dynamics and Optimization* ed J L Van Hemmen and I Morgenstern (Berlin: Springer)
Van Hemmen J L 1987 *Phys. Rev.* A **36** 1959
Van Hemmen J L and Van Enter A C D 1986 *Phys. Rev.* A **34** 2509
Virasoro M A 1988 *Europhys. Lett.* **7** 293
Willshaw D J, Buneman O P and Longuet-Higgins M C 1969 *Nature* **222** 960