Phase transitions in simple learning

1989 J. Phys. A: Math. Gen. 22 2133

(http://iopscience.iop.org/0305-4470/22/12/016)

J. Phys. A: Math. Gen. **22** (1989) 2133-2150. Printed in the UK

Phase transitions in simple learning

J A Hertz†, A Krogh‡ and G I Thorbergsson†

† Nordita, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark
‡ Niels Bohr Institute, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark

Received 6 February 1989

Abstract. We investigate learning in the simplest type of a layered neural network, the one layer perceptron. The learning process is treated as a statistical dynamical problem. Quantities we are interested in include the relaxation time (the learning time) and the capacity and how they depend on noise and constraints on the weights. The relaxation time is calculated as a function of the noise level and the number p of associations to be learned. We consider three different cases for input patterns that are random and uncorrelated. In the first, where the connection weights are constrained to satisfy $N^{-1}\Sigma_i\omega_i^2 = S^2$, there is a critical value of $p\ (<N)$ separating regimes of perfect and imperfect learning at zero noise. In contrast, the second model, unconstrained learning, exhibits a different kind of transition at $p = N$, and noise plays no role. In the third model, where the constraint is imposed only on the thermal fluctuations, there is a line of phase transitions terminating at $p = N$ and zero noise. We have also considered learning with correlated input patterns. The most important difference is the emergence of a second relaxation time, which we interpret as the time it takes to learn a prototype of the patterns.

1. Introduction

Layered neural networks have been the focus of much interest recently. Most of that work has been concerned with the learning process—what is the most efficient learning algorithm and the best cost function to use? One of the best known algorithms is the delta rule [1] which is based on gradient descent of the cost function. Learning takes place when connections between the units are changed in such a way as to descend the cost function surface. In this paper we study this learning process as a statistical dynamical problem.

A Langevin model is a natural choice for the study of learning in the presence of noise. It is well known from many physics problems and has obvious similarities to gradient descent. We will then study the effect of noise on learning and find the relaxation time of the learning process.

A brief description of some of this work has appeared previously [2]. The cost function that we use is

$$E = \tfrac{1}{2}\sum_{\mu i}\left[\zeta_i^\mu - f\left(\sum_j w_{ij}\xi_j^\mu\right)\right]^2 + \tfrac{1}{2}\lambda\sum_{ij} w_{ij}^2 - \sum_{ij} h_{ij}w_{ij} \tag{1}$$

where ζ_i^μ is the target for pattern $\mu = 1, \ldots, p$, and unit $i = 1, \ldots, N$ and ξ_j^μ is the input pattern for unit $j = 1, \ldots, N$, w_{ij} are the connections from unit j to unit i. In the last term in (1) h_{ij} is an auxiliary field that is needed in our calculations in the limit $h_{ij} \to 0$. The function f is the sigmoidal activation function of the output units, e.g.

0305-4470/89/122133+18$02.50

$f(x) = \tanh(x)$. In a one-layer perceptron the non-linearity of f is unimportant if the saturation value of f exceeds the target magnitude, so here we set $f(x) = x$ (Widrow-Hoff or adaline learning). We study only the simplest type of a layered network, the perceptron with one layer of connections. The cost function becomes

$$E = \tfrac{1}{2}\sum_{\mu}\left(\zeta^{\mu} - \frac{1}{\sqrt{N}}\sum_{j} w_j \xi_j^{\mu}\right)^2 + \tfrac{1}{2}\lambda \sum_{j} w_j^2 - \sum_{j} h_j w_j. \tag{2}$$

It separates in the output unit index so we omit this index. The output units can be treated separately.

The change in the connections is proportional to the negative gradient of E

$$\Delta w_{ij} \propto -\frac{\partial E}{\partial w_i} = \sum_{\mu}\left(\frac{1}{\sqrt{N}}\zeta^{\mu} - \frac{1}{N}\sum_{j} w_j \xi_j^{\mu}\right)\xi_i^{\mu} - \lambda w_i + h_i. \tag{3}$$

The chemical potential term with the constant λ has been added to the cost function to keep the connections from growing to infinity. It is often considered an advantage to have a weight decay term of this form. We will investigate what effect it has and the importance of the value of λ.

There is another reason for including this term. It might be necessary to restrict the w_i to a small number of values, e.g. binary ones $w_i = \pm S$. It is difficult to work with this condition. This difficulty is well known in the theory of magnetic systems, where a popular remedy is to replace the binary Ising model by the so-called spherical model where the spins can have a continuous range of values but their average magnitude is fixed. This condition is implemented by a term in the energy of the magnetic system like the λ term in our cost function. In the learning problem it is then natural to choose λ so that $[\langle w_i^2 \rangle]_{\xi\zeta} = S^2$. The bracket $[\ ]_{\xi\zeta}$ is an average over input and output patterns and $\langle\ \rangle$ is a noise average.

It is natural to characterise the asymptotic state in terms of the parameter

$$q \equiv \left[\frac{1}{N}\sum_{k}\langle w_k \rangle^2\right]_{\xi\zeta}. \tag{4}$$

We can derive a general formula for it in the static limit, i.e. at time long enough that the connections have relaxed to a time-independent value. The relaxation time depends on the parameter q.

We begin by studying random uncorrelated patterns. It turns out that noise can have the effect of decreasing the relaxation time. If there is no constraint on the connection strengths, $\lambda = 0$, the noise has no effect.

We also study correlated patterns characterised by a single common mutual overlap. Then two relaxation times appear. One of them is shorter and can be interpreted as the time it takes the network to learn the average features of the problem.

The relaxation time that we calculate is the time it takes the network to learn whatever it learns. But it may not learn the training set perfectly. Therefore one should not necessarily conclude that a shorter learning time is desirable.

2. The Langevin model

The delta rule is based on gradient descent by iteration of the cost function. If instead we introduce continuous time the learning process can be described by a differential

equation. In the presence of noise that is the Langevin equation [3] which models the relaxation of a physical system

$$\frac{\partial w_i}{\partial t} = -\gamma_0 \frac{\partial E}{\partial w_i} + \eta_i(t). \tag{5}$$

The learning rate parameter γ_0 determines the microscopic timescale of the problem. We let $\eta_i(t)$ be white noise with variance

$$\langle \eta_i(t)\eta_j(t')\rangle = 2T\gamma_0\delta_{ij}\delta(t-t') \tag{6}$$

where T is the noise level, the analogue of temperature.

The Langevin equation (5) with the cost function in (2) is

$$\frac{\partial w_i}{\partial t} = -\gamma_0\left(B_i - \sum_j A_{ij}w_j - \lambda w_i + h_i(t)\right) + \eta_i(t) \tag{7}$$

where

$$B_i = \frac{1}{\sqrt{N}}\sum_\mu \zeta^\mu \xi_i^\mu \tag{8}$$

and

$$A_{ij} = \frac{1}{N}\sum_\mu \xi_i^\mu \xi_j^\mu . \tag{9}$$

As in a physical system we are interested in the average behaviour in the limit of large N. Here we have to carry out two kinds of averages—first over noise (thermal average, denoted by $\langle\ \rangle$) and then over the distribution of patterns (average over disorder, denoted by $[\ ]_{\xi\zeta}$). We begin by considering the static limit of (7).

3. Statics

First we will look at the equilibrium properties of the system. Setting the auxiliary field $h_{ij} = 0$ and averaging the Langevin equation (7) over noise, we find

$$0 = -\gamma_0[B_i - (A_{ij} + \lambda\delta_{ij})\langle w_j\rangle] \tag{10}$$

(sum on free indices—here j). The unaveraged response function g_{ij} is defined by

$$\mathbf{g} = (\lambda\mathbf{I} + \mathbf{A})^{-1}. \tag{11}$$

We have then

$$\langle w_i\rangle = g_{ij}B_j \tag{12}$$

where g_{ij} has the expansion

$$g_{ij} = \lambda^{-1} - \lambda^{-2}A_{ij} + \lambda^{-3}A_{ik}A_{kj} - \cdots. \tag{13}$$

This function describes the behaviour of the system in the limit of infinite time.

3.1. The projection method

For completeness we will show how this reduces to the projection method [4, 5, 6]. If we use the definitions of A_{ij} and B_i and (13) then (12) becomes

$$\begin{aligned}\langle w_i\rangle &= \xi_i^\mu[\lambda^{-1} - \lambda^{-2}N^{-1}\xi_j^\mu\xi_j^\nu + \lambda^{-3}N^{-2}\xi_k^\mu\xi_k^\sigma\xi_j^\sigma\xi_j^\nu - \cdots]N^{-1/2}\zeta^\nu \\ &= N^{-1/2}\xi_i^\mu\zeta^\nu[\lambda^{-1} - \lambda^{-2}Q_{\mu\nu} + \lambda^{-3}Q_{\mu\sigma}Q_{\sigma\nu} - \cdots]\end{aligned} \tag{14}$$

where $Q_{\mu\nu} = (1/N)\,\Sigma_i \xi_i^\mu \xi_i^\nu$ is the overlap matrix. If $(\lambda\delta_{\mu\nu} + Q_{\mu\nu})$ is regular it can be summed to

$$\langle w_i \rangle = N^{-1/2} \xi_i^\mu (\lambda \mathbf{I} + \mathbf{Q})^{-1}_{\mu\nu} \zeta^\nu. \tag{15}$$

In the case of $\lambda = 0$ it reduces to the usual form of the projection method. If the input patterns are linearly independent $\mathbf{Q}$ is regular and (15) can be used. If the p inputs are generated randomly it is unlikely that they will be linearly dependent for $p < N$, but for $p > N$ there will of course be patterns that are linearly dependent and not all the patterns can be learned correctly. For a large system this shows that the capacity of the network for random patterns is at most $p = N$.

3.2. Fixing the chemical potential

We have chosen the spherical constraint on the size of the couplings

$$[\langle w_i^2 \rangle]_{\xi\zeta} = S^2. \tag{16}$$

The average of w^2 is part of the autocorrelation function

$$C \equiv [\langle w_i^2 \rangle - \langle w_i \rangle^2]_{\xi\zeta}. \tag{17}$$

This is connected to the response function through the fluctuation-response theorem [7]

$$C = TN^{-1}[\operatorname{tr} \mathbf{g}]_{\xi\zeta}. \tag{18}$$

Now from (16) and (17)

$$TN^{-1}[\operatorname{tr} \mathbf{g}]_{\xi\zeta} = S^2 - q \tag{19}$$

where q is the parameter that was defined in (4). The parameter q is similar to the Edwards-Anderson order parameter in spin glasses.

From (12) it is found that

$$q = [N^{-2} \zeta^\mu \xi_i^\mu \zeta^\nu \xi_j^\nu g_{ik} g_{kj}]_{\xi\zeta} \tag{20}$$

(sum on all indices). From the definition of g

$$\begin{aligned} g_{ik}g_{kj} &= (\lambda^{-1} - \lambda^{-2}A_{ik} + \lambda^{-3}A_{il}A_{lk} - \cdots)(\lambda^{-1} - \lambda^{-2}A_{kj} + \lambda^{-3}A_{kl}A_{lj} - \cdots) \\ &= \lambda^{-2} - 2\lambda^{-3}A_{ij} + 3\lambda^{-4}A_{ik}A_{kj} - \cdots \\ &= -\frac{\partial}{\partial\lambda} g_{ij}. \end{aligned} \tag{21}$$

Then

$$q = -N^{-2}\left[\zeta^\mu \xi_i^\mu \zeta^\nu \xi_j^\nu \frac{\partial}{\partial\lambda} g_{ij}\right]_{\xi\zeta}. \tag{22}$$

If we denote the expression in the brackets of (14) by $f_{\mu\nu}$, i.e.

$$f_{\mu\nu} = \lambda^{-1} - \lambda^{-2}Q_{\mu\nu} + \lambda^{-3}Q_{\mu\sigma}Q_{\sigma\nu} - \cdots \tag{23}$$

then it is easily realised that $\lambda f_{\mu\nu} = 1 - N^{-1}\xi_i^\mu g_{ij} \xi_j^\nu$ so that q can be written as

$$q = N^{-1}\left[\zeta^\mu \zeta^\nu \frac{\partial}{\partial\lambda}(\lambda f_{\mu\nu})\right]_{\xi\zeta}. \tag{24}$$

In what follows we will mainly be concerned with random outputs, i.e. the probability of ζ being $+1$ and -1 is the same. Then (22) can be averaged over the outputs and

$$q = -N^{-1}\left[\frac{\partial}{\partial\lambda}(A_{ij}g_{ij})\right]_{\xi} = -N^{-1}\left[\frac{\partial}{\partial\lambda}\operatorname{tr}(\mathbf{Ag})\right]_{\xi}. \tag{25}$$

From (13) it is seen that $\mathbf{g}$ obeys the Dyson equation

$$\mathbf{g} = \lambda^{-1}\mathbf{I} - \lambda^{-1}\mathbf{Ag} \tag{26}$$

which simplifies the expression for q

$$q = N^{-1}\frac{\partial}{\partial\lambda}\operatorname{tr}(\lambda\mathbf{G}) \tag{27}$$

where $\mathbf{G}$ is the pattern average of $\mathbf{g}$

$$\mathbf{G} = [\mathbf{g}]_{\xi\zeta}. \tag{28}$$

Equation (27) implicitly determines λ.

We have not made any assumption about the input patterns in deriving all these equations.

4. Random uncorrelated input patterns

Let the input patterns be random and uncorrelated, $\xi_i^\mu = \pm 1$, with equal probability. It is first shown that q does not depend on the distribution of the output patterns.

If we write out (22) we get

$$q = N^{-2}\frac{\partial}{\partial\lambda}[\lambda^{-1}\xi_i^\mu\zeta^\mu\xi_i^\nu\zeta^\nu - N^{-1}\lambda^{-2}\xi_i^\mu\zeta^\mu\xi_i^\sigma\zeta^\sigma\xi_j^\sigma\zeta^\sigma\xi_j^\nu\zeta^\nu + \cdots]_{\xi\zeta} \tag{29}$$

where pairs of ζ are put in with every pair of ξ ($\zeta^\mu\zeta^\mu = 1$). When the inputs are random, so are the $\xi\zeta$ and therefore the distribution of the output patterns is unimportant. The expression in the $[\cdot]_\xi$ will only have terms that are diagonal in both upper and lower index, $\mu = \nu$, $i = j$, so q becomes of the form (25) and is therefore given by (27).

Also, when the inputs are uncorrelated $\mathbf{G} = G\mathbf{I}$. Then

$$q = \frac{\partial}{\partial\lambda}(\lambda G). \tag{30}$$

4.1. Finding the response function

The summation of the series for the response function, G, can be done by diagrammatic methods (figure 1; for details of the method see the appendix). If we define the self-energy Σ as [3]

$$G^{-1} = \lambda + \Sigma \tag{31}$$

we only have to sum the irreducible diagrams. This diagram series can be simplified by 'dressing' the graphs, i.e. by drawing them with full Green functions G instead of the zero-order ones, λ^{-1}. Then we get the series in figure 2, which can easily be calculated. The calculation involves the evaluation of averages of the form

$$\left[N^{-q}\sum_{\mu}\sum_{i_1 i_2\cdots i_{q-1}}\xi_i^\mu\xi_{i_1}^\mu\cdots\xi_{i_{q-1}}^\mu\right]_{\xi} = \alpha \tag{32}$$

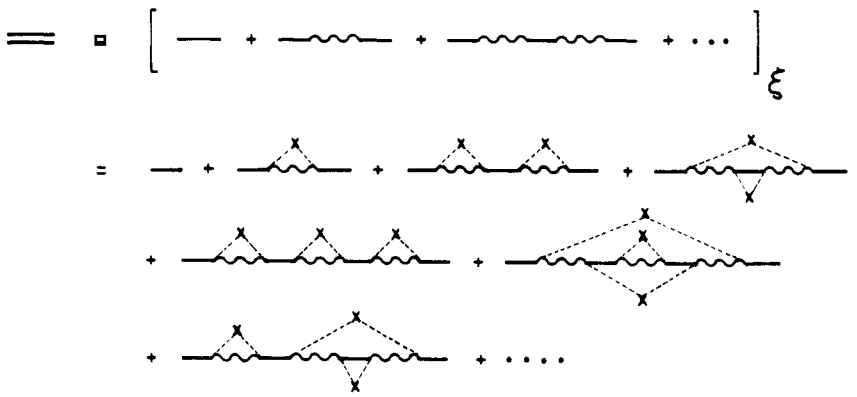

Figure 1. The double line is $-G$, the line is $-\lambda^{-1}$ and the wiggly line symbolises A. The broken lines with the '×' connect ends of the As to indicate that those ends have the same pattern and unit index. For more details consult the appendix.

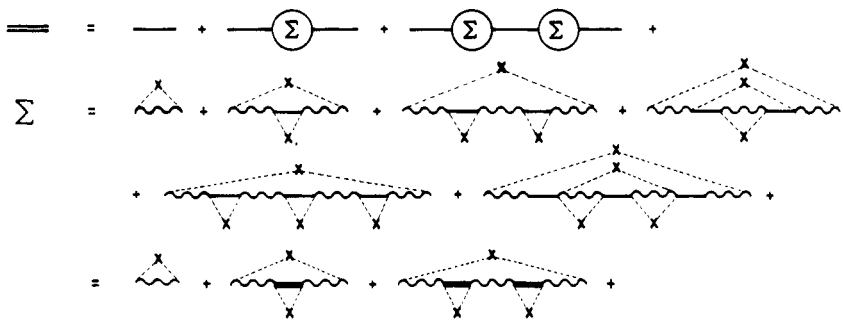

Figure 2. The response function written in terms of the self-energy Σ. Some of the diagrams for the self-energy and the dressing of the λ line are also shown.

where we have defined the load parameter $\alpha \equiv p/N$. The series for the self-energy, Σ, becomes

$$\sum = \alpha - \alpha G + \alpha G^2 - \cdots = \frac{\alpha}{1+G}. \tag{33}$$

Then the response function is

$$G^{-1} = \lambda + \frac{\alpha}{1+G}. \tag{34}$$

From the last expression we find

$$q = \frac{\partial}{\partial \lambda}(\lambda G) = \frac{\partial}{\partial \lambda}\left(\frac{-\alpha}{1+G^{-1}}\right) = \frac{\alpha}{(1+G^{-1})^2}\frac{\partial G^{-1}}{\partial \lambda}. \tag{35}$$

The derivative of G^{-1} is also found from (34) giving

$$\frac{\partial G^{-1}}{\partial \lambda} = \left(1 - \frac{\alpha}{(1+G^{-1})^2}\right)^{-1}. \tag{36}$$

Finally

$$q = S^2 - TG = \frac{\alpha}{(1+G^{-1})^2 - \alpha} \tag{37}$$

where (19) has been used. This expression relates G, S and α. It does not have a nice closed solution but we can examine the interesting limits analytically.

First a trivial limit: if the temperature is large we find $S^2 - TG \simeq \alpha G^2$. Only the positive solution makes sense and G is

$$G = S^2/T \qquad T \gg 1. \tag{38}$$

Thus $q = 0$, i.e. nothing is learned.

The low-noise limit is more relevant and interesting. First assume that G is large. Then for low T

$$S^2 - TG \simeq \frac{\alpha}{1-\alpha} \tag{39}$$

giving $TG = S^2 - \alpha/(1-\alpha)$. Because G becomes negative for $\alpha/(1-\alpha) > S^2$ this clearly only holds for α less than a critical value

$$\alpha_c = \frac{S^2}{1+S^2}. \tag{40}$$

If this inequality is satisfied,

$$G = \frac{1}{T}\left(S^2 - \frac{\alpha}{1-\alpha}\right) \tag{41}$$

at low temperatures. We have assumed that G is large. For this to hold, we require that $T \ll \alpha_c - \alpha$.

Right at $\alpha = \alpha_c$, or more generally, for $|\alpha - \alpha_c| \ll T \ll 1$, the response function G becomes

$$G = \left(\frac{2\alpha_c}{T}\right)^{1/2}(1-\alpha_c)^{-1} = \left(\frac{2S^2(1+S^2)}{T}\right)^{1/2}. \tag{42}$$

At $\alpha > \alpha_c$ we find G^{-1} by expanding around $G^{-1}(T=0) \equiv x_0 = \sqrt{\alpha/\alpha_c} - 1$. Then

$$G^{-1} \simeq x_0 + \frac{T(1-\alpha_c)^2}{2x_0\alpha_c\sqrt{\alpha_c/\alpha}} \qquad T \ll \alpha - \alpha_c \ll 1. \tag{43}$$

The transition at α_c has the following meaning: below α_c all the patterns can be learned perfectly in the low-noise limit. We can see this using (34) because with the solution (41) λ goes to zero as $T \to 0$ and (15) for the $\langle w_i \rangle$ reduces to the standard projection method result. Above α_c λ remains non-zero even for $T \to 0$, so the patterns are imperfectly learned.

4.2. Dynamics

To find the relaxation time of the learning process it is necessary to consider the dynamics. The reponse function that we have introduced in the static limit is more generally

$$G_{ij}(\omega) \equiv \left[\frac{\partial\langle w_i(\omega)\rangle}{\partial h_j}\right]_{\xi\zeta}. \tag{44}$$

The response function is diagonal when the patterns are uncorrelated, but we will need the general form in the next section.

The Fourier transform of the Langevin equation (7) is

$$w_i(\omega) = w_i^0(\omega) - G_0(\omega) \sum_j A_{ij} w_j(\omega) \tag{45}$$

where

$$G_0(\omega) = \frac{1}{\lambda - i\omega/\gamma_0} \tag{46}$$

and

$$w_i^0(\omega) = G_0(\omega)[B_i 2\pi\delta(\omega) + \eta_i(\omega)/\gamma_0 + h_i(\omega)]. \tag{47}$$

By iterating the equation we get a series expansion for w_i

$$w_i = w_i^0 - G_0 A_{ij} w_j^0 + G_0^2 A_{ij} A_{jk} w_k^0 - \cdots. \tag{48}$$

The response function (44) is then

$$G_{ij}(\omega) = [G_0(\omega) - G_0^2(\omega) A_{ij} + G_0^3(\omega) A_{ik} A_{kj} - \cdots]_{\xi\zeta}. \tag{49}$$

Note that the series in (49) is similar to the series for the unaveraged response function g_{ij} in (13). Apart from the pattern average, equation (13) is the static limit, $\omega \to 0$, of (49) because $G_0(0) = \lambda^{-1}$. We can use the results of the previous section by replacing λ by $\lambda - i\omega/\gamma_0$. From the response function we can find the characteristic relaxation time

$$\tau = \frac{\int_0^\infty tG(t)\,dt}{\int_0^\infty G(t)\,dt} = G(0) \lim_{\omega\to 0} \frac{\partial G^{-1}(\omega)}{\partial(-i\omega)}. \tag{50}$$

Note that

$$\frac{\partial G^{-1}(\omega)}{\partial(-i\omega)} = \frac{1}{\gamma_0} \frac{\partial G^{-1}}{\partial\lambda} \tag{51}$$

so from (36) we find

$$\tau = \frac{G(0)}{\gamma_0}\left(1 - \frac{\alpha}{(1+G^{-1}(0))^2}\right)^{-1}. \tag{52}$$

The static limit $G(0)$ of the response function which was determined before is needed to find the relaxation time.

There are three interesting cases, corresponding to eq (41)-(43). When $\alpha < \alpha_c$ the response function is given in the low-T limit by (41). The relaxation time is then

$$\tau = \frac{1}{\gamma_0 T} \frac{S^2(1-\alpha) - \alpha}{(1-\alpha)^2} \qquad T \ll \alpha_c - \alpha \ll 1. \tag{53}$$

In the critical region ($|\alpha - \alpha_c| \ll T \ll 1$) we use (42) and find

$$\tau = \left(\frac{2S^2(1+S^2)}{T}\right)^{1/2} \frac{1}{\gamma_0(1-\alpha)}. \tag{54}$$

Finally when $\alpha > \alpha_c$ we find from (43)

$$\tau = \left[\gamma_0 x_0 \left(1 + \frac{T(1-\alpha_c)^2}{2x_0^2 \alpha_c \sqrt{\alpha_c/\alpha}} \right) (1-\alpha_c) \left(1 + T\frac{1-\alpha_c}{x_0} \right) \right]^{-1} \tag{55}$$

for $T \ll \alpha - \alpha_c \ll 1$.

These equations also tell us something interesting about the learning process. From (53) we see that although the patterns can in principle be learned perfectly for $\alpha < \alpha_c$, it takes forever to do so. At finite temperature the learning time is finite but at the cost of imperfect learning (because then λ is non-zero). Note also that for fixed T the learning time shrinks as α approaches α_c, but this does not mean that the learning is improving. Rather the asymptotic performance is getting worse, so it takes less time to reach this state.

Right at α_c the learning time does not diverge as rapidly as $T \to 0$ (like $T^{-1/2}$ instead of T^{-1}), but this advantage is of course offset by a correspondingly more severe degradation of the asymptotic performance by small amounts of noise.

Above α_c the learning time is always finite even at zero temperature (though it diverges as $(\alpha - \alpha_c)^{-1}$). Correspondingly the asymptotic performance is never perfect.

5. Unconstrained learning

When there is no constraint on the couplings, $\lambda = 0$, the response function in (34) becomes

$$G^{-1}(\omega) = -i\omega/\gamma_0 + \frac{\alpha}{1+G(\omega)}. \tag{56}$$

Some care is needed when the relaxation time is determined from this equation. The response function has a pole which indicates that there is a non-relaxing contribution. We subtract this pole to get a finite relaxation time. Putting $z = -i\omega/\gamma_0$ the solution of (56) is

$$G(z) = \frac{1-\alpha-z+\sqrt{(z+\alpha-1)^2+4z}}{2z}. \tag{57}$$

Consider first the case $\alpha < 1$. In the limit $z \to 0$

$$G(z) = \frac{1-\alpha}{z}. \tag{58}$$

This pole has the following meaning: for $\lambda = 0$ and no noise the dynamics (7) takes place only in the subspace spanned by the patterns. Any component of the initial state in the complementary subspace will not relax. The residue of the pole in $\mathrm{tr}G$ at $z = 0$ is just the number of these non-relaxing components.

To study the dynamics in the subspace spanned by the patterns we subtract this pole off, defining the function

$$\hat{G}(z) = G(z) - \frac{1-\alpha}{z} = \frac{\alpha-1-z+\sqrt{(z+a-1)^2+4z}}{2z}. \tag{59}$$

The relaxation time is then

$$\tau = \frac{\hat{G}(0)}{\gamma_0} \lim_{z\to 0} \frac{\partial \hat{G}^{-1}(z)}{\partial z} = \frac{1}{\gamma_0(1-\alpha)^2}. \tag{60}$$

When $\alpha > 1$ there is no pole in G. The relaxation time is

$$\tau = \frac{G(0)}{\gamma_0} \lim_{z \to 0} \frac{\partial G^{-1}(z)}{\partial z} = \frac{\alpha}{\gamma_0(\alpha - 1)^2}. \tag{61}$$

To summarise

$$\tau = \begin{cases} 1/[\gamma_0(1-\alpha)^2] & \text{when } \alpha < 1 \\ \alpha/[\gamma_0(\alpha-1)^2] & \text{when } \alpha > 1 \end{cases}. \tag{62}$$

The critical value of the load parameter is $\alpha_c = 1$ in agreement with results by Opper [8]. Note that with $\lambda = 0$ the noise does not play any role in the problem.

6. Learning with constrained thermal fluctuations

There is an alternative model to the one we have considered which is also interesting. We can assume that the thermal fluctuations rather than the connection strengths are constrained. This is implemented by the following condition

$$[\langle w_i^2 \rangle - \langle w_i \rangle^2]_{\xi\zeta} = S^2. \tag{63}$$

The expression on the left-hand side is the autocorrelation function for the connections. Now the fluctuation-response relation is simply

$$S^2 = C(t=0) = TG(0). \tag{64}$$

This was the model we considered in [2], though there was an error in equation (19) of that paper. The relaxation time in (50) is then

$$\tau = \frac{S^2}{\gamma_0 T}\left(1 - \frac{\alpha}{(1+T/S^2)^2}\right)^{-1}. \tag{65}$$

The relaxation time diverges at a critical value of the load parameter

$$\alpha_c = (1 + T/S^2)^2. \tag{66}$$

In our previous paper [2], we were mostly interested in the transition as a function of α for fixed finite T, so we ignored the factor $1/T$ in (65).

A few words need to be said about the value of $\alpha_c(T)$: the system cannot learn more than N patterns because they are then linearly dependent. However, there is nothing that prevents us from investigating the relaxation process above $\alpha = 1$. What the equations we have derived mean is that the relaxation process breaks down at $\alpha = \alpha_c(T)$. They do not imply that the system can learn more than N patterns.

The breakdown at $\alpha = \alpha_c$ can be seen clearly by considering the effect of the requirement that the relaxation time be real and positive. This limits the range of α in which the above discussion is true. The solution of (34) for G (at $\omega = 0$) can be written as

$$G(\lambda) = \frac{1}{2\lambda}\left(1 - \alpha - \lambda \pm \sqrt{(1-\alpha-\lambda)^2 + 4\lambda}\right). \tag{67}$$

We must choose the + branch to get a positive value of τ. The analytic properties of this function depend on the value of α. When $\alpha < 1$ the function $G(\lambda)$ has a pole at

$\lambda = 0$. It is then always possible to satisfy the equation $G = S^2/T$. When $\alpha > 1$ there is no pole at $\lambda = 0$. The response function G has a maximum at λ_0 where

$$\lambda_0 = -(1-\sqrt{\alpha})^2. \tag{68}$$

It follows that S^2/T must be less than the maximum or

$$\frac{S^2}{T} < \frac{1-\alpha-\lambda_0}{2\lambda_0} = \frac{1}{\sqrt{\alpha}-1}. \tag{69}$$

The condition $\alpha < (1+T/S^2)^2 = \alpha_c$ follows. There is no stable solution to (63) for $\alpha > \alpha_c$.

7. Correlated patterns

We now consider input patterns that are biased, letting the probability that $\xi_i^\mu = \pm 1$ be $(1 \pm a)/2$ so the average of the ξ is a. In this case the response function in (13) is no longer diagonal but the series expansion for G_{ij} is still valid. The diagram technique used in the previous section must then be generalised.

7.1. The eigenvalues of the response function

First, notice that $\mathbf{G}$ is of the form

$$\mathbf{G} = \begin{bmatrix} g_0 & g_1 & \cdots & g_1 \\ g_1 & g_0 & & g_1 \\ \vdots & & & \\ g_1 & & \cdots & g_0 \end{bmatrix}. \tag{70}$$

This matrix has the eigenvalues $\gamma_1 = g_0 + (N-1)g_1$ and $\gamma_2 = g_0 - g_1$, which is $(N-1)$-fold degenerate.

We sum the series using the following diagram technique (details are in the appendix). The state of an input unit for a pattern μ can be written $\xi_i^\mu = x_i^\mu + a$. The average of the new variable, x_i^μ, is zero. The series for G_{ij} is drawn as before and in figure 3(a) we explain how the averaging diagrams are drawn. The self-energy $\mathbf{\Sigma}$ can be defined in a matrix equation equivalent to (31)

$$\mathbf{G}^{-1} = \lambda \mathbf{I} + \mathbf{\Sigma}. \tag{71}$$

The diagrams for the self-energy can be summed by 'dressing' both the response function and the pattern line A_{ij}. In figure 3(b) the dressing of the pattern line is shown. With this notation there are four diagrams that must be evaluated. We define these diagrams to be $\Sigma_{ij}^{(1)}, \ldots, \Sigma_{ij}^{(4)}$. The dressed pattern line in figure 3(b) is

$$\begin{aligned} A_{ij}^{\text{dressed}} &= A_{ij} - A_{ij} \sum_k \left[\frac{1}{N} x_k^\mu x_k^\mu \right]_\xi G_{kk} + A_{ij} \sum_{kl} \left[\frac{1}{N^2} x_k^\mu x_k^\mu x_l^\mu x_l^\mu \right]_\xi G_{ll} G_{kk} + \cdots \\ &= A_{ij} \left(1 - (1-a^2) \frac{1}{N} \operatorname{tr} G + \left((1-a^2) \frac{1}{N} \operatorname{tr} G \right)^2 - \cdots \right) \\ &= \frac{1}{1 + (1-a^2)(1/N) \operatorname{tr} G} A_{ij}. \end{aligned} \tag{72}$$

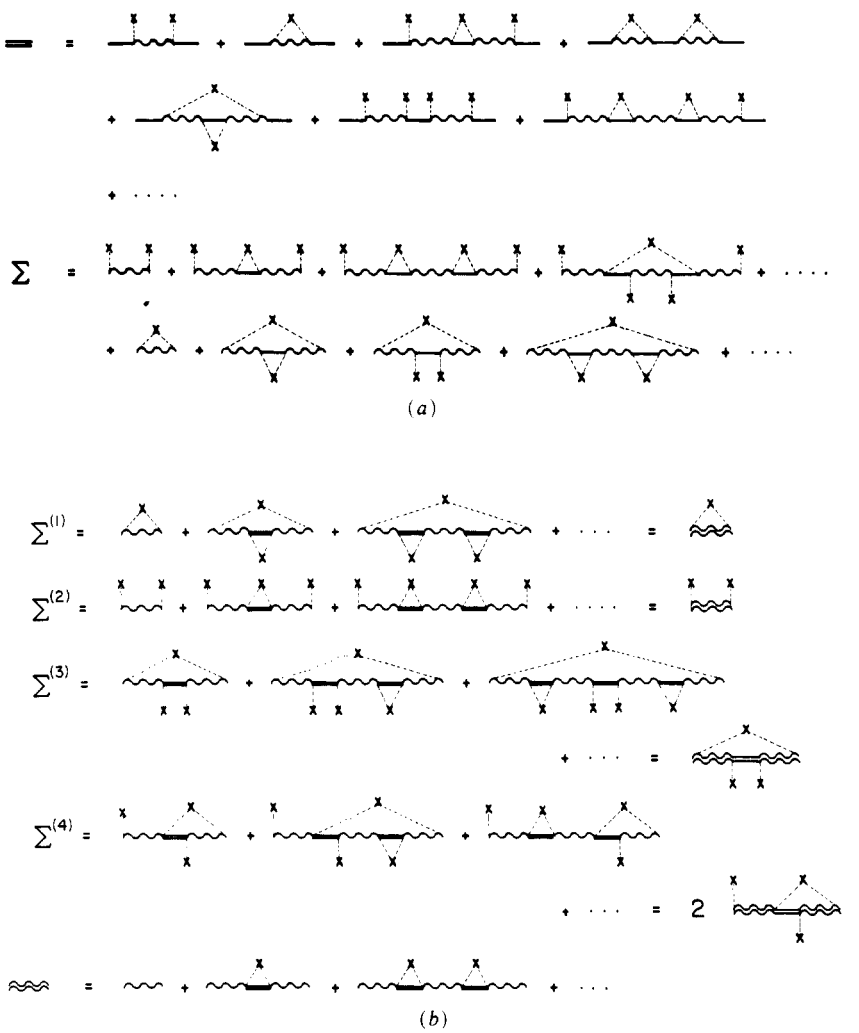

Figure 3. (*a*) Some of the diagrams for **G** when the patterns are correlated. A broken line ending in a × indicates an a. If two A are connected it means the same as before (uncorrelated patterns) except that ξ_i^μ is replaced by $x_i^\mu = \xi_i^\mu - a$. Some of the self-energy diagrams are also shown. (*b*) The four families of diagrams in the self-energy. They are expressed very compactly by dressing the A as shown in the bottom. The self-energy is $\Sigma = \Sigma^{(1)} + \cdots + \Sigma^{(4)}$, but only the first two are important, as described in the text.

We will now show that $\Sigma^{(3)}$ and $\Sigma^{(4)}$ can be ignored. Consider for the moment pattern lines that are not dressed. Then the two lowest-order diagrams in figure 3(*b*) give the contribution

$$\mathbf{\Sigma} = \alpha a^2 \mathbf{E} + \alpha(1 - a^2)\mathbf{I} \tag{73}$$

where $\mathbf{E}$ is the $N \times N$ matrix with every element equal to one and $\mathbf{I}$ is the $N \times N$ unit matrix. Since $\mathbf{G}^{-1}$ is given by (71) we have that its elements are

$$G_{ii}^{-1} = G_0^{-1} + \alpha \tag{74}$$

on the diagonal and

$$G_{ij}^{-1} = \alpha a^2 \tag{75}$$

off the diagonal. In this approximation the eigenvalues are $\eta_1 \equiv 1/\gamma_1 = G_0^{-1} + \alpha + (N-1)\alpha a^2$ and $\eta_2 \equiv 1/\gamma_2 = G_0^{-1} + \alpha(1-a^2)$. Obviously η_1 is much bigger than η_2 so γ_1 must be very small because of the factor N in η_1. We conclude that $\gamma_1 \simeq 0$ or $g_1 \simeq -g_0/N$. This allows us to drop the third term in $\boldsymbol{\Sigma}$ in figure 3(*b*). It is

$$\Sigma_{ij}^{(3)} = \frac{a^2}{N^2} \sum_{\mu lk} [x_i^\mu x_j^\mu]_\xi G_{lk}. \tag{76}$$

This involves the sum $\Sigma_k G_{lk}$ which is the small eigenvalue γ_1. A similar argument can be used for $\Sigma_{ij}^{(4)}$. It is then self-consistent to take $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)}$.

When the pattern lines have been dressed the self-energy matrix becomes

$$\boldsymbol{\Sigma} = R\alpha a^2 \mathbf{E} + R\alpha(1-a^2)\mathbf{I}. \tag{77}$$

The factor R comes from the dressing of the pattern line in (72). Since $\operatorname{tr} \mathbf{G} = (N-1)\gamma_2 + \gamma_1 \simeq N\gamma_2$ it is

$$R = \frac{1}{1+(1-a^2)\gamma_2}. \tag{78}$$

The eigenvalues of $\mathbf{G}^{-1}$ are now

$$\eta_1 = \lambda + R\alpha[1+(N-1)a^2] \tag{79}$$

and

$$\eta_2 = \lambda + R\alpha(1-a^2). \tag{80}$$

These eigenvalues describe the static properties of the system. In the limit of $a = 0$ (unbiased patterns) the two become identical and equal to the G^{-1} we found previously (34). Also note that the expressions for η_2 and G^{-1} are the same except for a scale factor $1-a^2$, which actually turns out to make the calculations of η_2 identical to our earlier calculations of G^{-1}.

7.2. Random output

For simplicity we consider first the case of random uncorrelated outputs. Then we can use (27) to find q:

$$q = (\partial/\partial\lambda)\lambda\gamma_2. \tag{81}$$

We now rescale, defining

$$\tilde{\gamma} = (1-a^2)\gamma_2 \qquad \tilde{\lambda} = \frac{\lambda}{(1-a^2)} \qquad \tilde{S}^2 = (1-a^2)S^2 \qquad \tilde{q} = (1-a^2)q \tag{82}$$

and get equations similar to (19), (34) and (81) to solve for $\tilde{\gamma}$:

$$\tilde{\gamma}^{-1} = \tilde{\lambda} + \alpha/(1+\tilde{\gamma}) \tag{83}$$

$$T\tilde{\gamma} = \tilde{S}^2 - \tilde{q} \tag{84}$$

$$\tilde{q} = (\partial/\partial\tilde{\lambda})\tilde{\lambda}\tilde{\gamma}. \tag{85}$$

These equations are the same as for the case of uncorrelated inputs and therefore the equation for $\tilde{\gamma}$ is the same as for G (37), but of course in terms of the rescaled variables.

The critical load capacity, α_c, now becomes (from (40))

$$\alpha_c = \frac{\tilde{S}}{1+\tilde{S}^2} = \frac{(1+a^2)S^2}{1+(1-a^2)S^2} \tag{86}$$

and all other properties—static and dynamic—governed by the second eigenvalue γ_2 scale accordingly. In particular the relaxation times of this mode are given by equations (53)-(55).

To find the relaxation time for the other mode determined by η_1, observe that from (79) and (80) it follows that

$$\eta_1 = \eta_2 + NR\alpha a^2. \tag{87}$$

Differentiating this and using that

$$\tau_2 = \frac{\gamma_2}{\gamma_0}\frac{\partial \eta_2}{\partial \lambda} \tag{88}$$

(see (50) and (51)) we find to leading order in $1/N$

$$\tau_1 = \frac{\gamma_1}{\gamma_0}\frac{\partial \eta_1}{\partial \lambda} = \tau_2 \frac{1-a^2}{1-a^2+\eta_2}. \tag{89}$$

The relative magnitude of the relaxation times is interesting. When $\eta_2 > 0$, which is always true for finite T, τ_1 is smaller than τ_2. We interpret τ_1 to be the time it takes the system to find a 'prototype solution' to the learning problem before finding the finer details.

7.3. General output

If the outputs are correlated we cannot use (81) to find q. We will not go into detailed calculations for other distributions of the outputs, but argue that it will always give basically the same results as above except when all the outputs (except from a finite number) are the same.

The Green function $f_{\mu\nu}$ defined in (23) is very similar to g_{ij} and the method of averaging is the same. The average $F_{\mu\nu} = [f_{\mu\nu}]_\xi$ looks exactly like G_{ij} except from some factors of α in various places. That means that (24) will be dominated by the large eigenvalue (corresponding to η_2) except when $\zeta^\mu \zeta^\nu = 1$ when the first eigenvalue is the only surviving part in (24). This is because the first eigenvalue has the eigenvector $(1, 1, 1, \ldots, 1)$.

This heuristic argument makes it reasonable to say that the small eigenvalue describes the relaxation of the system when essentially all the outputs are the same.

7.4. Unconstrained learning

When there is no constraint on the couplings we can proceed in a similar way as for uncorrelated patterns (see § 5). Notice that (80) for η_2 is of the same form as (56). By rescaling as above, $\tilde{\gamma} = \gamma_2(1-a^2)$ and $\tilde{z} = z/(1-a^2) = -i\omega/(1-a^2)$, equation (80) can be written

$$\tilde{\gamma}^{-1} = \tilde{z} + \alpha/(1+\tilde{\gamma}) \tag{90}$$

which is the same as (56). The eigenvalue η_2 gives us the relaxation time τ_2 (cf (60) and (61))

$$\tau_2 = \gamma_0^{-1} \hat{\gamma}_2(0) \lim_{z \to 0} \frac{\partial \hat{\eta}_2}{\partial z} \tag{91}$$

where $\hat{\gamma}_2$ is γ_2 with the pole subtracted. This equation holds for $\alpha < 1$ and for $\alpha > 1$ if γ_2 is substituted for $\hat{\gamma}_2$. As in § 5 we find

$$\tau_2 = \begin{cases} 1/[\gamma_0(1-a^2)(1-\alpha)^2] & \text{when } \alpha < 1 \\ \alpha/[\gamma_0(1-a^2)(\alpha-1)^2] & \text{when } \alpha > 1 \end{cases}. \tag{92}$$

Now the other relaxation time given by γ_1 is found. For $\alpha < 1$, γ_2 has a pole in $z = 0$:

$$\gamma_2 \simeq \frac{1-\alpha}{\tilde{z}} = \frac{(1-\alpha)(1-a^2)}{z}. \tag{93}$$

This pole carries over to γ_1 and has to be subtracted to find τ_1. From (87) γ_1 is found close to $z = 0$:

$$\gamma_1 = \frac{1+(1-a^2)\gamma_2}{N\alpha a^2 + \eta_2 + 1 - a^2} \simeq \frac{1+(1-a^2)\gamma_2}{N\alpha a^2}. \tag{94}$$

Subtracting the pole from γ_1 gives

$$\hat{\gamma}_1 = \frac{1+(1-a^2)\hat{\gamma}_2}{N\alpha a^2}. \tag{95}$$

For $\alpha > 1$ there is no pole and γ_1 is found from (87) to have the same expression to leading order in $1/N$ (just remove the hats in (95)).

The relaxation time is then found by differentiation:

$$\tau_1 = \begin{cases} \tau_2 \alpha & \text{when } \alpha < 1 \\ \tau_2/\alpha & \text{when } \alpha > 1 \end{cases}. \tag{96}$$

The critical value of the load parameter is $\alpha_c = 1$ as for uncorrelated inputs and again τ_1 is always less than or equal to τ_2.

8. Conclusion

Constrained weight decay and noise thus have very interesting effects on learning in simple perceptrons. In the model we have concentrated on most, where $[\langle w_i^2 \rangle]_{\xi\zeta}$ is constrained to a fixed value S^2, there is a critical point in the T, α plane at $\alpha = \alpha_c(S)$, $T = 0$. This is the transition between perfect and imperfect learning. At finite T (like finite external field in a ferromagnet) there is no transition. Learning is always imperfect ($\lambda \neq 0$), but there is a crossover which gets very sharp for low T.

Near the transition the learning time and asymptotic accuracy vary singularly strongly with the load (α) and noise (T) parameters, and there is always a complementarity between rapid and accurate learning. The critical region is a qualitatively optimal operating region for learning, where the competing priorities of speed and accuracy are well balanced. It therefore might be wise to tune S or λ so that α is near α_c.

In unconstrained learning, this transition is absent, and noise plays no role whatever. There is instead critical slowing down and a dynamical freezing transition at a new critical value, $\alpha = 1$, which is the maximum capacity for this algorithm.

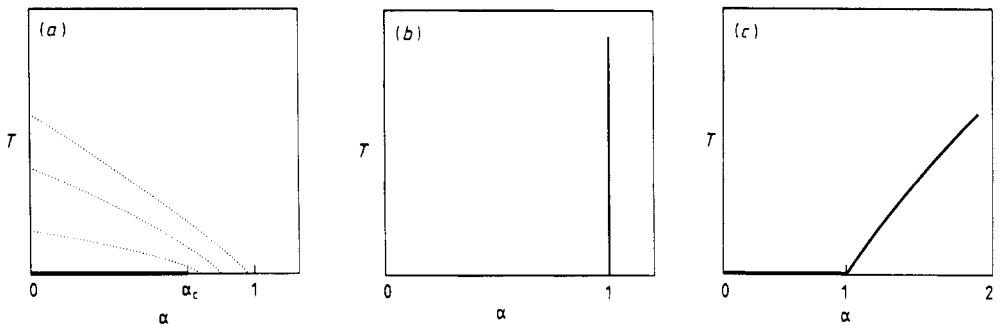

Figure 4. Phase diagrams for the three models discussed in the text. (*a*) Constraint $[\langle w_i^2\rangle]_{\xi\zeta} = S^2$. Along the heavy line the relaxation time is infinite. The dotted lines indicate qualitatively contours of constant τ. (*b*) Unconstrained learning. (*c*) Constained thermal fluctuations, $[\langle(w_i - \langle w_i\rangle)^2\rangle]_{\xi\zeta} = S^2$.

In the third model we studied, where the constraint is on the thermal fluctuations of w_i, independent of its equilibrium value, there is now a line of critical-slowing-down transitions in the α, T plane, ending for $T \to 0$ at the capacity $\alpha = 1$.

These results are summarised in the 'phase diagrams' (in α, T space) of figure 4.

We have succeeded in generalising much of this picture to the learning of correlated patterns. The important new feature is the emergence of a second relaxation time, which can be interpreted as the time it takes to learn a prototype (mean) of the group of patterns.

It would of course be highly desirable to see what kinds of phase transition occur in learning in less trivial networks with non-linear units and more layers. Perhaps the discovery of such rich structure in the simple linear perceptron can be a guide to some things to look for in these systems.

Appendix: The diagram technique

A1. Uncorrelated patterns

In G we have averages of the form (here shown for four A)

$$\left[N^{-4} \sum_{\mu_1 \cdots \mu_4} \sum_{i_1, i_2, i_3} \xi_i^{\mu_1} \xi_{i_1}^{\mu_1} \xi_{i_1}^{\mu_2} \xi_{i_2}^{\mu_2} \xi_{i_2}^{\mu_3} \xi_{i_3}^{\mu_3} \xi_{i_3}^{\mu_4} \xi_j^{\mu_4}\right]_\xi . \tag{A1}$$

We have to pair the ξ to get anything different from zero, since

$$[\xi_i^\mu \xi_j^\nu]_\xi = \delta_{ij}\delta_{\mu\nu}. \tag{A2}$$

In the diagrams such a pairing is shown by connecting the terms by a broken line with a $\times$. For example, putting $\mu_1 = \mu_2$ and $i = j$ in (A1) corresponds to connecting the two ends of the diagram for that term. All ends of the A have to be paired.

The first important observations is that 'crossing' diagrams are of order $1/N$ or less compared to the non-crossing ones. By crossing we mean that the broken lines cross as in the diagram shown in figure 5(*b*). Calculating the size of the diagrams in figure 5(*a*) gives (except for λ^{-5})

$$\left[N^{-4} \sum_{\mu_1, \mu_2} \sum_{i_1, i_2} \xi_i^{\mu_1} \xi_{i_1}^{\mu_1} \xi_{i_1}^{\mu_2} \xi_{i_1}^{\mu_2} \xi_{i_1}^{\mu_1} \xi_{i_2}^{\mu_1} \xi_{i_2}^{\mu_1} \xi_i^{\mu_1}\right]_\xi = N^{-4} \sum_{\mu_1, \mu_2} \sum_{i_1, i_2} 1 = \alpha^2 \tag{A3}$$

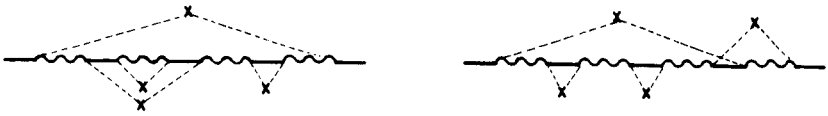

Figure 5. The two diagrams described in the text. (*a*) contributes to G but (*b*) does not because it is a crossing diagram.

and for the diagram in figure 5(*b*) gives

$$\left[N^{-4}\sum_{\mu}\sum_{i_1,i_2}\xi_i^\mu\xi_{i_1}^\mu\xi_{i_1}^\mu\xi_{i_2}^\mu\xi_{i_2}^\mu\xi_i^\mu\xi_i^\mu\xi_i^\mu\right]_\xi = N^{-4}\sum_{\mu}\sum_{i_1,i_2}1=\alpha/N. \tag{A4}$$

From now on we therefore ignore all crossing diagrams.

The second important observation is that we always have to put $i=j$ for the pairing to be possible without getting crossing graphs. That makes all the terms diagonal and then G is diagonal also.

This diagram technique counts some special parts of the average many times. For example, the term from having $i=i_1=i_2=i_3$ and $\mu_1=\cdots=\mu_4$ in the average (A1) is included in *all* the diagrams we draw with four A. It is easy to see that in the limit of large p double-counting terms like this are unimportant.

By *reducible diagrams* we mean diagrams that can be split into two or more diagrams without breaking any dotted lines. The sum of all possible (surviving) irreducible diagrams is called the self-energy, Σ and is shown in figure 2 where it is also shown how G can be written in terms of Σ. The power series for G is now easily summed:

$$G=G_0\left[1-G_0\Sigma+(G_0\Sigma)^2-(G_0\Sigma)^3+\cdots\right]=\frac{G_0}{1+G_0\Sigma} \tag{A5}$$

which can also be written as

$$G^{-1}=G_0^{-1}+\Sigma. \tag{A6}$$

The self-energy can be expressed in terms of 'dressed' diagrams as shown in figure 2. Dressing means that the diagrams are changed by using double lines for G instead of lines for λ^{-1}. This makes the series much simpler and the averages occurring (32) are now trivial.

A.2. Correlated patterns

The trick here is to write the patterns ξ_i^μ as a sum of their average a and the rest, x_i^μ, which then has average zero and variance

$$[x_i^\mu x_j^\nu]_\xi=(1-a^2)\delta_{ij}\delta_{\mu\nu}. \tag{A7}$$

As before, the x_i^μ has to be paired and it is drawn the same way as above. Having an a appearing in an average is drawn with a dotted line ending in a $\times$. Then, as an example, the average

$$\left[\sum_{\mu_1,\mu_2}\sum_k(a+x_i^{\mu_1})(a+x_k^{\mu_1})(a+x_k^{\mu_2})(a+x_i^{\mu_2})\right]_\xi \tag{A8}$$

can be drawn as in figure 6. Here we have not drawn terms that have an odd number of x—they will always average to zero—and crossing graphs that can again be neglected.

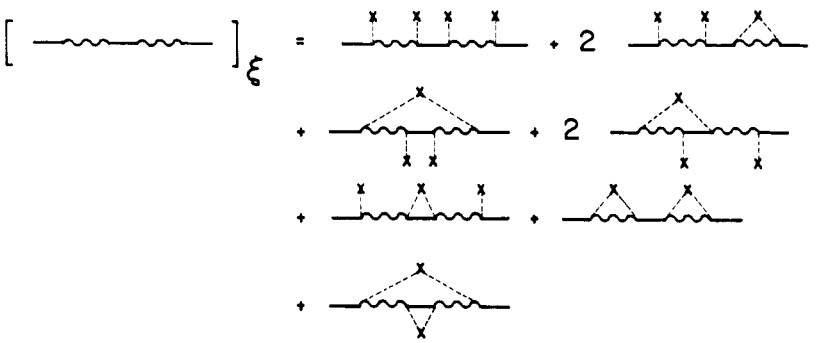

Figure 6. All diagrams with two As that contribute to G.

The self-energy is defined as before but has off-diagonal elements also, so all the above equations should now be read as matrix equations. Also G is of course a matrix and therefore the double lines in the diagrams now carry different indices in the two ends.

There are now four families of graphs in the self-energy. A few examples from each family are drawn in figure 3(b). Only the first two will survive as described in § 7.

The first family is the same as we had in the case of uncorrelated patterns, thus the only one that exists in the limit of small a. The second family is the off-diagonal equivalent of the first. By 'renormalising' G all families can be expressed very compactly as also shown in figure 3(b).

References

[1] Rumelhart D E, Hinton G E and Williams R J 1987 *Parallel Distributed Processing* ed D E Rumelhart and J L McClelland (Cambridge, MA: MIT Press) p 318
[2] Hertz J A, Thorbergsson G I and Krogh A 1989 *Phys. Scr.* **T25** 149
[3] Ma S-K 1976 *Modern Theory of Critical Phenomena* (New York: Benjamin)
[4] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
[5] Kohonen T 1984 *Self-organisation and Associate Memory* (Berlin: Springer)
[6] Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359
[7] Landau L D and Lifshitz E M 1980 *Statistical Physics* (Oxford: Pergamon)
[8] Opper M 1989 *Europhys. Lett.* **8** 389