TEN

Formal Statistical Mechanics of Neural Networks

As we have stressed in earlier chapters, neural networks are large interacting systems of simple units, like the physical systems we study in statistical mechanics. The formal methods and concepts of statistical physics are therefore natural tools to use for neural networks. In this chapter we illustrate the use of such methods in two different problems that we encountered earlier in the book: the recall of stored patterns in the Hopfield associative memory network, and the capacity of a simple perceptron.

This chapter is not for everyone. Up to this point, this book has been a fairly general introduction to the theory of neural networks, and has not required specialized knowledge of formal techniques. While what follows is also self-contained, it is much more formal mathematically, and readers without other exposure to techniques of this sort will probably find it hard going. We include these calculations here anyway because they illustrate how such theoretical methods can be brought to bear on problems in neural computation, with nontrivial results.

We do assume in this chapter some basic knowledge of statistical mechanics. The necessary ideas are reviewed briefly in the Appendix.

10.1 The Hopfield Model

In Chapter 2 we described the stochastic Hopfield model and obtained a number of its properties heuristically. The starting points were the Hebb rule (2.9) for the connection strengths, and the dynamics based on the stochastic evolution rule (2.40). We then calculated the average activations $\langle S_i \rangle$ in the heuristically motivated

mean field scheme (2.50). Here we will take a more systematic approach, obtaining the quantities we want by first calculating the **partition function**

$$Z = \mathrm{Tr}_S \exp(-\beta H\{S_i\}) \tag{10.1}$$

where the trace, Tr_S, means a sum over all possible states, $\{S_i = \pm 1\}$.[1] We can then take appropriate derivatives to obtain quantities of more direct interest, as outlined in the Appendix. Our treatment follows closely the classic article by Amit, Gutfreund, and Sompolinsky [1987a].

We start with the energy function

$$H_0 = -\frac{1}{2N}\sum_{\mu=1}^{p}\Big(\sum_i S_i\xi_i^\mu\Big)^2 + \frac{p}{2} \tag{10.2}$$

which is the same as (2.30) except for the second term (a constant), which cancels out the diagonal S_i^2 contributions from the first term. As we showed in (2.31), this energy function has the Hebbian connection strengths (2.9).

We again define $\alpha = p/N$, the ratio of the number of desired memories to the size of the system. We consider only the large N limit (the so-called **thermodynamic limit** $N \to \infty$), so when we discuss $\alpha \neq 0$ we mean that the number of patterns p scales proportionally with the number of units N. But first we discuss the simpler case $\alpha = 0$ with a fixed number p of patterns, independent of N.

Mean Field Theory for $\alpha = 0$

We start by adding to H_0 a set of "external fields" $h^\mu\xi_i^\mu$, one for each pattern ξ_i^μ:

$$H = H_0 - \sum_\mu h^\mu \sum_i \xi_i^\mu S_i\,. \tag{10.3}$$

We will set all the strengths h^μ to zero later, after these fields have served their purpose.

The partition function (10.1) is now

$$Z = e^{-\beta p/2}\,\mathrm{Tr}_S \exp\left[\frac{\beta}{2N}\sum_\mu\Big(\sum_i S_i\xi_i^\mu\Big)^2 + \beta\sum_\mu h^\mu\sum_i \xi_i^\mu S_i\right]. \tag{10.4}$$

This would be easy to evaluate if both the terms in the exponent were linear in the S_i's. Then the trace would simply factor into a product of N independent terms, one for each i, every term being a simple sum over $S_i = +1$ and $S_i = -1$.

[1] Usually the trace of an operator (or matrix) means the sum of all the diagonal elements. This way of using it originates in quantum statistics.

Unfortunately the first term is quadratic. But we can use the "Gaussian integral trick" to make it linear, at the expense of some other complications. This trick exploits the identity

$$\int_{-\infty}^{\infty} dx\, e^{-ax^2 \pm bx} = \sqrt{\pi/a}\, e^{b^2/4a} \tag{10.5}$$

which can be used to turn an exponential in b^2 (on the right) into an exponential linear in b (on the left). The cost, of course, is the introduction of the auxiliary variable x, and the integral over it. Our price is actually p times higher, because (10.4) contains p quadratic terms, one for each μ. So we introduce p auxiliary variables m^μ, and take $a = \beta N/2$ and $b^\mu = \beta \sum_i S_i \xi_i^\mu$ to give

$$\begin{aligned} Z &= e^{-\beta p/2} \left(\frac{\beta N}{2\pi}\right)^{p/2} i \\ &\quad \times \mathrm{Tr}_S \prod_\mu \int dm^\mu \exp\left(-\tfrac{1}{2}\beta N (m^\mu)^2 + \beta(m^\mu + h^\mu) \sum_i \xi_i^\mu S_i\right). \end{aligned} \tag{10.6}$$

Now let us adopt a shorthand vector notation, taking $\mathbf{m}$, $\mathbf{h}$, and $\boldsymbol{\xi}_i$ to be p-component vectors with components m^μ, h^μ, and ξ_i^μ respectively. Then (10.6) becomes

$$Z = e^{-\beta p/2} \left(\frac{\beta N}{2\pi}\right)^{p/2} \int d\mathbf{m}\, e^{-\beta N \mathbf{m}^2/2} \prod_i \mathrm{Tr}_{S_i}\, e^{\beta(\mathbf{m}+\mathbf{h})\cdot\boldsymbol{\xi}_i S_i}. \tag{10.7}$$

The trace is now easy because the exponent is linear in S_i. Using $e^x + e^{-x} = 2\cosh x$, we obtain after a little reorganization

$$Z = \left(\frac{\beta N}{2\pi}\right)^{p/2} \int d\mathbf{m}\, e^{-\beta N f(\beta, \mathbf{m})} \tag{10.8}$$

with

$$f(\beta, \mathbf{m}) = \tfrac{1}{2}\alpha + \tfrac{1}{2}\mathbf{m}^2 - \frac{1}{\beta N} \sum_i \log\left(2\cosh[\beta(\mathbf{m}+\mathbf{h})\cdot\boldsymbol{\xi}_i]\right). \tag{10.9}$$

We still have a p-fold integral to do, but the fact that the exponent in (10.8) is proportional to N allows us to evaluate it in the limit of large N. The bigger N is, the more the integral is dominated by contributions from the region where f is smallest. So we can approximate it by finding the value of $\mathbf{m}$ which minimizes f, and expanding the integrand around there. This is called the **saddle-point method**, and is best understood through a simple example.

Suppose that we had a one-dimensional integral of the form

$$I = \sqrt{N} \int dx\, e^{-Ng(x)}. \tag{10.10}$$

Then expanding the exponent around the point x_0 where $g(x)$ is minimized we get

$$I = \sqrt{N} \int dx \exp\left(-N[g(x_0) + \tfrac{1}{2}g''(x_0)(x - x_0)^2 + \cdots]\right) \tag{10.11}$$

using $g'(x_0) = 0$. If we truncate the expansion at this point, the integral is just a Gaussian one, so

$$I = \sqrt{N}e^{-Ng(x_0)}\sqrt{\frac{2\pi}{Ng''(x_0)}} = e^{-Ng(x_0)}\sqrt{\frac{2\pi}{g''(x_0)}}. \tag{10.12}$$

For large N this result is dominated by the exponential factor, as can be clearly seen by putting it in the form

$$\begin{aligned} -\frac{1}{N}\log I &= g(x_0) + \frac{1}{2N}(\log g''(x_0) - \log 2\pi) \\ &\xrightarrow{N\to\infty} g(x_0). \end{aligned} \tag{10.13}$$

Thus all we need to do is to find x_0; this is often called the **saddle point**, from behavior in the complex x plane.

For (10.8) we use a p-dimensional version of the same idea, thereby obtaining

$$-\frac{1}{N}\log Z = \beta \min_{\mathbf{m}} f(\beta, \mathbf{m}) \tag{10.14}$$

in the $N \to \infty$ limit. Comparing this with (A.9) we see that

$$F/N = \min_{\mathbf{m}} f(\beta, \mathbf{m}) \tag{10.15}$$

where F is the free energy, so $\min_{\mathbf{m}} f(\beta, \mathbf{m})$ gives us the free energy per unit.

We now have to minimize $f(\beta, \mathbf{m})$, which requires

$$0 = \frac{\partial f}{\partial m^\mu} = m^\mu - \frac{1}{N}\sum_i \xi_i^\mu \tanh[\beta(\mathbf{m}+\mathbf{h})\cdot\boldsymbol{\xi}_i]. \tag{10.16}$$

Note that this is a set of p nonlinear simultaneous equations for the p unknowns m^μ. These equations appear to depend on the random patterns $\boldsymbol{\xi}_i$, but in fact the system is **self-averaging**; we can replace the average $N^{-1}\sum_i$ over *units* by an average over *patterns* at any one site, yielding

$$m^\mu = \langle\!\langle \xi^\mu \tanh[\beta(\mathbf{m}+\mathbf{h})\cdot\boldsymbol{\xi}] \rangle\!\rangle \tag{10.17}$$

where $\langle\!\langle \cdots \rangle\!\rangle$ indicates an average over the random distribution of $\boldsymbol{\xi}$ patterns. Similarly (10.9) becomes (with $\alpha \to 0$)

$$f = \tfrac{1}{2}\mathbf{m}^2 - \beta^{-1}\langle\!\langle \log(2\cosh[\beta(\mathbf{m}+\mathbf{h})\cdot\boldsymbol{\xi}]) \rangle\!\rangle. \tag{10.18}$$

It is easy to see how the self-averaging property arises. As we go from unit to unit in the sum on i in (10.9) or (10.16), we are choosing N independent $\boldsymbol{\xi}_i$'s from the distribution $P(\boldsymbol{\xi})$, which we take to be uniform over the 2^p possibilities. So if N

is large compared to 2^p the average over sites is equivalent to an average over the distribution. This requires $p \ll \log N$, which is valid in our present $\alpha = 0$ case, but not for the $\alpha \neq 0$ case considered later.

The values of m^μ at the saddle point given by (10.17) admit a simple physical interpretation. To see this, we start from the free energy $F = -\beta^{-1} \log Z$ and differentiate with respect to h^μ. Using the original expression (10.4) for Z leads—as in (A.10)—to

$$\frac{\partial F}{\partial h^\mu} = -\beta^{-1} \frac{\partial \log Z}{\partial h^\mu} = -\sum_i \langle S_i \rangle \xi_i^\mu \tag{10.19}$$

whereas (10.15), (10.17), and (10.18) give us

$$\frac{\partial F}{\partial h^\mu} = N \frac{\partial f}{\partial h^\mu} = -N \langle\!\langle \xi^\mu \tanh[\beta(\mathbf{m} + \mathbf{h}) \cdot \boldsymbol{\xi}] \rangle\!\rangle = -N m^\mu . \tag{10.20}$$

We can thus identify

$$m^\mu = \frac{1}{N} \sum_i \xi_i^\mu \langle S_i \rangle \tag{10.21}$$

so the saddle-point value of m^μ is just the average overlap of the network configuration with pattern number μ.

It was to derive (10.21) that we needed the external field terms inserted in (10.3). Now they are no longer needed and we henceforth set $h^\mu = 0$. Thus the mean field equation (10.17) becomes simply

$$m^\mu = \langle\!\langle \xi^\mu \tanh(\beta \mathbf{m} \cdot \boldsymbol{\xi}) \rangle\!\rangle . \tag{10.22}$$

There are many solutions of (10.22). The simplest and most important are the **memory states**, which have a finite correlation with just *one* of the patterns ξ_i^μ. So from (10.21) we expect the $\mathbf{m}$ vector for these solutions to have the form

$$\mathbf{m} = (m, 0, 0, \ldots) \tag{10.23}$$

if we order the indices μ so that the "condensed" pattern is first. Then (10.22) reduces to

$$m^\mu = \langle\!\langle \xi^\mu \tanh \beta m \xi^1 \rangle\!\rangle = \langle\!\langle \xi^\mu \xi^1 \rangle\!\rangle \tanh \beta m = \delta_{\mu 1} \tanh \beta m . \tag{10.24}$$

So (10.23) *does* give a solution—putting (10.23) into the right-hand side of (10.22) produces the same form on the left—provided the magnitude m of the average overlap with pattern 1 satisfies

$$m = \tanh \beta m . \tag{10.25}$$

This is identical to the equation (2.54) that we found in our simpler analysis in Chapter 2. It implies stable memory states for $T < T_c$ with $T_c = 1$, and tells us what fraction of the bits will be correct at any such temperature; see Fig. 2.14.

TABLE 10.1
Critical Temperatures

n	T_n
1	1
3	0.46
5	0.39
7	0.35

There are also more complicated solutions of the mean field equations, corresponding to the **spurious states**. The simplest of these are the **symmetric mixture states** in which the $\mathbf{m}$ vector has the form

$$\mathbf{m} = (\underbrace{m, m, m, \ldots, m}_{n}, \underbrace{0, 0, \ldots, 0}_{p-n}) \tag{10.26}$$

with n nonzero entries of $\mathbf{m}$ equal to some value m. Note that there are $\binom{n}{p}$ ways we might have placed the nonzero elements, corresponding to many such spurious states. There is actually a further degeneracy factor of 2^n, because solutions like $(\pm m, \pm m, \pm m, 0, \ldots, 0)$ are all possible.

If we insert the form (10.26) into the mean field equations (10.22) we obtain

$$m^\mu = \left\langle\!\left\langle \xi^\mu \tanh\left(\beta m \sum_{\nu=1}^{n} \xi^\nu\right)\right\rangle\!\right\rangle \tag{10.27}$$

which vanishes if $\mu > n$ (because $\langle\!\langle \xi^\mu \xi\nu \rangle\!\rangle = 0$ for $\mu \neq \nu$), and otherwise gives

$$m = \langle\!\langle z \tanh \beta m z \rangle\!\rangle / n \tag{10.28}$$

where z is the random variable

$$z = \sum_{\mu=1}^{n} \xi^\mu \tag{10.29}$$

which has a binomial distribution. Thus our symmetric combination pattern (10.26) solves the mean field equations if m satisfies (10.28). This has solutions at any n, as long as $T < 1$.

However, not all these solutions are stable. We want $\mathbf{m}$ to produce a *minimum* of $f(\beta, \mathbf{m})$, whereas our mean field equations only guarantee a stationary point, $\partial f / \partial m^\mu = 0$. So we also need the eigenvalues of the matrix

$$A_{\mu\nu} = \frac{\partial^2 f}{\partial m^\mu \partial m^\nu} \tag{10.30}$$

to be positive. This turns out to be satisfied only if n is odd, and then only if the temperature T is below a **critical temperature** T_n. The first few T_n's are shown in table 10.1.

There are also **asymmetric mixture states**, such as

$$\mathbf{m} = (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, 0, 0, 0, \ldots). \tag{10.31}$$

None of these is stable above T_3, however. This means that one can avoid *all* mixture states by going to temperatures above T_3. Of course raising the temperature from $T = 0$ degrades the memory states somewhat, but the amount is actually very small; $\langle N_{\text{correct}} \rangle$ falls only very slowly from N with increasing T, as seen in Fig. 2.14. At $T = 0.47$ we find from (10.25) that $\langle N_{\text{correct}} \rangle \approx 0.97N$, so only about 3% of the bits will be recalled incorrectly if we work just above T_3.

Mean Field Theory for $\alpha \neq 0$

As we observed already in Chapter 2, the crosstalk between different patterns on account of their random overlap begins to affect the recall of a given pattern when p becomes of the order of N. We now examine the statistical mechanics for this case. The self-averaging we used in the $\alpha = 0$ calculation breaks down, and we are forced to do the averaging over the distribution of patterns more systematically.

As always, the basic quantity we start from is $\log Z$. Now Z depends on the particular set of patterns used to compute the weights w_{ij} using the Hebb rule (2.9). What is of interest to us is the *average* $\langle\langle \log Z \rangle\rangle$ over the distribution of all random binary patterns; this gives us the average free energy whose derivatives give the average quantities we want to know, such as m^μ. Unfortunately this average is very hard to calculate directly, and is *not* the same thing as $\log\langle\langle Z \rangle\rangle$, which would be much easier. To get meaningful results we must average the relevant quantity, which is $\log Z$, not Z.

Luckily there is a technique, called the **replica method**, that lets us circumvent averaging $\log Z$. It is based on the identity

$$\log Z = \lim_{n \to 0} \frac{Z^n - 1}{n} \tag{10.32}$$

which allows us to compute $\langle\langle \log Z \rangle\rangle$ from knowledge of $\langle\langle Z^n \rangle\rangle$. Note that we need this for the parameter n close to 0, but we ignore that for a while and focus on $\langle\langle Z^n \rangle\rangle$ for *integer* n. In that case we can think of Z^n as the partition function of n copies, or **replicas**, of the original system, writing

$$Z^n = \mathrm{Tr}_{S^1} \mathrm{Tr}_{S^2} \ldots \mathrm{Tr}_{S^n} e^{-\beta(E\{S_i^1\} + \cdots + E\{S_i^n\})}. \tag{10.33}$$

Each copy is labelled by a superscript **replica index** on its S_i's, running from 1 to n.

Proceeding as we did in the $\alpha = 0$ case for (10.6), using the Gaussian integral trick for each pattern and each replica, we now find

$$\langle\langle Z^n \rangle\rangle = e^{-\beta p n/2} \Big\langle\Big\langle \mathrm{Tr}_S \prod_{\mu=1}^{p} \prod_{\rho=1}^{n} \int dm_\rho^\mu \Big(\frac{\beta N}{2\pi}\Big)^{1/2} \times \exp\Big(-\tfrac{1}{2}\beta N (m_\rho^\mu)^2 + \beta m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho\Big)\Big\rangle\Big\rangle \tag{10.34}$$

where ρ labels the different replicas. Note that the pattern average $\langle\langle\cdots\rangle\rangle$ is still over the Np variables ξ_i^μ; there is no replica index on the patterns. We have omitted any external fields h^μ this time, although it would be easy to include them.

Henceforth we focus on states in which the configuration has appreciable overlap with only a *finite* number s of the p stored patterns, called the **condensed patterns**. Specifically we assume that the m_ρ^μ's[2] are only appreciable in size when $\mu \le s$, with s independent of N. This will eventually allow us to use the self-averaging trick just on these μ's. For $\mu > s$ we assume $m_\rho^\mu \ll 1$.

Let us consider the contribution of the last term in the integrand of (10.34) for a particular $\mu > s$, one of the *small* m_ρ^μ's:

$$\begin{aligned}
\Big\langle\Big\langle \prod_\rho \exp\Big(\beta m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho\Big)\Big\rangle\Big\rangle &= \prod_i \Big\langle\Big\langle \exp\Big(\beta \xi_i^\mu \sum_\rho m_\rho^\mu S_i^\rho\Big)\Big\rangle\Big\rangle \\
&= \prod_i \cosh\Big(\beta \sum_\rho m_\rho^\mu S_i^\rho\Big) \\
&= \exp\Big[\sum_i \log\cosh\Big(\beta \sum_\rho m_\rho^\mu S_i^\rho\Big)\Big] \\
&\approx \exp\Big[\sum_i \frac{1}{2}\Big(\beta \sum_\rho m_\rho^\mu S_i^\rho\Big)^2\Big] \\
&= \exp\Big(\frac{\beta^2}{2} \sum_{i\rho\sigma} S_i^\rho S_i^\sigma m_\rho^\mu m_\sigma^\mu\Big)
\end{aligned} \tag{10.35}$$

where the approximation involved $\log\cosh x \approx x^2/2$ for small x. If we now define an $n \times n$ matrix $\tilde{\Lambda}_{\rho\sigma}$ by

$$\tilde{\Lambda}_{\rho\sigma} \equiv \delta_{\rho\sigma} - (\beta/N) \sum_i S_i^\rho S_i^\sigma \tag{10.36}$$

we can write the whole exponential factor (for fixed $\mu > s$) in (10.34) as

$$E = \exp\Big(-\frac{\beta N}{2} \sum_{\rho\sigma} \tilde{\Lambda}_{\rho\sigma} m_\rho^\mu m_\sigma^\mu\Big) . \tag{10.37}$$

[2]Strictly speaking: the *saddle-point* values of the m_ρ^μ's. That is, we will again evaluate the multi-dimensional integral by the saddle-point method, and the values of the m_ρ^μ that will matter will be those at the saddle point.

This leaves us with an n-dimensional Gaussian integral, giving

$$\int \Big(\prod_\rho dm_\rho^\mu \Big(\frac{\beta N}{2\pi}\Big)^{1/2}\Big) E \;=\; \Big(\frac{\beta N}{2\pi}\Big)^{n/2} \sqrt{\frac{\pi^n}{\det(\frac{1}{2}\beta N\tilde{\Lambda})}} \;=\; (\det\tilde{\Lambda})^{-1/2}. \quad (10.38)$$

We get a contribution exactly like this for every value of μ greater than s (about p in all, since $p \gg s$), giving an overall factor

$$\begin{aligned}(\det\tilde{\Lambda})^{-p/2} \;&=\; \exp\big(-\tfrac{1}{2}p\log\det\tilde{\Lambda}\big) \;=\; \exp\Big(-\tfrac{1}{2}p\log\prod_\rho \tilde{\lambda}_\rho\Big) \\ &=\; \exp\Big(-\tfrac{1}{2}p\sum_\rho \log\tilde{\lambda}_\rho\Big) \end{aligned} \quad (10.39)$$

where $\tilde{\lambda}_\rho$ are the eigenvalues of $\tilde{\Lambda}$.

The extra complications we encounter for $\alpha > 0$ all come from this factor (10.39), which, together with the other parts of (10.34), now has to be summed over all the S_i^ρ. Unfortunately the S-dependence is buried in the eigenvalues $\tilde{\lambda}_\rho$, and the trace is far from easy. So now we use some more auxiliary variable tricks. First let us define a generalized version of $\tilde{\Lambda}_{\rho\sigma}$:

$$\Lambda_{\rho\sigma} \equiv (1-\beta)\delta_{\rho\sigma} - \beta q_{\rho\sigma}\,. \quad (10.40)$$

This is equal to $\tilde{\Lambda}_{\rho\sigma}$ if

$$q_{\rho\sigma} = \begin{cases} N^{-1}\sum_i S_i^\rho S_i^\sigma & \text{for } \rho \neq \sigma\,; \\ 0 & \text{otherwise.} \end{cases} \quad (10.41)$$

Thus we can write any function $G\{\tilde{\lambda}_\rho\}$ of the eigenvalues of $\tilde{\Lambda}$ in the form

$$G\{\tilde{\lambda}_\rho\} = \int\Big[\prod_{(\rho\sigma)} dq_{\rho\sigma}\,\delta\Big(q_{\rho\sigma} - \frac{1}{N}\sum_i S_i^\rho S_i^\sigma\Big)\Big] G\{\lambda_\rho\} \quad (10.42)$$

using a Dirac delta function, where the λ_ρ's are the eigenvalues of Λ, and are functions of the $q_{\rho\sigma}$'s. There are $n(n-1)/2$ integrals (the notation $(\rho\sigma)$ means all distinct pairs), and we leave it as understood that $q_{\rho\sigma} = q_{\sigma\rho}$ and $q_{\rho\rho} = 0$.

Now we introduce yet another set of auxiliary variables, this time for an integral representation of the delta-function:

$$\delta(x) = \int_{-i\infty}^{i\infty} \frac{d\tilde{r}}{2\pi i}\, e^{-\tilde{r}x}. \quad (10.43)$$

We need to use this $n(n-1)/2$ times, giving us

$$G\{\tilde{\lambda}_\rho\} \propto \int\Big[\prod_{(\rho\sigma)} dq_{\rho\sigma}dr_{\rho\sigma}\,\exp\Big(-N\alpha\beta^2 r_{\rho\sigma}q_{\rho\sigma} + \alpha\beta^2 r_{\rho\sigma}\sum_i S_i^\rho S_i^\sigma\Big)\Big] G\{\lambda_\rho\} \quad (10.44)$$

where we have left out unimportant prefactors and scaled the r variables by a factor of $N\alpha\beta^2$ for later convenience.

When we apply the transformation (10.44) to (10.39), we can write our full expression (10.34) for $\langle\langle Z^n \rangle\rangle$ as

$$\begin{aligned}\langle\langle Z^n \rangle\rangle \propto\ & e^{-\beta p n/2} \int \Big(\prod_{\mu\rho} dm_\rho^\mu \Big(\frac{\beta N}{2\pi} \Big)^{1/2} \Big) \Big(\prod_{(\rho\sigma)} dq_{\rho\sigma} dr_{\rho\sigma} \Big) \\ & \times \exp\Big(-\tfrac{1}{2}\beta N \sum_{\mu\rho} (m_\rho^\mu)^2 - \frac{\alpha N}{2} \sum_\rho \log \lambda_\rho - \tfrac{1}{2} N\alpha\beta^2 \sum_{\rho\sigma} r_{\rho\sigma} q_{\rho\sigma} \Big) \\ & \times \Big\langle\!\Big\langle \mathrm{Tr}_S \exp\Big(\beta \sum_{\mu\rho} m_\rho^\mu \sum_i \xi_i^\mu S_i^\rho + \tfrac{1}{2}\alpha\beta^2 \sum_{i\rho\sigma} r_{\rho\sigma} S_i^\rho S_i^\sigma \Big) \Big\rangle\!\Big\rangle \end{aligned} \tag{10.45}$$

where the sums over μ now run only over the condensed patterns: $\mu = 1, 2, \ldots, s$. We have also written $\frac{1}{2}\sum_{\rho\sigma}$ instead of $\sum_{(\rho\sigma)}$ and again left it understood that diagonal $\rho\rho$ terms are zero.

Now at last we can get rid of the i indices through self-averaging. The last line of (10.45) is the pattern average of an expression with the form

$$X \equiv \mathrm{Tr}_S \exp\Big(\sum_i F\{\xi_i, S_i\} \Big) \tag{10.46}$$

$$= \prod_i \mathrm{Tr}_{S_i} \exp F\{\xi_i, S_i\} \tag{10.47}$$

$$= \exp\Big(\sum_i \log \mathrm{Tr}_{S_i} \exp F\{\xi_i, S_i\} \Big) . \tag{10.48}$$

The function F depends on ξ_i^1–ξ_i^s and S_i^1–S_i^n, but only one index i is needed at a time. The trace in (10.46) is over all the S_i^ρ's (all i's and all ρ's), but in (10.47) and (10.48) is over only the S_i^ρ's for a particular i. The result of the trace would be exactly the same for each i except for the dependence of F on the ξ_i^μ's, because i is otherwise a dummy index. But since N (the number of i's) is much larger as $N \to \infty$ than 2^s (the number of possible sets $\{\xi_i^\mu\}$ at fixed i), the sum over i is equivalent to an average over patterns. Thus

$$X = \exp\Big(N \big\langle\!\big\langle \log \mathrm{Tr}_S \exp F\{\xi_i, S_i\} \big\rangle\!\big\rangle \Big) \tag{10.49}$$

where now all i's have disappeared and we have in effect a single unit with n different S^ρ's and p different ξ^μ's. Note that in the end we did not need the outer average $\langle\langle \cdots \rangle\rangle$ in (10.45), because the self-averaging of the inner i sum already performs all the pattern averaging. So we may drop the outer average.

Now we can write the whole expression for $\langle\langle Z^n \rangle\rangle$ as an integral of the exponential of something proportional to N:

$$\langle\langle Z^n \rangle\rangle \propto e^{-\beta p n/2} \int \Big(\prod_{\mu\rho} dm_\rho^\mu \Big(\frac{\beta N}{2\pi} \Big)^{1/2} \Big) \Big(\prod_{(\rho\sigma)} dq_{\rho\sigma} dr_{\rho\sigma} \Big) e^{-N\beta f\{m,q,r\}} \tag{10.50}$$

where

$$f\{m,q,r\} = \tfrac{1}{2}\sum_{\mu\rho}(m_\rho^\mu)^2 + \frac{\alpha}{2\beta}\sum_\rho \log\lambda_\rho + \tfrac{1}{2}\alpha\beta\sum_{\rho\sigma} r_{\rho\sigma}q_{\rho\sigma} - \frac{1}{\beta}\Big\langle\!\Big\langle \log \mathrm{Tr}_S \exp\Big(\beta\sum_{\mu\rho} m_\rho^\mu \xi^\mu S^\rho + \tfrac{1}{2}\alpha\beta^2\sum_{\rho\sigma} r_{\rho\sigma}S^\rho S^\sigma\Big)\Big\rangle\!\Big\rangle. \quad (10.51)$$

The factor of N in the exponent allows us to use the saddle-point method again, minimizing this time with respect to the q's and r's as well as the m's. Thus we obtain the free energy per unit

$$F/N = -\frac{1}{\beta N}\langle\!\langle \log Z\rangle\!\rangle = -\frac{1}{\beta N}\lim_{n\to 0}\frac{1}{n}\big(\langle\!\langle Z^n\rangle\!\rangle - 1\big) \quad (10.52)$$

$$= -\frac{1}{\beta N}\lim_{n\to 0}\frac{1}{n}\log\langle\!\langle Z^n\rangle\!\rangle$$

$$= \frac{\alpha}{2} + \lim_{n\to 0}\frac{1}{n}\min f\{m,q,r\}. \quad (10.53)$$

In replacing $\langle\!\langle Z^n\rangle\!\rangle - 1$ by $\log\langle\!\langle Z^n\rangle\!\rangle$ we just assumed that $\langle\!\langle Z^n\rangle\!\rangle$ goes to 1 as $n \to 0$, as it must; that is why we didn't bother to keep all the prefactors earlier.

The location of the saddle point is determined by the equations

$$\frac{\partial f}{\partial m_\rho^\mu} = 0 \quad (10.54)$$

$$\frac{\partial f}{\partial q_{\rho\sigma}} = 0 \quad (10.55)$$

$$\frac{\partial f}{\partial r_{\rho\sigma}} = 0. \quad (10.56)$$

As in the simpler $\alpha = 0$ case, these equations lead to interpretations of the **order parameters** m_ρ^μ, $q_{\rho\sigma}$, and $r_{\rho\sigma}$ at the saddle point:

$$m_\rho^\mu = \frac{1}{N}\sum_i \xi_i^\mu \langle S_i^\rho\rangle \quad (10.57)$$

$$q_{\rho\sigma} = \Big\langle\!\Big\langle \frac{1}{N}\sum_i \langle S_i^\rho\rangle\langle S_i^\sigma\rangle\Big\rangle\!\Big\rangle \quad (10.58)$$

$$r_{\rho\sigma} = \frac{1}{\alpha}\sum_{\mu>s}\langle\!\langle m_\rho^\mu m_\sigma^\mu\rangle\!\rangle. \quad (10.59)$$

We omit the detailed derivations of these results, which require the inclusion of external field terms h^μ as in the $\alpha = 0$ case. Equation (10.59), which comes from $\partial f/\partial q_{\rho\sigma} = 0$, also involves rewriting the $\log\lambda_\rho$ term as a Gaussian integral. Note that (10.57) is just like the $\alpha = 0$ result (10.21) apart from the presence of the replica index ρ.

To proceed further we have to make an *ansatz* without *a priori* justification: that of **replica symmetry**. This means that we assume that the saddle-point values of the order parameters do not depend on their replica indices:

$$m_\rho^\mu = m^\mu \tag{10.60}$$

$$q_{\rho\sigma} = q \tag{10.61}$$

$$r_{\rho\sigma} = r. \tag{10.62}$$

The validity of this assumption can be tested afterwards, and one finds that it is exactly true except at very low temperatures, and that even there it is a good approximation.

With this simplification the meaning of the order parameters (10.57)–(10.59) is evident, and consistent with the heuristic treatment in Chapter 2: m^μ is (as before) the overlap between the network configuration and the μth pattern, q is the mean squared magnetization, and αr is the mean squared value of the overlap with the uncondensed patterns ($\mu > s$). Each m^μ for $\mu > s$ is of order $1/\sqrt{N}$, but r remains finite as $N \to \infty$ because there are of order N terms in the sum (10.59).

Using the replica symmetric *ansatz* the expression (10.51) for $f(m, q, r)$ simplifies to

$$f(\mathbf{m}, q, r) = \tfrac{1}{2}n\mathbf{m}^2 + \frac{\alpha}{2\beta}\sum_\rho \log\lambda_\rho + \tfrac{1}{2}n(n-1)\alpha\beta rq + \tfrac{1}{2}n\alpha\beta r$$

$$-\frac{1}{\beta}\left\langle\!\!\left\langle \log \mathrm{Tr}_S \exp\left[\beta\mathbf{m}\cdot\boldsymbol{\xi}\sum_\rho S^\rho + \tfrac{1}{2}\alpha\beta^2 r\left(\sum_\rho S^\rho\right)^2\right]\right\rangle\!\!\right\rangle \tag{10.63}$$

where the last term on the first line is to cancel the diagonal part of the $(\sum_\rho S^\rho)^2$ term. We still have to evaluate the sum of the $\log\lambda_\rho$'s and compute the average of the Tr over the S^1–S^n but, thanks to the replica symmetry, these can now be done without too much trouble.

Let us first deal with the eigenvalue sum. The matrix $\Lambda_{\rho\sigma}$ now has the simple form

$$\Lambda_{\rho\sigma} = \begin{cases} 1-\beta & \text{if } \rho = \sigma; \\ -\beta q & \text{otherwise.} \end{cases} \tag{10.64}$$

It is an elementary exercise to show that such a matrix has eigenvalues

$$\lambda_1 = 1 - \beta - (n-1)\beta q \tag{10.65}$$

with multiplicity 1 and

$$\lambda_2 = 1 - \beta(1-q) \tag{10.66}$$

with multiplicity (i.e., number of eigenvectors with this eigenvalue) $n-1$. Thus the sum over the logs of the eigenvalues becomes

$$\frac{1}{n}\sum_\rho \log\lambda_\rho = \frac{1}{n}\{\log[1-\beta-(n-1)\beta q] + (n-1)\log[1-\beta(1-q)]\}$$

$$\xrightarrow{n\to 0} \log[1-\beta(1-q)] - \frac{\beta q}{1-\beta(1-q)}. \tag{10.67}$$

To evaluate the Tr over the S's, we again use the Gaussian integral trick:

$$\exp\Big[\tfrac{1}{2}\alpha\beta^2 r\Big(\sum_\rho S^\rho\Big)^2\Big] = \int \frac{dz}{\sqrt{2\pi}} \exp\Big(-\tfrac{1}{2}z^2 + \beta\sqrt{\alpha r}z \sum_\rho S^\rho\Big) \tag{10.68}$$

giving for the trace $X \equiv \mathrm{Tr}\exp[\cdots]$ in (10.63):

$$\begin{aligned} X &= \mathrm{Tr}_S \int \frac{dz}{\sqrt{2\pi}} \exp\Big(-\tfrac{1}{2}z^2 + \beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})\sum_\rho S^\rho\Big) \\ &= \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}\Big(2\cosh\beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})\Big)^n \\ &= \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}\exp\Big(n\log[2\cosh\beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})]\Big)\,. \end{aligned} \tag{10.69}$$

We actually want $1/n$ times the average of the log of this, in the $n \to 0$ limit. Expanding for small n gives

$$\begin{aligned} \frac{1}{n}\langle\!\langle \log X\rangle\!\rangle &= \frac{1}{n}\Big\langle\!\Big\langle \log\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}\Big(1 + n\log[2\cosh\beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})] + \cdots\Big)\Big\rangle\!\Big\rangle \\ &= \frac{1}{n}\Big\langle\!\Big\langle n\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}\log[2\cosh\beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})] + \cdots\Big\rangle\!\Big\rangle \\ &\xrightarrow{n\to 0} \langle\!\langle \log[2\cosh\beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})]\rangle\!\rangle \end{aligned} \tag{10.70}$$

where now we take $\langle\!\langle\cdots\rangle\!\rangle$ to mean both the average over the condensed patterns $\mu \le s$ and the Gaussian average over z. Physically this means averaging over *all* the patterns, since the Gaussian random field z came from representing the effects of the uncondensed patterns $\mu > s$.

All we have left is to collect the terms from (10.53), (10.63), (10.67), and (10.70) to give the average free energy per site in the form

$$\begin{aligned} F/N &= \tfrac{1}{2}\alpha + \tfrac{1}{2}\mathbf{m}^2 + \frac{\alpha}{2\beta}\Big(\log[1-\beta(1-q)] - \frac{\beta q}{1-\beta(1-q)}\Big) \\ &\quad + \tfrac{1}{2}\alpha\beta r(1-q) - \frac{1}{\beta}\langle\!\langle \log[2\cosh\beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})]\rangle\!\rangle\,. \end{aligned} \tag{10.71}$$

The saddle-point equations (10.54)–(10.56) are equivalent to setting the derivatives of F/N to zero, giving

$$m^\mu = \langle\!\langle \xi^\mu \tanh\beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})\rangle\!\rangle \tag{10.72}$$

$$q = \langle\!\langle \tanh^2\beta(\sqrt{\alpha r}z + \mathbf{m}\cdot\boldsymbol{\xi})\rangle\!\rangle \tag{10.73}$$

$$r = \frac{q}{[1-\beta(1-q)]^2}\,. \tag{10.74}$$

Only the second of these, which comes from $\partial F/\partial r = 0$, is a little tricky, needing the identity

$$\int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} z f(z) = \int \frac{dz}{\sqrt{2\pi}} e^{-z^2/2} f'(z) \tag{10.75}$$

for any bounded function $f(z)$.

Equation (10.72) is just like (10.22) for the $\alpha = 0$ case, except for the addition of the effective Gaussian random field term, which represents the crosstalk from the uncondensed patterns. For $\alpha = 0$ it reduces directly to (10.22). Equation (10.73) is the obvious equation for the mean square magnetization. Equation (10.74) gives the (nontrivial) relation between q and the mean square value of the random field, and is identical to (2.67).

For **memory states**, i.e., m-vectors of the form $(m, 0, 0, \ldots)$, the saddle-point equations (10.72) and (10.73) become simply

$$m = \langle\langle \tanh \beta(\sqrt{\alpha r} z + m) \rangle\rangle_z \tag{10.76}$$

$$q = \langle\langle \tanh^2 \beta(\sqrt{\alpha r} z + m) \rangle\rangle_z \tag{10.77}$$

where the averaging is solely over the Gaussian random field. These are are identical to (2.65) and (2.68) that we found in the heuristic theory of Section 2.5. Their solution, and the consequent phase diagram of the model in $\alpha - T$ space, can be studied as we sketched there. Spurious states, such as the symmetric combinations (10.26), can also be analyzed at finite α using the full equations (10.72)–(10.74).

There are several subtle points in this replica method calculation:

- We started by calculating $\langle\langle Z^n \rangle\rangle$ for *integer* n but eventually interpreted n as a real number and took the $n \to 0$ limit. This is not the only possible continuation from the integers to the reals; we might for example have added a function like $\sin \pi n / n$.
- We treated the order of limits and averages in a cavalier fashion, and in particular reversed the order of $n \to 0$ and $N \to \infty$.
- We made the replica symmetry approximation (10.60)–(10.62) which was really only based on intuition.

Experience has shown that the replica method usually does work, but there are few rigorous mathematical results. It can be shown for the Sherrington-Kirkpatrick spin glass model, and probably for this one too, that the reversal of limits is justified, and that the replica symmetry assumption is correct *for integer* n [van Hemmen and Palmer, 1979]. But for some problems, including the spin glass, the method sometimes gives the wrong answer. This can be blamed on the integer-to-real continuation, and can be corrected by **replica symmetry breaking**, in which the replica symmetry assumption is replaced by a more complicated assumption. Then the natural continuation seems to give the right answer.

For the present problem Amit et al. showed that the replica symmetric approximation is valid except at very low temperatures where there is replica symmetry breaking. This seems to lead only to very small corrections in the results. However,

the predicted change in the capacity—α_c becomes 0.144 instead of 0.138—can be detected in numerical simulations [Crisanti et al., 1986].

10.2 Gardner Theory of the Connections

The second classic statistical mechanical *tour de force* in neural networks is the computation by Gardner [1987, 1988] of the capacity of a simple perceptron. The calculation applies in the same form to a Hopfield-like recurrent network for auto-associative memory if the connections are allowed to be asymmetric.

This theory is very general; it is not specific to any particular algorithm for determining the connections. On the other hand, it does not provide us with a specific set of connections even when it has told us that such a set exists. As in Section 6.5, the basic idea is to consider the fraction of **weight space** that implements a particular input-output function; recall that weight space is the space of all possible connection weights $\mathbf{w} = \{w_{ij}\}$.

In Section 6.5 we used relatively simple methods to calculate weight space volumes. The present approach is more complicated, though often more powerful. We use many of the techniques introduced in the previous section, including replicas, auxiliary variables, and the saddle-point method.

We consider a simple perceptron with N binary inputs $\xi_j = \pm 1$ and M binary threshold units that compute the outputs

$$O_i = \mathrm{sgn}\Big(N^{-1/2}\sum_j w_{ij}\xi_j\Big)\,. \tag{10.78}$$

The $N^{-1/2}$ factor will be discussed shortly. Given a desired set of associations $\xi_j^\mu \to \zeta_i^\mu$ for $\mu = 1, 2, \ldots, p$, we want to know in what fraction of weight space the equations

$$\zeta_i^\mu = \mathrm{sgn}\Big(N^{-1/2}\sum_j w_{ij}\xi_j^\mu\Big) \tag{10.79}$$

are satisfied (for all i and μ). Or equivalently, in what fraction of this space are the inequalities

$$\zeta_i^\mu N^{-1/2}\sum_j w_{ij}\xi_j^\mu > 0 \tag{10.80}$$

true?

It is also interesting to ask the corresponding question if the condition (10.80) is strengthened so there is a **margin size** $\kappa > 0$ as in (5.20):

$$\zeta_i^\mu N^{-1/2}\sum_j w_{ij}\xi_j^\mu > \kappa\,. \tag{10.81}$$

A nonzero κ guarantees correction of small errors in the input pattern.