Phase Space Techniques in Neural Network Models

Hon Wah Yau

Submitted for the degree of
Doctor of Philosophy

Department of Physics
University of Edinburgh

June 1992

Declaration

I declare that I am the sole author of this thesis, and that all the research detailed is my own except where otherwise stated. Part of the work has been published in the journal:

H W Yau and D J Wallace. Enlarging the attractor basins of neural networks with noisy external fields. *Journal of Physics A: Maths and General*, 24:5639–5650, 1991.

H W Yau

June 1992

I wish to dedicate this thesis to

my mother Yuk Chun Lee, and my father Tin Kuen

in gratitude for all they have done

Acknowledgements

The support I have received during my time in Edinburgh has never been less than generous, and I hope this page goes some way towards expressing my gratitude. Firstly, I must thank my supervisor David Wallace, who has supported me in everything I have pursued, and whose inspiration initiated this work. For invaluable advice and patience I wish to thank Michael Wong, who offered me the opportunity to work with him in Oxford. I, for one, believe this thesis will be half the size and/or twice as tedious without Michael's kindness.

There are many people in and outside Edinburgh who deserve a mention here, but I must offer special thanks to David Sherrington for arranging my visits to Oxford, and to my fellow office inmate Martin Simmen for the many profound discussions on LaTeX, politics, and each other's theses. I am also grateful for contributing comments from Alan Dunmur, Roger Blackburn, and Martin Evans.

In addition I wish to thank Fiona Pardey for her infectious *joie de vivre*, Caroline Howell for the myriad manifestations of her cake recipe, and Susanne Riehemann for her friendship. I must also thank Ken Hawick, Charles Beveridge, Stephen Booth, Simon Tett, and Claire Jones for those occasional and enjoyable forays to the bright city lights.

I am, of course, also indebted to my parents and my sister Cynthia, without whom many more things than this thesis will have been impossible.

Finally, I wish to thank *mon amie*, Carolyn Maguire, simply for all the quieter moments.

Financial support from the Science and Engineering Research Council, and a CASE award from the MEiKO company are gratefully acknowledged.

Abstract

We present here two calculations based on the *phase-space of interactions* treatment of neural network models.

As a way of introduction we begin by discussing the type of neural network models we wish to study, and the analytical techniques available to us from the branch of disordered systems in statistical mechanics. We then detail a neural network which models a *content addressable memory*, and sketch the mathematical methods we shall use. The model is a mathematical realisation of a neural network with its synaptic efficacies optimised in its phase space of interactions through some *training function*.

The first model looks at how the basin of attraction of such a content addressable memory can be enlarged by the use of noisy external fields. These fields are used separately during the training and retrieval phases, and their influences compared. Expressed in terms of the number of memory patterns which the network's dynamics can retrieve with a microscopic initial overlap, we shall show that content addressability can be substantially improved.

The second calculation concerns the use of *dual distribution functions* for two networks with different constraints on their synapses, but required to store the same set of memory patterns. This technique allows us to see how the two networks accommodate the demands imposed on them, and whether they arrive at radically different solutions. The problem we choose is aimed at, and eventually succeeds in, resolving a paradox in the sign-constrained model.

Contents

List of Figures

Chapter 1

Neural Network Models and the Phase Space of Interactions

In principle, neural network models are nothing less than an attempt to capture the salient properties of biological neural tissues. Whether this programme is anywhere near to accommodating the enormous body of neurological information is a debate we shall not enter. But what we can certainly say is that the introduction of neural networks has spawned a variety of fields across the disparate disciplines of computer science, electronic engineering, theoretical physics, cognitive science, and perhaps even in neurobiology itself. However, it is the theoretical physics viewpoint which will be the main concern of this work, and in particular a demonstration of the power of the mathematical tools developed in the field of statistical mechanics.

The object of this introductory chapter is to bring together the relevant features of neural network models and statistical mechanics. We shall avoid a lengthy treatise on either fields by focusing our attention on a neural network based model of a *content addressable memory*, and its analysis by the use of *quenched disorder* statistical mechanics. The characteristics of this model and the calculation techniques to be described will both be of relevance in the two chapters which follow.

1.1 Neural Network Models and Statistical Mechanics

Much has been written in the literature across the entire spectrum of interests which falls under the heading of 'Neural Networks', and it is not simply a matter of expedience here to request the reader to refer to, for example, references [Ami89] and [HKP91] for suitably broad treatments. Instead, in this section we shall discuss what we mean by a neural network, with a deliberate bias to the areas we shall require for the subsequent chapters.

The model of the neuron we shall use is that introduced by McCulloch and Pitts in 1943. This device models the two possible states of neural activity with a simple update rule dependent on the state of the interacting neighbours and the *synaptic efficacy* between them. The values of these synapses are determined during the *training* phase, by presenting the network with some set of training data based on the problem we wish the network to solve. In the case of a memory model, it will be the binary images of patterns we wish the network to store.

Neural networks are categorised by the architecture which seems best suited to the tasks we require to solve. The one we shall study consists of a single layer of many inter-connected neurons, with each site acting as a bit of information for a content addressable memory. Such an architecture is believed to exist in the hippocampus region of the brain [TR90]. Fortunately, this model has strong analogies with magnetic spin systems, and in certain cases is amenable to analysis by the methods of statistical mechanics. Indeed, so much of the analytical work done on neural networks has been based on statistical mechanics that we have freely borrowed many of its imageries and terms. It is therefore necessary to devote the next subsection to discussing the machinery of the techniques we shall encounter.

1.1.1 Quenched Disorder Systems

Statistical mechanics is the study of systems with many components all behaving in a simple way. The quantities it calculates reflect the nature of such systems by demanding that they be statistically relevant to be of any importance. That is, the *typical* value of an observable is equal to its *average* with a vanishing degree of uncertainty. This is another way of saying any relevant quantity we calculate must be invariant from sample to sample, and is hence independent of microscopically fine details.

To calculate the observables of a statistical mechanical system the usual starting point is to write it in terms of a *configurational sum*, suitably weighted by a Boltzmann term. This is of course, the celebrated central axiom of statistical mechanics which equates the equilibrium values of a system with a configurational average over the important region of phase space. If we now denote the *thermal average* of an observable A by angled brackets we can see this configurational sum by schematically writing down the weighted sum over the configurations S as

$$\langle A\rangle_\beta \equiv \frac{1}{Z}\sum_S A(S)\exp\left(-\beta\mathcal{H}(S)\right), \tag{1.1}$$

where the normalising constant Z is the *partition function*, β^{-1} is the temperature of the system, and $\mathcal{H}(S)$ is the *Hamiltonian* which contains the details of the model. Upon taking the low temperature limit of ($\beta \to \infty$), the sum will be dominated by the configuration which yields the lowest value for the Hamiltonian. This procedure is analogous to slowly cooling down the system and hence is often referred to as *annealed optimisation.*

There are a variety of ways to calculate the thermal averages of equation (1.1), depending on the class of model under scrutiny. A common first steps is to insert any physics we know into the problem by identifying the relevant *overlap measures*, so called because they are a succinct statement of the system's configuration with a chosen state. Given a sensible set of overlap measures, we can replace the configurational sum with an integration over the overlap measures by use of the 'extraction' technique of equation (A.7). The resulting integrals are not likely to be any easier to exactly calculate than the original configurational sum, but the problem is now amenable to approximation by the saddle-point method [Cop65, Arf85]. This approximation becomes exact in the *thermodynamic limit*, that is, when the co-ordination number of all the sites is large. Following reference [Ste89], we shall elevate the overlap measure at the saddle-point by calling it an *order parameter* of the system. We do this because the order parameter is a physically relevant, thermal-averaged observable which characterises the system, and not just a quantity we introduce to ease the mathematics. In passing we should mention this nomenclature is not universal, and the literature has referred to our overlap measure as their order parameter.

The Replica Technique

The title of this subsection refers to the particular branch of statistical mechanics we shall be interested in. This is the study of models where there is some disorder frozen, *i.e.*, quenched, into the system. The origins of this concept lie in magnetic spin-glass systems, where it is believed the interactions between sites have been frozen into a random configuration [EA75, KS75]. References [Pal89, Ste89] provide good introductions to this subject, while the book [MPV87] gives the most authoritative treatment to date.

In the spirit of statistical mechanics we want to take the average example of disorder in our system and feel justified in believing it is also a typical example. This requires us to do the averaging on the quenched disorder of the system, but since this disorder is supposed to be fixed throughout all time the averaging must be over a physically observable quantity. Another way of putting this is that we wish to quenched average over an extensive object whose fluctuations from system to system will vanish in the thermodynamic limit. Unfortunately, this averaging is only simple for the partition function, but which has exponentially large fluctuations. To properly do the quenched averaging we need to use the *replica technique* first introduced in reference [EA75].

The archetypal quantity we shall quenched average over is the logarithm of the partition function. This extensive object is related to the physically observable free energy by a factor $-\beta$ which we shall neglect. The crucial first step is to write the logarithm in terms of $a = 1 \ldots n$ replicas of the system via the identity

$$\langle\!\langle \ln Z \rangle\!\rangle = \lim_{n \to 0} \frac{1}{n} \ln \left\langle\!\!\left\langle \prod_{a=1}^{n} Z^a \right\rangle\!\!\right\rangle \tag{1.2}$$

with the double-angled brackets denoting the quenched average. Upon performing the quenched average we find it is necessary to introduce an overlap measure between the replicas. This overlap is written as q_{ab} and it encodes how the phase-space of the model is being divided up by the disorder. The overlap q_{ab} is again evaluated at the saddle-point in the thermodynamic limit, whilst taking the *replica-symmetric ansatz* of $q_{ab} = q \ \forall a \neq b$. We can identify this saddle-point of q with the Edwards-Anderson order parameter q_{EA}, a quantity originally proposed to indicate broken ergodicity in a system. The appendix chapter B discusses this object in more detail, albeit in another context.

Given equation (1.2), which appears to display full symmetry under permutations of the replica index, we may naïvely expect the replica-symmetric ansatz to be the only sensible move to make. However in highly disordered models such as spin-glasses it was quickly realised that replica-symmetry gave incorrect solutions, with physically questionable results such as negative entropies and re-entrant behaviour in the phase-diagram. Hence for many quenched disorder calculations involving the replica technique a query often made by the cognoscenti is whether the replica-symmetric ansatz is a valid assumption to make, or whether we should use something with a richer structure. There are at present three means of answering this question and their merits and pitfalls are outlined here.

Historically the first, the method introduced by de-Almeida and Thouless [dAT78] was in response to problems with the low temperature limit of the Sherrington-Kirkpatrick spin-glass model [KS75, KS78]. It acknowledges the use of the method of steepest descent in performing the integrations over the replica indexed overlap measures and examines whether the replica-symmetric ansatz is actually a stable fixed-point. That is, whether the determinant of the Hessian at that point is positive definite. The replicas provide the Hessian with a non-trivial structure, from which we can reduce the question of stability to any arbitrary fluctuation to that of just three specific modes. The calculation of the elements of the Hessian matrix can be somewhat tedious, but this is the method used in calculations that follow.

Unfortunately, having a stable replica-symmetric ansatz does not necessarily mean it will be *the* correct solution, for there may be another non replica-symmetric ansatz with a lower free energy[1] which is the true saddle-point solution. The obvious way to test for this eventuality is to insert an alternative ansatz and compare it against the replica-symmetric case. Finding an appropriate alternative is not as hopeless as we may expect, for there is in the Parisi replica-symmetry-breaking scheme[2] a proven and possibly unique [Lau91, Fra91] route to generating the *ultrametric structure* expected. Hence we can in principle begin by testing with the so-called one-step replica-symmetry breaking ansatz, and progress with further steps towards ever more elaborate schemes.

Lastly we can in certain cases calculate the (replica-symmetric) entropy of the model. Finding the line of zero-entropy below which the replica-symmetric ansatz

[1] The word 'energy' can be used in a metaphorical sense to denote some cost-function.

[2] Reference [MPV87] contains a section devoted to the Parisi ansatz.

should be rejected is the most straightforward of the methods discussed here, and can be extended to consider the validity of the replica-symmetry-broken ansätze as well. Unfortunately this seems only possible in models where the space of the annealing dynamics is discrete, such as in spin-glass models, and cannot be extended to continuous models.

1.1.2 Neural Network Models as Quenched Disorder Systems

We shall now pull the threads of the previous two subsections together and discuss how the statistical mechanics of quenched disorder systems have anything to say about neural network models. Once again the literature contains good coverage of the subject, in for example references [Ami89, She90].

The initial breakthrough came from the writing down of a Lyapunov function for the neural-dynamics by Hopfield [Hop82], which made the problem amenable to analysis by equilibrium statistical mechanics. This model was quickly elaborated upon and culminated in the full replica treatment by references [AGS85, AGS86]. This model maps the neural activity of the system onto spin-sites, with a Hamiltonian based on Hopfield's Lyapunov function. The interactions amongst the sites are given by a prescribed matrix which is motivated by the proposed mechanism of Hebbian learning in biological synapses[Heb49]. By being able to store and retrieve a given set of memory patterns, the network is indeed an example of a content addressable memory,

The second approach was introduced by Gardner and is distinctively non-biological in motivation. Instead it asks what are the properties of a network optimised to a certain task, namely the task of storing a set of memory patterns. The choice of the word 'optimise' is deliberate, for the calculation is based upon an annealing process in the synaptic efficacy. That is, in lieu of the neurons' spin configuration we have the values of the synaptic connections as the dynamical variable, which in the limit of a zero anneal temperature limit settles on the optimal solution. Hence instead of a sum over the spin model's configurations by its partition function, we have here an exploration over the *volume of the phase-space of interactions.* For the work which follows, we shall be interested in the properties of a network possessing such an anneal-optimised set of interactions. To actually uncover these optimal connections Gardner proposed a perceptron-like training algorithm first introduced by Rosenblatt. She then proceeded to prove that provided a set of optimal connections actually exists, the algorithm will converge in a finite number of steps [Gar88].

How the synaptic connections are optimised is dependent on the physical features of the network, and to elaborate on the Gardner ethos, it is now necessary to turn our attention to the physical definition of a neural network model.

1.2 A Neural Network Model of Associative Memory

The aim here is to write down mathematically the physical realisation of the neural network model, and in particular an auto-associative memory trained to learn a set of memory patterns. Variations on this model will be studied further in following chapters.

The network consists of N time-dependent binary perceptrons, the state of which are represented by the vector $\boldsymbol{S}(t) \equiv \{S_i(t)\}, i = 1 \ldots N$ where each site can be either $S_i(t) = \pm 1$. Each neuron interacts with C others via the connection matrix $\{\boldsymbol{J}^i\} \equiv \{J^{ij}\}, i = 1 \ldots N, j = 1 \ldots C$ $(j \neq i)$ and obeys the zero-temperature parallel update

$$S_i(t+1) = \mathrm{sign}[h_i(t)] \tag{1.3}$$

dependent on the local field at the previous time-step. The added complexity of an external field will be discussed in full later, but for now we shall assume the local field is given simply by the scalar product

$$h_i(t) = \sum_{j \neq i}^{C} \frac{J^{ij}}{|\boldsymbol{J}^i|} S_j(t) \tag{1.4}$$

where the normaliser is defined through $|\boldsymbol{J}^i|^2 = \sum_{j \neq i}^{C} J_{ij}^2$. The inclusion of this normaliser is a reflection of the redundancy in the weights' magnitude when using equation (1.3) for the update. If the $\{J_{ii}\}$ diagonal elements are sufficiently large and positive, we can see that *any* given state will be stable to the update rule. This obviously distorts the storage capacity of the model with spurious states possessing attractor basins of zero sizes, so we shall choose to explicitly remove them from the model.

The task given to the network is the storage of P uncorrelated patterns $\boldsymbol{\xi}^\mu \equiv \{\xi_i^\mu\}, i = 1 \ldots N, \mu = 1 \ldots P$, and this is successfully done if they form fixed-points of the above dynamics, that is if

$$\xi_i^\mu = \mathrm{sign}[\boldsymbol{J}^i \cdot \boldsymbol{\xi}^\mu / |\boldsymbol{J}^i|] \qquad \forall \mu = 1 \ldots P \tag{1.5}$$

where we have taken the liberty of defining the scalar product as the sum over the input nodes $j = 1 \ldots C, (j \neq i)$ entering the i^{th} neuron. The above equation can be alternatively expressed by defining an 'alignment field'

$$\Lambda_i^\mu \equiv \xi_i^\mu (\boldsymbol{J}^i \cdot \boldsymbol{\xi}^\mu) / |\boldsymbol{J}^i| \tag{1.6}$$

for each pattern $\boldsymbol{\xi}^{\mu}$, and requiring $\Lambda_i^{\mu} > 0 \; \forall i$ for it to be memorised.

In an associative memory it is obviously imperative to monitor the dynamics of the N neurons as they evolve in time. A natural measure to use for this is the overlap with a chosen pattern, say $\boldsymbol{\xi}^{\mu=1}$,

$$m^1(t) \equiv \frac{1}{N} \sum_{i=1}^{N} \xi_i^1 S_i(t) \tag{1.7}$$

at time-step t. We are typically interested in starting the network at a state close to one of the nominated patterns, and whether the dynamics will draw the neurons towards that pattern. In this case the convergence of the quantity (1.7) to unity will indicate the successful retrieval of that pattern.

This completes the physical definition of an associative memory. For the calculations to proceed, we must specify the nature of the phase-space of connections and detail the importance of the alignment field. But first it is useful to build on the above and make a brief but necessary digression into the case of learning in a single perceptron.

1.2.1 The Single Perceptron

Here we are considering just a single neuron[3] connected to C inputs by a synaptic vector $\boldsymbol{J} \equiv \{J^j\}$. The problem then is not so much memorising a set of patterns as learning a given rule as defined by a set of binary input-output mappings $(\boldsymbol{\xi}^{\mu}, \mathrm{T}^{\mu})$, where $\mu = 1 \dots P$ now enumerates over the so-called training examples. With a binary perceptron we can make use of the alignment field (1.6) and succinctly say

[3] For a reference on the properties of single perceptrons from a mainly geometric perspective please refer to the standard text by Minsky and Papert [MP88].

the perceptron is successfully trained when

$$\Lambda^\mu \equiv \mathrm{T}^\mu(\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu)/|\boldsymbol{J}| > 0 \tag{1.8}$$

for all the P input-output relations.

The computational ability of a single binary perceptron is not unlimited and we can very easily think of input-output rules which cannot be learnt, such as the celebrated 'Exclusive-Or problem'[MP88]. These 'unlearnable' problems occur whenever the inputs cannot be linearly separated into two regions corresponding to the outputs $\mathrm{T}^\mu = +1$ and -1. However this is hardly a reasonable criterion in real world problems and the need for its circumvention has spawned networks with richer architectures, but with less analytically tractable behaviour.

Returning to the case of an associative memory, we can consider each of the N sites as being trained with P input-output relations. For the storage of uncorrelated patterns these relations are simply examples of the random binary mapping problem. Hence for a pattern $\boldsymbol{\xi}^\mu$ the required output T^μ is given by a specific bit ξ_i^μ and it is successfully memorised when this rule is observed across all the $i = 1 \dots N$ sites.

For all but the most trivial case the random binary mapping problem will become more likely to be unlearnable as the number of memory patterns increases, and the point at which this occurs defines the storage-capacity of the network. This important parameter gives a readily accessible measure of the performance of a network and its analytic calculation is one of the more obvious successes of the statistical mechanics approach.

1.3 The Phase Space of Interactions

The task here is to formulate the process of learning in a network of perceptrons as an annealed search through the phase-space of connections. This process of annealing can be expressed by writing down a phase-space volume associated with each site $S_i(t)$ with the appropriate Boltzmann weighting. This weighting is a reflection of the training task and is only dependent upon the alignment field, that is some 'training-function' $g(\Lambda_i^\mu)$ which rewards correct learning [WS90a] for each of the $\mu = 1 \ldots P$ patterns. Hence for a network performing as an associative memory the volume of phase-space for the interactions into the i^{th} site looks like

$$\mathrm{V}^i = \int \mathrm{D}(\boldsymbol{J}^i) \exp \left(\beta \sum_\mu^P g(\Lambda_i^\mu) \right) \tag{1.9}$$

where β is the inverse annealing temperature, such that in the zero temperature ($\beta \rightarrow \infty$) limit the optimal synaptic configuration will be found. The volume element $\mathrm{D}(\boldsymbol{J}^i)$ is over the connectivity-C dimensions with some prior distribution and constraint. The above equation (1.9) is the volume for a single site in the network; for the whole network the total volume is simply $\prod_i^N \mathrm{V}^i$ as the dimensionality of the problem is increased. For the case of uncorrelated input-output mappings considered here, we can remove the superfluous site index i and along with it the distinction between an associative-memory network and a single perceptron. As far as each neuron is concerned, memorising uncorrelated patterns is merely the consequence of attempting to learn a set of random binary-mapping problems.

The patterns we require the network to memorise are random, uncorrelated binary images. In the spirit of statistical mechanics we take average patterns as being the *typical* patterns, with any observables having vanishing fluctuations between differing samples. These patterns are our source of disorder, remaining fixed over the time-scale of the annealing process. As outlined in §1.1.1 any averaging over them must hence be over extensive quantities, and this means we do not average equation (1.9) but instead quantities such as the logarithm of the volume. Replicating the connections, which are here the annealed variables, this means we wish to perform a quenched average of the archetypal form

$$\langle\langle \mathrm{V}^n \rangle\rangle = \left\langle\left\langle \int \prod_a^n \left\{ \mathrm{D}(\boldsymbol{J}_a) \exp\left(\beta \sum_\mu^P g(\Lambda_a^\mu)\right)\right\}\right\rangle\right\rangle \tag{1.10}$$

with the alignment field for weight $\boldsymbol{J}_a$ having also picked up a replica index.

Upon doing the quenched average we introduce interactions amongst the replicas and this is dealt with by introducing the overlap measure between the replicas

$$q_{ab} = \frac{1}{C}(\boldsymbol{J}_a \cdot \boldsymbol{J}_b), \qquad \forall a, b \, (a < b), \tag{1.11}$$

The restriction in the indices give this quantity $n(n-1)/2$ unique values, but having said that, the so-called replica-symmetric ansatz is normally chosen as soon as possible to facilitate any further analytical progress. This ansatz sets the above elements to the same value, *i.e.*, $q_{ab} = q, \forall a \neq b$. Finally, this is a mean-field calculation, and in the thermodynamic limit the integrations with respect to q_{ab} can be replaced by the saddle-point value through use of the method of steepest descent [Cop65, Arf85].

Thus far we have avoided the issue of the nature of the phase-space of interactions. But in order to perform the calculation in equation (1.10) we must specify the nature of the volume element $\mathrm{D}(\boldsymbol{J}^i)$, and with it any prior distributions in the weights. In the work that follows, the synapses will take on continuous values with the phase-space restricted to lie on a unit hypersphere by enforcing the spherical constraint

$$|\boldsymbol{J}|^2 = C \tag{1.12}$$

over the connectivity, hence setting the diagonal elements of the correlation matrix (1.11) to unity.

In passing we should mention the existence of another phase-space in common use. This is the case where the weights are restricted to $J^j = \pm 1$ in lieu of constraint (1.12), giving an annealing landscape similar to that seen in the dynamics of spin-models [GD88] with the integration in equation (1.10) replaced by a configurational sum. Furthermore there have been extensions beyond the binary weights case, where the weights have multiple discrete states [BDvM91].

1.4 Summary

We have given in this chapter brief introductions to neural network models and statistical mechanics, with the aim of allowing us to use them in the calculations that follow. This then led to the formulation of learning in a neural network model as an annealed optimisation process. By considering annealed optimised synapses, learning can be treated as a statistical mechanical problem, with the partition function replaced by a volume of phase-space. For proper treatment the calculation requires a quenched average over the uncorrelated input-output relationships, which necessitates the use of the replica method. An important simplification can be made in the notation by recognising the memorising of patterns in an associative memory can be viewed from the perspective of a single perceptron. That is, an associative memory for uncorrelated patterns is merely the random binary mapping problem, a problem which is eventually unlearnable in the limit of a large number of patterns.

The use of the phase-space of interactions, the alignment field, and the simplification to a single perceptron will be key concepts in the following chapters.

Chapter 2

Enlarging the Attractor Basin in Neural Network Models by External Fields

This chapter is an elaboration of previously published results [YW91], with the majority of the additional material given to detailing the calculations involved. We shall build on sections made in the last chapter by presenting here an explicit example of an associative memory neural network model, allowing insights into the analytical methods albeit with an unavoidable loss of generality. As a way of introducing this work, we shall begin by giving an idea of what we wish to achieve, followed by a discussion on the quantities to be calculated. There will then be sections on the calculation itself, results, and a concluding discussion.

2.1 Motivations for an External Field

A key attraction of statistical mechanical models of neural networks is their ability to function as associative memories. This is a two stage process with the network first *trained* to store a set of memory patterns, followed by a *retrieval* stage defined by the neurons' update dynamics. However, retrieval of a stored pattern can only occur if the system is initiated sufficiently close to it. That is, if the initial state of the networks is inside the *basin of attraction* of the nominated memory pattern's *attractor*. Expressed in this way, we can see that content addressability is merely the consequence of a memory state having a non-zero basin of attraction.

The principal motivation of this chapter is to examine how the basin of attraction can be enlarged by the use of *external fields* which are noisy representations of the memory patterns stored. Independent work has shown the beneficial effects of applying noisy external fields throughout retrieval [EBKS90, RSW91] but as stated above, a network is defined in two stages and their rôle during the training phase should also be explored. Moreover, both simulation [GSW89] and analytical [WS90b, WS90a] results have shown that training a network with *ensembles* of noisy representations also improves content addressability, so it may be advantageous to include noise in our training formalism.

For these two reasons this work calculates the properties of a network trained with *ensembles of noisy external fields.* The retrieval dynamics under a persistent, noisy external field is also examined, and the effects these two fields have on content addressability compared. To do this we shall look at the fixed-point behaviour of the dynamics which reveals the attractor structure, and from this we judge whether content addressability has been improved. Finally comparisons are made for the three cases when external fields are applied during training only, during retrieval only, and during both stages.

2.2 The External Field Model

The associative memory model we have is a single layered network of N time-dependent binary spin neurons, required to store P uncorrelated patterns, much as defined in §1.2. To considerably simplify calculation of the dynamics, the number C of connections into each site is set at $C \ll \ln N$, or equivalently $(\ln N / \ln C \to$

∞) [KZ]. This high level of dilution in the synapses simplifies the dynamics at each time step by allowing self-averaging at all time steps to be assumed. The reasoning for this can be found in the literature [DGZ87, Ami89].

The space of interactions is continuous and bounded by the spherical constraint given in equation (1.12). The dynamics of the network is conducted by zero temperature parallel update, with each site acting deterministically on the sign of its local field

$$h_i(t) = \sum_{j \neq i}^{C} \frac{J^{ij}}{|\boldsymbol{J}^i|} S_j(t) + \tau_{\mathrm{R}} \zeta_i \xi_i^1 \tag{2.1}$$

which differs from the original equation (1.4) by a persistent external field applied throughout the retrieval phase. This added field of strength τ_{R} can be thought of as a corrupted version of the nominated pattern $\boldsymbol{\xi}^{\mu=1}$ we wish to retrieve. The noise comes from the ζ_i term which follows the discrete probability distribution

$$\mathcal{P}(\zeta_i) = (1 - f_{\mathrm{R}})\delta[\zeta_i - 1] + f_{\mathrm{R}}\delta[\zeta_i + 1] \tag{2.2}$$

where f_{R} is the mean fraction of erroneous sites in the applied field. The resulting dynamical progress in retrieving the designated pattern is again measured by the network's overlap with the nominated pattern, as defined in equation (1.7).

The stated aim of this work was to consider the performance of networks already trained with ensembles of noisy external fields. In deference to Gardner, the network we consider has its connections annealed-optimised, with the noise and stored patterns the two sources of quenched disorder [Gar88]. These connections are optimised to maximise a performance function, in much the same way as a magnetic spin system optimises by seeking out its lowest energy configuration. Moreover, Gardner gave a convergent iterative algorithm to train the network to these optimal connections, one which reflects the performance function used.

Hence the performance function will be intuitively better referred to as the *training function*, a name which stresses its rôle in determining the network's properties.

The training function chosen in the original Gardner model required the state of the network to be invariant to the updating process once the sites match a stored memory pattern. This is more concretely expressed by requiring the alignment field defined in equation (1.6) to be positive definite for all the patterns to be memorised. We can also make this demand more stringent by requiring the alignment field to be larger than some positive parameter κ, which will set the minimal stability of the network. Increasing this stability constant allows the basin of attraction to be enlarged, but its usefulness is limited by a corresponding decrease in the storage capacity [Gar88, For88].

Further enlarging the basin of attraction by an improved training function is the goal of this work, and the idea is to train the network with ensembles of noisy external fields, in anticipation of later retrieval with a statistically similar field. This can be achieved by having

$$g(\Lambda_i^\mu) = \frac{1}{Q}\sum_s^Q \theta[\Lambda_i^\mu + \tau_{\rm T}\zeta_i^{\mu,s} - \kappa] \tag{2.3}$$

for the training function in §1.3, where $\tau_{\rm T}$ is the external training field strength and $\{\zeta_i^{\mu,s}\} = \pm 1$ the noise factor. The noise terms are enumerated over the $s = 1 \ldots Q$ ensembles for each pattern, and follow the same probability distribution as in equation (2.2) above, but with the external field strength and mean fraction of incorrect bits given by the parameters $\tau_{\rm T}$ and $f_{\rm T}$ respectively.

2.3 Quantities to Calculate

Since we are interested in improving the content-addressability of a neural-network model, it will be very helpful if we have a quantitative inkling as to the size of the network's basin of attraction. The way this is done here is by finding an equation which describes the dynamics of the model, and then look for all the fixed-points with respect to the time-step [KA88, Gar89]. We can then define the attractor boundaries and centres by the unstable and stable fixed-points of the dynamics, respectively.

Upon calculating the dynamical equation we find that the network is strongly dependent on the probability distribution of the alignment field. This distribution is entirely determined by how the network is trained, and in the optimally trained case is calculated by the Gardner phase-space treatment [WS90a]. Furthermore it is strongly dependent on two parameters: the *storage capacity* which is simply the number of memory patterns divided by the connectivity of the network ($P \div C$), and the *storage error* which is the fraction of erroneous bits with which each pattern is stored. As we shall see, these quantities are useful in providing ways of controlling the behaviour of the network, allowing its performance to be assessed.

Finally the validity of the assumptions in the replica calculation for the alignment field distribution will be tested by finding the stability of the solution to small perturbations. The method is based on the approach used for the Sherrington-Kirkpatrick spin-glass model [dAT78], with elements of a (in the author's opinion) clearer treatment by Lautrup [Lau88].

2.4 The Iterative Dynamics

We require an expression for describing the retrieval dynamics of the network as measured by the system's overlap with a nominated pattern, as defined in equation (1.7). We shall settle on a simple iterative description, that is one where the overlap at time step $(t+1)$ is some function of the overlap at time t. The derivation is essentially the same as originally done by Abbott and Kepler [KA88] with minor modifications for the external field, so only an outline will be presented.

Since only one memory pattern —the one nominated to be retrieved— is ever considered in this section, the pattern index is superfluous and will be removed from the notation. The conditional probability that a binary perceptron network with an initial overlap m_0 will be taken to an overlap m_1 after one update is

$$\begin{aligned}
&\mathcal{P}(m_1 \mid m_0) = \\
&\sum_{\{S_i\}} \left\{ \delta\left[m_0 - \frac{1}{N}\sum_j^N \xi_j S_j \right] \delta\left[m_1 - \frac{1}{N}\sum_i^N \xi_i \operatorname{sign}\left(\sum_j^N \frac{J^{ij}}{\sqrt{C}} S_j + \tau_{\mathrm{R}} \zeta_i \xi_i \right)\right]\right\} \\
&\quad \div \sum_{\{S_i\}} \left\{ \delta\left[m_0 - \frac{1}{N}\sum_j^N \xi_j S_j \right]\right\}
\end{aligned} \tag{2.4}$$

where the spherical constraint (1.12) and external field have been explicitly inserted into the local field (2.1). This expression is better handled by first introducing a set of fields $\{w_i\}$ at each site and rewriting it as

$$\mathcal{P}(m_1 \mid m_0) = \int \prod_i^N \left[\mathrm{d}w_i \; p(w_i \mid m_0) \right] \delta\left[m_1 - \frac{1}{N}\sum_j^N \operatorname{sign}[w_j] \right], \tag{2.5}$$

leaving the conditional probabilities

$$\begin{aligned} p(w_i \mid m_0) &= \sum_{\{S_i\}} \left\{ \delta\left[m_0 - \frac{1}{N}\sum_j^N \xi_j S_j \right] \delta\left[w_i - \left(\xi_i \sum_j^N \frac{J^{ij}}{\sqrt{C}} S_j + \tau_{\mathrm{R}} \zeta_i \right) \right] \right\} \\ &\quad \div \sum_{\{S_i\}} \left\{ \delta\left[m_0 - \frac{1}{N}\sum_j^N \xi_j S_j \right] \right\} \end{aligned} \tag{2.6}$$

to deal with. This is done by first writing the delta-functions in their integral forms and performing the configurational sums over the spins $\{S_i\}$, giving

$$\begin{aligned} & p(w_i \mid m_0) = \\ & \quad \int \frac{\mathrm{d}x\mathrm{d}y}{2\pi} \exp\left(\mathrm{i}xNm_0 + \mathrm{i}y(w_i - \tau_{\mathrm{R}}\zeta_i) + \sum_j^N \ln\cos\left[x\xi_j + y\xi_i \frac{J^{ij}}{\sqrt{C}} \right] \right) \\ & \quad \div \int \mathrm{d}x \exp\left(\mathrm{i}xNm_0 + \sum_j^N \ln\cos\left[x\xi_j \right] \right) . \end{aligned}$$

Taking the connectivity to the thermodynamic limit and making use of the alignment field Λ_i defined in equation (1.6) (again without the superfluous pattern index), the cosine term is then expanded to give

$$\begin{aligned} & p(w_i \mid m_0) = \int \frac{\mathrm{d}x\mathrm{d}y}{2\pi} \\ & \quad \exp\left(\mathrm{i}xNm_0 + \mathrm{i}y(w_i - \tau_{\mathrm{R}}\zeta_i) + N\ln\cos x - \frac{1}{2}y^2(1+\tan^2 x) - y\Lambda_i \tan x \right) \\ & \quad \div \int \mathrm{d}x \exp\left(\mathrm{i}xNm_0 + \sum_j^N \ln\cos\left[x\xi_j \right] \right) \end{aligned}$$

which allows the x-integrations to be done by the method of steepest descent. The denominator subsequently divides out leaving a straightforward Gaussian integral in y to do. Inserting the result of this into expression (2.5), and writing another integral-representation for the remaining delta-function, we get

$$
\begin{aligned}
\mathcal{P}(m_1 \mid m_0) &= \int \frac{\mathrm{d}z}{2\pi} \exp(\mathrm{i}zm_1) \\
&\times \int \prod_i^N \left\{ \frac{\mathrm{d}w_i}{\sqrt{2\pi}} \frac{1}{\sqrt{1-m_0^2}} \exp\left(-\frac{1}{2}\left[\frac{w_i - m_0\Lambda_i - \tau_{\mathrm{R}}\zeta_i}{\sqrt{1-m_0^2}} \right]^2 - \frac{\mathrm{i}z}{N}\mathrm{sign}(w_i) \right) \right\}
\end{aligned}
$$

which can be rewritten using the error-function defined in equation (A.3). In the large system size N limit the integral over z reduces to a delta-function to give the overlap at the first time-step

$$
m_1 = \frac{1}{N} \sum_i^N \mathrm{erf}\left[\frac{m_0\Lambda_i + \tau_{\mathrm{R}}\zeta_i}{\sqrt{2(1-m_0^2)}} \right]
$$

with probability one. Finally, in the limit of a large system size N we can use self-averaging to rewrite this as an average over the alignment field and external-field noise,

$$
m_1 = \left\langle\!\!\left\langle \int \mathrm{d}\Lambda\ \rho(\Lambda)\ \mathrm{erf}\left[\frac{m_0\Lambda + \tau_{\mathrm{R}}\zeta}{\sqrt{2(1-m_0^2)}} \right] \right\rangle\!\!\right\rangle_{\zeta} \tag{2.7}
$$

where $\rho(\Lambda)$ is the probability distribution of the alignment field and the double-angled brackets denote an average over the noise in the external field. The task of finding the probability distribution for the alignment field is the subject of the next subsection.

The above expression (2.7) is exact for the first time step, and also in the case of low connectivity as mentioned earlier in §2.2. The validity of this simplification has been confirmed by direct numerical simulations [Hen91] of dilute networks. Furthermore, earlier work simulating fully connected networks [KA88] has stressed the importance of the first time step dynamics by showing it to be highly indicative of the network's ultimate fate, a result which broadens the generality of the iterative map calculation.

2.5 The Alignment Field Distribution

The previous section's derivation of equation (2.7) describing the model's iterative dynamics ended with the introduction of the alignment field distribution. This distribution should be thought of as the encapsulation of an optimally trained network, and be independent of the retrieval dynamics we care to choose.

Averaging Over the External Field Noise

We shall first make use of the simplification to a single perceptron as justified in §1.2.1, and write the distribution function of the alignment field as

$$\begin{aligned} \rho(\Lambda) \;&=\; \left\langle\!\!\left\langle \int \prod_j^C \mathrm{d}J^j \delta[\Lambda - \Lambda^1] \delta[\sum_j (J^j)^2 - C] \exp\left(\beta \sum_\mu^P g(\Lambda^\mu)\right) \right.\right. \\ &\left.\left. \div \int \prod_j^C \mathrm{d}J^j \delta[\sum_j (J^j)^2 - C] \exp\left(\beta \sum_\mu^P g(\Lambda^\mu)\right) \right\rangle\!\!\right\rangle_{\zeta,\xi} \qquad (2.8) \end{aligned}$$

where one delta function picks out the alignment field for a typical pattern $\boldsymbol{\xi}^{\mu=1}$, and the other explicitly enforces the spherical constraint. This function is an extensive quantity, so it is a suitable object for which to perform the quenched average over, as denoted by the double angled brackets introduced in §1.1.1. Replicating over the annealed variables, equation (2.8) can be written as

$$\begin{aligned} \rho(\Lambda) \;&=\; \lim_{n\to 0} \left\langle\!\!\left\langle \int \prod_{a,j}^{n,C} \mathrm{d}J_a^j \delta[\Lambda - \Lambda_1^1] \right.\right. \\ &\left.\left. \prod_a^n \left\{ \delta[\sum_j (J_a^j)^2 - C] \exp\left(\beta \sum_\mu^P g(\Lambda_a^\mu)\right) \right\} \right\rangle\!\!\right\rangle_{\zeta,\xi} \qquad (2.9) \end{aligned}$$

which now picks out the alignment field for a typical replica $a = 1$ as well as a typical pattern $\mu = 1$. We can now perform the quenched average over the fixed disordered quantities, beginning with the noise term in the external training field $\{\zeta^{\mu,s}\}$ which appears in the form

$$\left\langle\!\!\left\langle \prod_{a}^{n} \exp\left(\beta \sum_{\mu}^{P} g(\Lambda_a^\mu)\right)\right\rangle\!\!\right\rangle_{\{\zeta^{\mu,s}\}} = \left\langle\!\!\left\langle \exp\left(\frac{\beta}{Q} \sum_{a,\mu,s}^{n,P,Q} \theta[\Lambda_a^\mu + \tau_{\mathrm{T}}\zeta^{\mu,s} - \kappa]\right)\right\rangle\!\!\right\rangle_{\{\zeta^{\mu,s}\}}$$

where these angled-brackets are specifically over the noise. Performing this average via the probability distribution given in equation (2.2) we arrive at

$$\prod_{\mu,s}^{P,Q} \left\langle\!\!\left\langle \exp\left(\frac{\beta}{Q} \sum_{a}^{n} \theta[\Lambda_a^\mu + \tau_{\mathrm{T}}\zeta^{\mu,s} - \kappa]\right)\right\rangle\!\!\right\rangle_{\{\zeta^{\mu,s}\}} = \prod_{\mu}^{P} \left\{ \sum_{r=0}^{Q} \binom{Q}{r} (1 - f_{\mathrm{T}})^r {f_{\mathrm{T}}}^{Q-r} \right.$$
$$\left. \times \exp\left(\frac{\beta}{Q} \sum_{a}^{n} \{r\theta[\Lambda_a^\mu + \tau_{\mathrm{T}} - \kappa] + (Q - r)\theta[\Lambda_a^\mu - \tau_{\mathrm{T}} - \kappa]\}\right)\right\}. \quad (2.10)$$

In the limit of large number of ensembles Q, the binomial sum in equation (2.10) can be considerably simplified by substituting the $(Q + 1)$ terms with the single mean term $\bar{r} = Q(1 - f_{\mathrm{T}})$ to give

$$\left\langle\!\!\left\langle \prod_{a}^{n} \exp\left(\beta \sum_{\mu}^{P} g(\Lambda_a^\mu)\right)\right\rangle\!\!\right\rangle_{\{\zeta^{\mu,s}\}} = \prod_{\mu}^{P} \exp\left(\beta \sum_{a}^{n} \bar{g}(\Lambda_a^\mu)\right).$$

where we have defined a mean training function $\bar{g}(\lambda)$

$$\bar{g}(\lambda) \equiv (1 - f_{\mathrm{T}})\ \theta[\lambda + \tau_{\mathrm{T}} - \kappa] + f_{\mathrm{T}}\ \theta[\lambda - \tau_{\mathrm{T}} - \kappa]. \quad (2.11)$$

The remaining disordered quantity to quenched average are the memory patterns. This quantity only occurs in the alignment field and can be extracted from the training function via equation (A.7) to give

$$\begin{aligned}\rho(\Lambda) &= \lim_{n\to 0}\int\prod_{a,\mu}^{n,P}\frac{\mathrm{d}\lambda_a^\mu \mathrm{d}x_a^\mu}{2\pi}\int\prod_{a,j}^{n,C}\mathrm{d}J_a^j \\ &\quad \delta[\Lambda-\lambda_1^1]\prod_a^n\delta[\sum_j^C(J_a^j)^2-C]\prod_\mu^P\exp\sum_a^n[\mathrm{i}x_a^\mu\lambda_a^\mu+\beta\bar{g}(\lambda_a^\mu)] \\ &\quad \times\left\langle\!\!\left\langle\prod_\mu^P\exp\left(-\mathrm{i}\sum_a^n x_a^\mu\Lambda_a^\mu\right)\right\rangle\!\!\right\rangle_{\{\xi_i^\mu\}} . \end{aligned} \qquad (2.12)$$

The term in equation (2.12) dependent on the patterns can be treated in isolation, by first writing in the input and output patterns

$$\left\langle\!\!\left\langle\prod_\mu^P\exp\left(-\mathrm{i}\sum_a^n x_a^\mu\Lambda_a^\mu\right)\right\rangle\!\!\right\rangle_{\{\xi_i^\mu\}}=\prod_{\mu,j}^{P,C}\left\langle\!\!\left\langle\exp\left(-\mathrm{i}\sum_a^n x_a^\mu\mathrm{T}^\mu\frac{J_a^j}{\sqrt{C}}\xi_j^\mu\right)\right\rangle\!\!\right\rangle_{\{\xi_i^\mu\}}$$

where T^μ can be viewed as the i^{th} bit of pattern $\boldsymbol{\xi}^\mu$. Upon performing the average over all the input patterns $\{\boldsymbol{\xi}^\mu\}$ and (uncorrelated) output targets $\{\mathrm{T}^\mu\}$ we can, in the thermodynamic limit of the connectivity C going to infinity, expand and exponentiate the resulting cosine function to get

$$\begin{aligned}\left\langle\!\!\left\langle\prod_\mu^P\exp\left(-\mathrm{i}\sum_a^n x_a^\mu\Lambda_a^\mu\right)\right\rangle\!\!\right\rangle_{\{\xi_i^\mu\}} &= \prod_{\mu,j}^{P,C}\cos\sum_a^n x_a^\mu\frac{J_a^j}{\sqrt{C}} \\ &= \prod_\mu^P\exp\left(-\frac{1}{2C}\sum_{a,b,j}^{n,n,C}x_a^\mu J_a^j J_b^j x_b^\mu\right) . \end{aligned} \qquad (2.13)$$

This term (2.13) contains an interaction between the various replicas, which in the spirit of the replica-method is dealt with by introducing the inter-replica correlation parameter q_{ab} as defined by equation (1.11) in §1.3. Inserting this inter-replica correlation parameter and writing in integral form the delta function which enforces the spherical constraint by equation (A.6), we find

$$\begin{aligned}
\rho(\Lambda) \;\; = \;\; & \lim_{\substack{n\to 0\\ C\to\infty}} \int \prod_{a}^{n} \frac{\mathrm{d}E_a}{4\pi} \int \prod_{a<b} \frac{\mathrm{d}F_{ab}\mathrm{d}q_{ab}}{2\pi/C} \int \prod_{a,\mu}^{n,P} \frac{\mathrm{d}\lambda_a^\mu \mathrm{d}x_a^\mu}{2\pi} \int \prod_{a,j}^{n,C} \mathrm{d}J_a^j \\
& \delta[\Lambda - \lambda_1^1] \prod_{a,j}^{n,C} \exp\left(-\frac{\mathrm{i}}{2} E_a[(J_a^j)^2 - 1]\right) \\
& \times \prod_{a<b} \exp\left(-\mathrm{i}F_{ab}[Cq_{ab} - \sum_j^C J_a^j J_b^j]\right) \\
& \times \prod_{a,\mu}^{n,P} \exp\left(\mathrm{i}x_a^\mu \lambda_a^\mu + \beta\bar{g}(\lambda_a^\mu)\right) \prod_{\mu}^{P} \exp\left(-\frac{1}{2}\sum_a^n (x_a^\mu)^2 - \sum_{a<b} x_a^\mu q_{ab} x_b^\mu\right) .
\end{aligned}$$

This expression can be collected into a form explicitly amenable for integration by the method of steepest descent,

$$\begin{aligned}
\rho(\Lambda) \;\; = \;\; & \lim_{\substack{n\to 0\\ C\to\infty}} \int \prod_{a}^{n} \frac{\mathrm{d}E_a}{4\pi} \int \prod_{a<b} \frac{\mathrm{d}F_{ab}\mathrm{d}q_{ab}}{2\pi/C} \\
& \exp C \left\{ G_J(\{E_a, F_{ab}\}) + \alpha G_\Lambda(\{q_{ab}\}) + G_0(\{E_a, F_{ab}, q_{ab}\}) \right\} \\
& \times \int \prod_{a}^{n} \frac{\mathrm{d}\lambda_a^1 \mathrm{d}x_a^1}{2\pi} \delta[\Lambda - \lambda_1^1] \\
& \times \exp\left\{ -\frac{1}{2}\sum_a^n (x_a^1)^2 - \sum_{a<b} x_a^1 q_{ab} x_b^1 + \sum_a^n \left[\mathrm{i}x_a^1\lambda_a^1 + \beta\bar{g}(\lambda_a^1)\right] \right\} . \quad (2.14)
\end{aligned}$$

The functions in the exponential have taken advantage of any symmetries in the site and pattern indices and are defined as

$$G_J(\{E_a, F_{ab}\}) \;\; \equiv \;\; \ln \int \prod_a^n \mathrm{d}J_a \exp\left(-\frac{\mathrm{i}}{2}\sum_a^n E_a J_a^2 + \mathrm{i}\sum_{a<b} J_a F_{ab} J_b\right) ,$$

$$\begin{aligned} G_\Lambda(\{q_{ab}\}) &\equiv \ln \int \prod_a^n \frac{\mathrm{d}\lambda_a \mathrm{d}x_a}{2\pi} \exp\left(-\frac{1}{2}\sum_a^n x_a^2 - \sum_{a<b} x_a q_{ab} x_b \right. \\ &\qquad \left. + \sum_a^n [\mathrm{i}x_a\lambda_a + \beta\bar{g}(\lambda_a)] \right), \\ G_0(\{E_a, F_{ab}, q_{ab}\}) &\equiv \frac{\mathrm{i}}{2}\sum_a^n E_a - \mathrm{i}\sum_{a<b} F_{ab} q_{ab}. \end{aligned} \tag{2.15}$$

In equation (2.14) the parameter α is the storage capacity of the network and is defined in §2.3 as $\alpha \equiv P \div C$.

The Saddle Point Equations

In the thermodynamic limit the integrations over the $\{E_a, F_{ab}, q_{ab}\}$ variables in equation (2.14) can be performed by the method of steepest descent. By differentiating equations (2.15) with respect to these three sets of variables, the stationary point can be found by solving the equations

$$\begin{aligned} 0 &= -\tfrac{\mathrm{i}}{2}\{J_a^2\}_{G_J} + \tfrac{\mathrm{i}}{2}, && \forall a, \\ 0 &= \mathrm{i}\{J_a J_b\}_{G_J} - \mathrm{i}q_{ab}, && \forall a, b\ (a<b), \\ 0 &= -\alpha\{x_a x_b\}_{G_\Lambda} - \mathrm{i}F_{ab}, && \forall a, b\ (a<b), \end{aligned} \tag{2.16}$$

where we have introduced braces operators in the interactions and alignment field. Taking the following replica symmetric ansatz for the three steepest-descent integration variables

$$\begin{aligned} E_a &= E, \ \forall a, \\ F_{ab} &= F, \ \forall a, b\ (a<b), \\ q_{ab} &= q, \ \forall a, b\ (a<b), \end{aligned} \tag{2.17}$$

these two operators can be written down as

$$
\begin{aligned}
\{(\cdots)\}_{G_J} &= \int\prod_a^n \mathrm{d}J_a\,(\cdots)\,\exp\left(-\frac{\mathrm{i}}{2}(E+F)\sum_a^n J_a^2+\frac{\mathrm{i}F}{2}\left[\sum_a J_a\right]^2\right)\\
&\quad\div\int\prod_a^n \mathrm{d}J_a\exp\left(-\frac{\mathrm{i}}{2}(E+F)\sum_a^n J_a^2+\frac{\mathrm{i}F}{2}\left[\sum_a J_a\right]^2\right),\\
\{(\cdots)\}_{G_\Lambda} &= \int\prod_a^n \frac{\mathrm{d}\lambda_a\mathrm{d}x_a}{2\pi}\,(\cdots)\,\exp\left(-\frac{1}{2}(1-q)\sum_a^n x_a^2-\frac{q}{2}\left[\sum_a x_a\right]^2\right.\\
&\qquad\left.+\sum_a^n\left[\mathrm{i}x_a\lambda_a+\beta\bar{g}(\lambda_a)\right]\right)\\
&\quad\div\int\prod_a^n \frac{\mathrm{d}\lambda_a\mathrm{d}x_a}{2\pi}\exp\left(-\frac{1}{2}(1-q)\sum_a^n x_a^2-\frac{q}{2}\left[\sum_a x_a\right]^2\right.\\
&\qquad\left.+\sum_a^n\left[\mathrm{i}x_a\lambda_a+\beta\bar{g}(\lambda_a)\right]\right). \qquad (2.18)
\end{aligned}
$$

The integrations over $\{J_a\}$ and $\{x_a\}$ are done by first performing Hubbard-Stratonovich transformations (A.2) on the numerators and denominators in order to linearise the summed-squared terms in the exponentials. It can then be seen that the denominators will involve some finite expression raised to the number of replicas-n, hence going to unity in the $(n \to 0)$ limit. This same limit allows the two simpler operators

$$
\begin{aligned}
\langle(\cdots)\rangle_J &\equiv \int \mathrm{d}J\,(\cdots)\,\exp\left(-\frac{\mathrm{i}}{2}(E+F)J^2+uJ\sqrt{\mathrm{i}F}\right)\\
&\quad\div\int\mathrm{d}J\exp\left(-\frac{\mathrm{i}}{2}(E+F)J^2+uJ\sqrt{\mathrm{i}F}\right),\\
\langle(\cdots)\rangle_\Lambda &\equiv \int\frac{\mathrm{d}\lambda\mathrm{d}x}{2\pi}\,(\cdots)\,\exp\left(-\frac{1}{2}(1-q)x^2+\mathrm{i}x(\lambda-z\sqrt{q})+\beta\bar{g}(\lambda)\right)\\
&\quad\div\int\frac{\mathrm{d}\lambda\mathrm{d}x}{2\pi}\exp\left(-\frac{1}{2}(1-q)x^2+\mathrm{i}x(\lambda-z\sqrt{q})+\beta\bar{g}(\lambda)\right) \qquad (2.19)
\end{aligned}
$$

to be defined, where the quantities u and z originated from the said linearising transformations. The saddle-point equations (2.16) can hence be written as

$$
\begin{aligned}
1 &= \left\{J_a^2\right\}_{G_J} \\
&= \int \mathcal{D}u \ \left\{\langle J^2\rangle_J\right\}, \\
q &= \left\{J_a J_b\right\}_{G_J} \\
&= \int \mathcal{D}u \ \left\{\langle J\rangle_J\right\}^2, \\
\mathrm{i}F &= -\alpha \left\{x_a x_b\right\}_{G_\Lambda} \\
&= -\alpha \int \mathcal{D}z \ \left\{\langle x\rangle_\Lambda\right\}^2 .
\end{aligned}
\tag{2.20}
$$

Expressions of the form $\langle J^k\rangle_J$ for an arbitrary moment k are straightforward Gaussian integrals. The two cases we require to solve equations (2.20) are

$$
\begin{aligned}
\langle J\rangle_J &= \frac{u}{\sqrt{\mathrm{i}E+\mathrm{i}F}}\sqrt{\frac{F}{E+F}}, \\
\langle J^2\rangle_J &= \frac{1}{\mathrm{i}E+\mathrm{i}F}\left(1+u^2\frac{F}{E+F}\right).
\end{aligned}
\tag{2.21}
$$

Some authors have removed the imaginary nature of E and F by transforming their integrations in equation (2.14) to lie along the imaginary line, but this is at the expense of some confusion when evaluating the stability of the saddle-points, which we shall prefer to avoid. The corresponding expression $\langle x^k\rangle_\Lambda$ for an arbitrary moment of x is a somewhat more complicated. The integrations over the x variable is a Gaussian and upon completing squares can be expressed as

$$
\begin{aligned}
\langle x^k\rangle_\Lambda = &\int \frac{\mathrm{d}\lambda}{\sqrt{2\pi}} \int \mathcal{D}x \ (1-q)^{-\frac{1}{2}(1+k)} \left[x + \mathrm{i}\left(\frac{\lambda - z\sqrt{q}}{\sqrt{1-q}}\right)\right]^k \\
&\times \exp\left(\beta\bar{g}(\lambda) + \mathrm{i}x\left[\frac{\lambda - z\sqrt{q}}{\sqrt{1-q}}\right]\right) \\
&\div \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi}}(1-q)^{-\frac{1}{2}} \exp\beta\left(\bar{g}(\lambda) - \frac{(\lambda - z\sqrt{q})^2}{2\beta(1-q)}\right).
\end{aligned}
\tag{2.22}
$$

We can see from equation (2.22) that in the limit of a zero annealed temperature ($\beta \rightarrow \infty$), the previously difficult integrations over λ are now amenable to solving by the method of steepest descent. In return this requires us to find a $\hat{\lambda}(z)$ which maximises the function

$$\begin{aligned}\mathcal{G}(\lambda) &\equiv \bar{g}(\lambda) - \frac{1}{\gamma^2}(\lambda - z\sqrt{q})^2, \\ \gamma^2 &\equiv 2\beta(1-q),\end{aligned} \tag{2.23}$$

where we have surreptitiously avoided the problem of the $1/\beta$ factor by introducing a parameter γ. From equation (1.11), we can interpret the saddle-point solution of q as the order-parameter of the overlaps between the differing solutions under identical training conditions. Thus when we take the ($q \rightarrow 1$) limit we are forcing all the solutions to converge together, a scenario Gardner called the optimal limit because, as we shall see in §2.7, it corresponds to the case of having the maximum number of memory patterns learnt. Introducing the parameter γ now allows us to take the limits of a zero annealed temperature ($\beta \rightarrow \infty$) and an optimal perceptron ($q \rightarrow 1$) together, as this corresponds to keeping γ finite.

Returning to equations (2.20), we find the required moment of equation (2.22) is

$$\langle x \rangle_\Lambda = \frac{\mathrm{i}}{\sqrt{1-q}} \left(\frac{\hat{\lambda}(z) - z}{\sqrt{1-q}} \right), \tag{2.24}$$

after having taken the zero temperature and optimal-solution limits. Collecting the pieces from equations (2.20)–(2.24), we arrive at the saddle-point equations now equal to

$$1 = \frac{1}{\mathrm{i}E + \mathrm{i}F}\left(1 + \frac{F}{E+F}\right),$$

$$\begin{aligned} q &= \frac{1}{\mathrm{i}E+\mathrm{i}F}\left(\frac{F}{E+F}\right), \\ \mathrm{i}F &= \frac{\alpha}{(1-q)^2}\int \mathcal{D}z\left[\hat{\lambda}(z)-z\right]^2 . \end{aligned} \tag{2.25}$$

Finally, taking the zero replica limit for the alignment field distribution of equation (2.14) gives

$$\begin{aligned} \rho(\Lambda) &= \int \mathcal{D}z\ \langle \delta[\Lambda-\lambda_1^1]\rangle_\Lambda \\ &= \int \mathcal{D}z\ \delta[\Lambda-\hat{\lambda}(z)] \end{aligned} \tag{2.26}$$

which has been written in terms of the function $\hat{\lambda}(z)$ which maximises equation (2.23). Finding this function $\hat{\lambda}(z)$ is in principle straightforward, but in practice has been quite involved, as will be seen in the following section.

2.6 Maximising the Function $\mathcal{G}(\lambda)$

The stated aim of this section is to find the values of $\lambda = \hat{\lambda}$ which maximise the function $\mathcal{G}(\lambda)$ defined in equation (2.23), corresponding to the stationary points in a steepest descent evaluation. Hence by definition (2.11) we are trying to find the λ's which maximises the function

$$\mathcal{G}(\lambda) = (1-f_{\mathrm{T}})\ \theta[\lambda+\tau_{\mathrm{T}}-\kappa] + f_{\mathrm{T}}\ \theta[\lambda-\tau_{\mathrm{T}}-\kappa] - \frac{1}{\gamma^2}(\lambda-z)^2. \tag{2.27}$$

in the optimal perceptron limit of ($q \to 1$). Unfortunately, we can quickly see that the interplay between the above quantities γ, the (Gaussian in origin) variable z, and the field parameters f_{T} (fraction of erroneous bits) and τ_{T} (field strength) considerably complicates the task of finding $\hat{\lambda}$. Firstly the stationary point is strongly dependent on z, with discontinuous jumps encountered as it moves along the real number line, in a manner analogous to first-order phase transitions between the lowest energy states. This point is also dependent on the remaining three parameters, but thankfully they can be grouped into distinct regimes of parameter space. The task then is to identify these regimes, and what the so-called *maximising function* $\hat{\lambda}(z)$ will look like in them as z is varied.

We shall first find what are the possible values $\hat{\lambda}(z)$ can take. Equation (2.27) has discontinuities at $\lambda = (\kappa - \tau_{\mathrm{T}})$ and $\lambda = (\kappa + \tau_{\mathrm{T}})$, hence if we define the Heaviside step function to return unity with a zero argument, we find $\hat{\lambda}(z)$ can be $\hat{\lambda}(z) = (\kappa - \tau_{\mathrm{T}})$ or $\hat{\lambda}(z) = (\kappa + \tau_{\mathrm{T}})$. In addition, when the quadratic term is dominant, equation (2.27) will be maximised by $\hat{\lambda}(z) = z$. These then are the three possible values the maximising function $\hat{\lambda}(z)$ can take.

The next step is to determine which values of z will cause a transition between these three values of the maximising function. This is done by substituting each of the possible values of $\hat{\lambda}(z)$ into equation (2.27) and solving for z. That is, by solving the three equations $\mathcal{G}(\hat{\lambda}(z) = \kappa \pm \tau_{\mathrm{T}}) = \mathcal{G}(\hat{\lambda}(z) = z)$ and $\mathcal{G}(\hat{\lambda}(z) = \kappa + \tau_{\mathrm{T}}) = \mathcal{G}(\hat{\lambda}(z) = \kappa - \tau_{\mathrm{T}})$ for z. These solutions of z give the possible transitions as occurring at six points, which are given in equation (C.1) in the appendix.

The order in which these six *transition points* occur along the z number line now needs to be determined by comparing fifteen inequalities. Appendix §C.1 gives the results of this, and concludes that the γ and external field parameters group themselves into six possible regimes, as given by expression (C.3) which presents them in terms of ranges in the external field strength τ_{T}.

For each of these six ranges in the field strength, we now need to determine $\hat{\lambda}(z)$. This is done by equating z with each of the six possible transition points and determining which one of the three $\hat{\lambda}(z) = \{(\kappa - \tau_{\mathrm{T}}), (\kappa + \tau_{\mathrm{T}}), z\}$ possible choices gives the largest value to the function (2.27) we wish to maximise. Further details of this procedure are given in appendix §C.2. The pieces are combined in §C.3 and we find the ranges (C.3) can actually be reduced to just three regimes, inside each of which $\hat{\lambda}(z)$ will look like

Regime 1 $\quad -\infty < 2\tau_{\mathrm{T}} < \gamma\left[1-\sqrt{1-f_{\mathrm{T}}}\right]$

$$\begin{aligned}
\hat{\lambda}(z\epsilon[-\infty, \kappa+\tau_{\mathrm{T}}-\gamma]) &= z, \\
\hat{\lambda}(z\epsilon[\kappa+\tau_{\mathrm{T}}-\gamma, \kappa+\tau_{\mathrm{T}}]) &= (\kappa+\tau_{\mathrm{T}}), \\
\hat{\lambda}(z\epsilon[\kappa+\tau_{\mathrm{T}}, \infty]) &= z.
\end{aligned}$$

Regime 2 $\quad \gamma\left[1-\sqrt{1-f_{\mathrm{T}}}\right] < 2\tau_{\mathrm{T}} < \gamma\sqrt{f_{\mathrm{T}}}$

$$\begin{aligned}
\hat{\lambda}(z\epsilon[-\infty, \kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}]) &= z, \\
\hat{\lambda}(z\epsilon[\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}, \kappa-\gamma^2 f_{\mathrm{T}}/(4\tau_{\mathrm{T}})]) &= (\kappa-\tau_{\mathrm{T}}), \\
\hat{\lambda}(z\epsilon[\kappa-\gamma^2 f_{\mathrm{T}}/(4\tau_{\mathrm{T}}), \kappa+\tau_{\mathrm{T}}]) &= (\kappa+\tau_{\mathrm{T}}), \\
\hat{\lambda}(z\epsilon[\kappa+\tau_{\mathrm{T}}, \infty]) &= z.
\end{aligned}$$

Regime 3 $\quad \gamma\sqrt{f_{\mathrm{T}}} < 2\tau_{\mathrm{T}} < \infty$

$$\begin{aligned}
\hat{\lambda}(z\epsilon[-\infty, \kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}]) &= z, \\
\hat{\lambda}(z\epsilon[\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}, \kappa-\tau_{\mathrm{T}}]) &= (\kappa-\tau_{\mathrm{T}}), \\
\hat{\lambda}(z\epsilon[\kappa-\tau_{\mathrm{T}}, \kappa+\tau_{\mathrm{T}}-\gamma\sqrt{f_{\mathrm{T}}}]) &= z, \\
\hat{\lambda}(z\epsilon[\kappa+\tau_{\mathrm{T}}-\gamma\sqrt{f_{\mathrm{T}}}, \kappa+\tau_{\mathrm{T}}]) &= (\kappa+\tau_{\mathrm{T}}), \\
\hat{\lambda}(z\epsilon[\kappa+\tau_{\mathrm{T}}, \infty]) &= z.
\end{aligned} \tag{2.28}$$

The second and third regimes differ from the original Gardner-Derrida result[1] [GD88] but, as we can verify, they do reduce to the referenced result in the limit of a zero external field strength. Having obtained $\hat{\lambda}(z)$ for the three differing regimes, we are now in a position to give concrete results for the quantities discussed in §2.3.

[1] Which looks like the first regime but with $(\kappa+\tau_{\mathrm{T}})$ replaced by κ.

2.7 The Optimal Storage Capacity

Returning to the saddle-point equations (2.25), we can see that by eliminating the quantities E and F we can obtain an expression for the storage capacity. As we approach the optimal perceptron limit of ($q \to 1$), the storage capacity increases up to the *optimal storage capacity* α_C, defined as

$$\alpha_C \equiv \left[\int \mathcal{D}z\Big(\hat{\lambda}(z) - z\Big)^2\right]^{-1}. \tag{2.29}$$

The maximising function $\hat{\lambda}(z)$ for the three identified regimes can then be inserted from equations (2.28) to give the optimal storage capacity as:

Regime 1 $\quad -\infty < 2\tau_{\mathrm{T}} < \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right]$

$$[\alpha_C]^{-1} = \int_{\kappa+\tau_{\mathrm{T}}-\gamma}^{\kappa+\tau_{\mathrm{T}}} \mathcal{D}z(z - (\kappa + \tau_{\mathrm{T}}))^2,$$

Regime 2 $\quad \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right] < 2\tau_{\mathrm{T}} < \gamma\sqrt{f_{\mathrm{T}}}$

$$[\alpha_C]^{-1} = \int_{\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}}^{\kappa-\frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}} \mathcal{D}z(z - (\kappa - \tau_{\mathrm{T}}))^2 + \int_{\kappa-\frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}}^{\kappa+\tau_{\mathrm{T}}} \mathcal{D}z(z - (\kappa + \tau_{\mathrm{T}}))^2,$$

Regime 3 $\quad \gamma\sqrt{f_{\mathrm{T}}} < 2\tau_{\mathrm{T}} < \infty$

$$[\alpha_C]^{-1} = \int_{\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}}^{\kappa-\tau_{\mathrm{T}}} \mathcal{D}z(z - (\kappa - \tau_{\mathrm{T}}))^2 + \int_{\kappa+\tau_{\mathrm{T}}-\gamma\sqrt{f_{\mathrm{T}}}}^{\kappa+\tau_{\mathrm{T}}} \mathcal{D}z(z - (\kappa + \tau_{\mathrm{T}}))^2. \tag{2.30}$$

2.8 The Fractional Storage Error

We may understandably feel uncomfortable at giving the parameter γ defined in equation (2.23) any significance other than as a mathematical artefact. Appendix B attempts to assuage this apprehension by making an analogy with spin-glass models, but we can also alleviate it by expressing γ in terms of some physically more obvious parameter, namely the *fractional storage error*,

$$\begin{aligned}\mathcal{F} \;\equiv\;& \frac{1}{P}\sum_{\mu}^{P}\left\langle\!\!\left\langle \int\prod_{j}^{C}\mathrm{d}J^{j}\;\theta[-\Lambda^{\mu}]\;\delta[\sum_{j}(J^{j})^{2}-C]\exp\left(\beta\sum_{\mu}^{P}g(\Lambda^{\mu})\right)\right.\right.\\ &\left.\left.\div\int\prod_{j}^{C}\mathrm{d}J^{j}\;\delta[\sum_{j}(J^{j})^{2}-C]\exp\left(\beta\sum_{\mu}^{P}g(\Lambda^{\mu})\right)\right\rangle\!\!\right\rangle_{\zeta,\xi}\end{aligned} \tag{2.31}$$

which is the average fraction of bits per pattern and (since we are explicitly using the single-perceptron simplification in our notation) per site that is erroneously stored by the network. This defining equation is treated on similar lines to that in the calculation of the alignment field, and we find we can write expression (2.31) for a typical replica-1 and a typical pattern-1 as

$$\begin{aligned}\mathcal{F} &= \int\mathcal{D}z\;\langle\theta[-\lambda_{1}^{1}]\rangle_{\Lambda}\\ &= \int\mathcal{D}z\;\theta[-\hat{\lambda}(z)]\\ &= \int\mathrm{d}\Lambda\int\mathcal{D}z\;\delta[\Lambda-\hat{\lambda}(z)]\;\theta[-\Lambda]\\ &= \int_{-\infty}^{0}\mathrm{d}\Lambda\;\rho(\Lambda),\end{aligned} \tag{2.32}$$

using the alignment field distribution in equation (2.26). It can be shown [GD88] that the fractional error $\mathcal{F}$ is only equal to zero in the limit of the parameter γ tending to infinity.

We may interpret equation (2.32) literally and regard the fractional storage error as a consequence of the negative and destabilising part of the alignment field distribution. Another way of viewing the storage error is as a way of controlling the strictness of the network's training, such that the further above zero it is —and the further γ moves away from infinity— the less strictly enforced is the training function.

The consequences of having a non-zero storage error will be discussed in the results for the alignment field distribution. Before that we shall examine the validity of the replica-symmetric ansatz by a stability analysis around the saddle-point.

2.9 Stability of Replica Symmetry

The question of whether the replica-symmetric ansatz is a valid one to use is addressed here. The analysis follows closely that by de-Almeida and Thouless [dAT78], and in essence examines whether the saddle-point solution of the replicated quantities is stable to small, local, replica-symmetry breaking fluctuations. In the original paper the saddle-integrations were over real variables, so the criterion for stability was re-cast by asking whether the stationary point was a maximum[2]; that is, whether the eigenvalues of the stability matrix were all negative definite[Wid61]. In this problem the stationary point is located in a complex-space and the stability condition determined by sign changes in the determinant of the stability matrix.

de-Almeida and Thouless found that the elements of the Hessian matrix had certain symmetry properties which not only simplified its treatment considerably but allowed the actual eigenvalues to be found. Appendix §D.1 illustrates how this is done from an alternative perspective [Lau88], from which the relevant re-

[2] In which case we were actually integrating by '*The method of Laplace*'.

sults will just be quoted here. The fluctuations arising from these symmetries are characterised by three types: *symmetric*, *weakly-asymmetric* and *asymmetric* fluctuations. However, §D.1.1 and §D.1.2 show the first two of these give identical determinants in the zero replica limit so we shall only examine the cases of symmetric and asymmetric fluctuations.

2.9.1 Symmetric Fluctuations

Appendix §D.1 gives the simplified Hessian for symmetric fluctuations as matrix (D.10), with the product of eigenvalues to these fluctuations given by the determinant

$$|\mathrm{Q}^{\mathrm{s}}| = \begin{vmatrix} (Q^A_{EE} - Q^B_{EE}) & -(Q^A_{EF} - Q^B_{EF}) & 0 \\ 2(Q^A_{EF} - Q^B_{EF}) & (Q^A_{FF} - 4Q^B_{FF} + 3Q^C_{FF}) & -\mathrm{i} \\ 0 & -\mathrm{i} & (Q^A_{qq} - 4Q^B_{qq} + 3Q^C_{qq}) \end{vmatrix} \tag{2.33}$$

where the matrix elements reflect the possible permutation symmetries in the replica indices, namely

$$\begin{aligned}
Q^A_{EE} &\equiv -\frac{1}{4}\int \mathcal{D}u \left\{\langle J^4\rangle_J\right\} + \frac{1}{4}\int \mathcal{D}u \left\{\langle J^2\rangle_J\right\} \int \mathcal{D}u \left\{\langle J^2\rangle_J\right\}, \\
Q^B_{EE} &\equiv -\frac{1}{4}\int \mathcal{D}u \left\{\langle J^2\rangle^2_J\right\} + \frac{1}{4}\int \mathcal{D}u \left\{\langle J^2\rangle_J\right\} \int \mathcal{D}u \left\{\langle J^2\rangle_J\right\}, \\
Q^A_{EF} &\equiv \frac{1}{2}\int \mathcal{D}u \left\{\langle J^3\rangle_J \langle J\rangle_J\right\} - \frac{1}{2}\int \mathcal{D}u \left\{\langle J^2\rangle_J\right\} \int \mathcal{D}u \left\{\langle J\rangle^2_J\right\}, \\
Q^B_{EF} &\equiv \frac{1}{2}\int \mathcal{D}u \left\{\langle J^2\rangle_J \langle J\rangle^2_J\right\} - \frac{1}{2}\int \mathcal{D}u \left\{\langle J^2\rangle_J\right\} \int \mathcal{D}u \left\{\langle J\rangle^2_J\right\}, \\
Q^A_{FF} &\equiv -\int \mathcal{D}u \left\{\langle J^2\rangle^2_J\right\} + \int \mathcal{D}u \left\{\langle J\rangle^2_J\right\} \int \mathcal{D}u \left\{\langle J\rangle^2_J\right\}, \\
Q^B_{FF} &\equiv -\int \mathcal{D}u \left\{\langle J^2\rangle_J \langle J\rangle^2_J\right\} + \int \mathcal{D}u \left\{\langle J\rangle^2_J\right\} \int \mathcal{D}u \left\{\langle J\rangle^2_J\right\},
\end{aligned}$$

$$\begin{aligned}
Q^C_{FF} &\equiv -\int \mathcal{D}u\left\{\langle J\rangle_J^4\right\} + \int \mathcal{D}u\left\{\langle J\rangle_J^2\right\}\int \mathcal{D}u\left\{\langle J\rangle_J^2\right\}, \\
Q^A_{qq} &\equiv \alpha\int \mathcal{D}z\left\{\langle x^2\rangle_\Lambda^2\right\} - \int \mathcal{D}z\left\{\langle x\rangle_\Lambda^2\right\}\int \mathcal{D}z\left\{\langle x\rangle_\Lambda^2\right\}, \\
Q^B_{qq} &\equiv \alpha\int \mathcal{D}z\left\{\langle x^2\rangle_\Lambda\langle x\rangle_\Lambda^2\right\} - \int \mathcal{D}z\left\{\langle x\rangle_\Lambda^2\right\}\int \mathcal{D}z\left\{\langle x\rangle_\Lambda^2\right\}, \\
Q^C_{qq} &\equiv \alpha\int \mathcal{D}z\left\{\langle x\rangle_\Lambda^4\right\} - \int \mathcal{D}z\left\{\langle x\rangle_\Lambda^2\right\}\int \mathcal{D}z\left\{\langle x\rangle_\Lambda^2\right\}. \qquad (2.34)
\end{aligned}$$

The braces and angled brackets operators follow the definitions in equations (2.18) and (2.19). The first $\langle J\rangle_J$ and second moments $\langle J^2\rangle_J$ are given in equations (2.21), but we also need the moments

$$\begin{aligned}
\langle J^3\rangle_J &= \frac{1}{(\mathrm{i}E+\mathrm{i}F)^{\frac{3}{2}}}\left[3u\sqrt{\frac{F}{E+F}} + u^3\left(\frac{F}{E+F}\right)^{\frac{3}{2}}\right], \\
\langle J^4\rangle_J &= \frac{1}{(\mathrm{i}E+\mathrm{i}F)^2}\left[3 + 6u^2\frac{F}{E+F} + u^4\left(\frac{F}{E+F}\right)^2\right]. \qquad (2.35)
\end{aligned}$$

Similarly we take the first moment $\langle x\rangle_\Lambda$ from equation (2.24) and use the second moment

$$\langle x^2\rangle_\Lambda = \frac{1}{1-q}\left[1-\left(\frac{\hat{\lambda}(z)-z}{\sqrt{1-q}}\right)^2 + \frac{1}{(1-q)\beta\mathcal{G}''(\hat{\lambda}(z))}\right] \qquad (2.36)$$

where the last term contains a double differentiation of equation (2.27) with respect to the variable λ evaluated at the saddle-point. This term is a necessary second-order correction to the integration method by Laplace [Cop65].

Placing these moments into equations (2.34) we get the stability matrix elements of matrix (2.33) for fluctuations in E and F to be

$$(Q^A_{EE} - Q^B_{EE}) = -\frac{1}{4}\int \mathcal{D}u\left\{\langle J^4\rangle_J - \langle J^2\rangle_J^2\right\}$$

$$
\begin{aligned}
&= -\frac{1}{4(\mathrm{i}E+\mathrm{i}F)^2}\left[2+\frac{4F}{E+F}\right], \\
(Q^A_{EF}-Q^B_{EF}) &= \frac{1}{2}\int\mathcal{D}u\left\{\langle J^3\rangle_J\langle J\rangle_J-\langle J^2\rangle_J\langle J\rangle^2_J\right\} \\
&= \frac{1}{(\mathrm{i}E+\mathrm{i}F)^2}\left[\frac{F}{E+F}\right], \\
(Q^A_{FF}-4Q^B_{FF}+3Q^C_{FF}) &= -\int\mathcal{D}z\left\{\left[\langle J\rangle^2_J-\langle J^2\rangle_J\right]\left[3\langle J\rangle^2_J-\langle J^2\rangle_J\right]\right\} \\
&= \frac{1}{(\mathrm{i}E+\mathrm{i}F)^2}\left[\frac{2F}{E+F}-1\right], \qquad (2.37)
\end{aligned}
$$

while the element for fluctuations in q is

$$
\begin{aligned}
&(Q^A_{qq}-4Q^B_{qq}+3Q^C_{qq}) = \alpha\int\mathcal{D}u\left\{\left[\langle x\rangle^2_\Lambda-\langle x^2\rangle_\Lambda\right]\left[3\langle x\rangle^2_\Lambda-\langle x^2\rangle_\Lambda\right]\right\} \\
&= \frac{\alpha}{(1-q)^2}\int\mathcal{D}z\left\{\left[1+\frac{1}{(1-q)\beta\mathcal{G}''(\hat\lambda(z))}\right]\right. \\
&\qquad\left.\left[1+\frac{1}{(1-q)\beta\mathcal{G}''(\hat\lambda(z))}+2\left(\frac{\hat\lambda(z)-z}{\sqrt{1-q}}\right)^2\right]\right\},
\end{aligned}
$$

which can be re-written as

$$
\begin{aligned}
&= \frac{\alpha}{(1-q)^2}\int\mathcal{D}z\left\{\left[1-\frac{1}{1-\gamma^2\bar g''(\hat\lambda(z))/2}\right]\right. \\
&\qquad\left.\left[1-\frac{1}{1-\gamma^2\bar g''(\hat\lambda(z))/2}+2\left(\frac{\hat\lambda(z)-z}{\sqrt{1-q}}\right)^2\right]\right\}. \qquad (2.38)
\end{aligned}
$$

in terms of the γ parameter of equation (2.23) and the mean training function in equation (2.11). Clearly the integrand is zero when $\bar g''(\hat\lambda(z)) = 0$, which for the training function in question will only *not* occur at $\hat\lambda(z) = \kappa - \tau_{\mathrm{T}}$ and $\kappa + \tau_{\mathrm{T}}$, whereupon we are left with

$$
(Q^A_{qq}-4Q^B_{qq}+3Q^C_{qq}) = \frac{\alpha}{(1-q)^2}\int_{\hat\lambda(z)=\kappa\pm\tau_{\mathrm{T}}}\mathcal{D}z\left[1+2\left(\frac{\hat\lambda(z)-z}{\sqrt{1-q}}\right)^2\right]. \qquad (2.39)
$$

We can now see the effect of symmetric fluctuations on replica-symmetry by writing down the product of the eigenvalues as given by the determinant (2.33)

$$
\begin{aligned}
|\mathrm{Q}^{\mathrm{s}}| \;&=\; (Q^A_{qq} - 4Q^B_{qq} + 3Q^C_{qq})\left\{(Q^A_{EE} - Q^B_{EE})(Q^A_{FF} - 4Q^B_{FF} + 3Q^C_{FF})\right. \\
&\qquad \left. +2(Q^A_{EF} - Q^B_{EF})^2\right\} + (Q^A_{EE} - Q^B_{EE})
\end{aligned}
$$

which by piecing together equations (2.37) and (2.39) as well as using the saddle-point equations (2.25) solves to

$$
\begin{aligned}
|\mathrm{Q}^{\mathrm{s}}| \;&=\; \frac{\alpha}{2}(1-q)^2 \int_{\hat{\lambda}(z)=\kappa\pm\mathcal{T}_{\mathrm{T}}} \mathcal{D}z \left[1 + 2\left(\frac{\hat{\lambda}(z) - z}{\sqrt{1-q}}\right)^2\right] \\
&\qquad -\frac{1}{4}(1-q)^2 \left[2 + \frac{1}{1-q}\left(1 - (1-q)^2\right)\right], \qquad (2.40)
\end{aligned}
$$

but this is of order $(1-q)$ so vanishes in the optimal perceptron limit of $(q \to 1)$. However, we can say is that for q just under the optimal limit, the terms of order $(q-1)$ is a constant for all storage capacities, hence symmetric and weakly-asymmetric fluctuations do not affect the stability of the replica-symmetric ansatz.

2.9.2 Asymmetric Fluctuations

For the stability matrix in the case of asymmetric fluctuations we find orthogonality conditions mean that the contribution from fluctuations of the quantity $\{E_a\}$ vanishes, resulting in the simpler matrix (D.15). Excluding unimportant constant terms, the determinant of this matrix is

$$
|\mathrm{Q}^{\mathrm{a}}| = \begin{vmatrix} (Q^A_{FF} - 2Q^B_{FF} + Q^C_{FF}) & -\mathrm{i} \\ -\mathrm{i} & (Q^A_{qq} - 2Q^B_{qq} + Q^C_{qq}) \end{vmatrix} \qquad (2.41)
$$

where the quantities are as defined in equations (2.34). Assuming a line of reasoning very similar to that of the previous section, we can immediately write down

$$\begin{aligned}
(Q^A_{FF} - 2Q^B_{FF} + Q^C_{FF}) &= -\int \mathcal{D}u \left[\langle J\rangle^2_J - \langle J^2\rangle_J\right]^2, \\
&= -\frac{1}{(\mathrm{i}E + \mathrm{i}F)^2} \\
(Q^A_{qq} - 2Q^B_{qq} + Q^C_{qq}) &= \alpha \int \mathcal{D}z \left[\langle x\rangle^2_\Lambda - \langle x^2\rangle_\Lambda\right]^2 \\
&= \frac{\alpha}{(1-q)^2} \int \mathcal{D}z \left[1 - \frac{1}{1 - \gamma^2 \bar{g}''(\hat{\lambda}(z))/2}\right]^2 \\
&= \frac{\alpha}{(1-q)^2} \int_{\substack{\hat{\lambda}(z)=\kappa-\tau_{\mathrm{T}} \\ \hat{\lambda}(z)=\kappa+\tau_{\mathrm{T}}}} \mathcal{D}z. \qquad (2.42)
\end{aligned}$$

The complex nature of the elements in the determinant (2.41) indicate we are sited on a saddle-point[Cop65, Arf85]. The condition for whether the replica-symmetric solution is to be stable is deemed when the determinant of the Hessian is positive definite. This is easily verified by the positivity of the determinant ($|\mathrm{Q}^{\mathrm{a}}| = 1$) at the zero storage capacity limit of $\alpha_C = 0$, if we are prepared to accept the veracity of the replica symmetric ansatz with zero quenched disorder in the model.

Hence we can write the condition for a positive determinant as

$$\alpha_C \int_{\hat{\lambda}(z)=\kappa\pm\tau_{\mathrm{T}}} \mathcal{D}z < 1 \qquad (2.43)$$

where the necessary ranges for $\hat{\lambda}(z) = \kappa \pm \tau_{\mathrm{T}}$ is read off table (2.28) in the optimal perceptron limits of ($q \to 1$) and hence ($\alpha \to \alpha_C$), for each of the three parameter regimes identified. Unlike with symmetric and weakly-asymmetric fluctuations, the determinant for asymmetric fluctuations does not vanish in the optimal perceptron limit. Hence we must beware of the network wandering into a region where the replica symmetric solution is unstable, and hence obviously incorrect. This was done for all the results to be presented, and we can pre-empt them slightly by stating that they all lie within the replica-symmetrically stable regime.

2.10 Results for the Alignment Field Distribution

Equations (2.7) for the iterative dynamics and (2.32) for the storage error are both dependent on the alignment field distribution. This distribution gives an indication of the training function's effect as the following external field parameters are varied: the training field noise level f_T and field strength τ_T, the stability constant κ, and the storage error $\mathcal{F}$.

Armed with equation (2.28), the somewhat terse expression (2.26) for the alignment field distribution function can be expanded for the three regimes into

Regime 1 $\quad -\infty < 2\tau_T < \gamma\left[1-\sqrt{1-f_T}\right]$

$$\rho(\Lambda) = \frac{1}{\sqrt{2\pi}}\left\{\theta[(\kappa+\tau_T-\gamma)-\Lambda]+\theta[\Lambda-(\kappa+\tau_T)]\right\}\exp\left(-\frac{1}{2}\Lambda^2\right)$$
$$+\delta[\Lambda-(\kappa+\tau_T)]\left\{\overline{\mathrm{H}}[\kappa+\tau_T]-\overline{\mathrm{H}}[\kappa+\tau_T-\gamma]\right\},$$

Regime 2 $\quad \gamma\left[1-\sqrt{1-f_T}\right] < 2\tau_T < \gamma\sqrt{f_T}$

$$\rho(\Lambda) = \frac{1}{\sqrt{2\pi}}\left\{\theta[(\kappa-\tau_T-\gamma\sqrt{1-f_T})-\Lambda]+\theta[\Lambda-(\kappa+\tau_T)]\right\}\exp\left(-\frac{1}{2}\Lambda^2\right)$$
$$+\delta[\Lambda-(\kappa-\tau_T)]\left\{\overline{\mathrm{H}}\left[\kappa-\frac{\gamma^2 f_T}{4\tau_T}\right]-\overline{\mathrm{H}}[\kappa-\tau_T-\gamma\sqrt{1-f_T}]\right\}$$
$$+\delta[\Lambda-(\kappa+\tau_T)]\left\{\overline{\mathrm{H}}[\kappa+\tau_T]-\overline{\mathrm{H}}\left[\kappa-\frac{\gamma^2 f_T}{4\tau_T}\right]\right\},$$

Regime 3 $\quad \gamma\sqrt{f_T} < 2\tau_T < \infty$

$$\rho(\Lambda) = \frac{1}{\sqrt{2\pi}}\left\{\theta[(\kappa-\tau_T-\gamma\sqrt{1-f_T})-\Lambda]+\theta[\Lambda-(\kappa-\tau_T)]\right.$$

$$-\theta[\Lambda - (\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}] + \theta[\Lambda - (\kappa + \tau_{\mathrm{T}})]\Big\} \exp\left(-\frac{1}{2}\Lambda^2\right)$$

$$+\delta[\Lambda - (\kappa - \tau_{\mathrm{T}})]\left\{\overline{\mathrm{H}}\,[\kappa - \tau_{\mathrm{T}}] - \overline{\mathrm{H}}[\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}}]\right\}$$

$$+\delta[\Lambda - (\kappa + \tau_{\mathrm{T}})]\left\{\overline{\mathrm{H}}[\kappa + \tau_{\mathrm{T}}] - \overline{\mathrm{H}}\left[\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}\right]\right\}, \qquad (2.44)$$

where the error function $\overline{\mathrm{H}}[\cdots]$ is as defined by equation (A.4) in the mathematics appendix. These three distributions reflect how well the two terms in the training function (2.11) are satisfied, a 'correct' term $(1-f_{\mathrm{T}})\theta[\Lambda+\tau_{\mathrm{T}}-\kappa]$ and an 'incorrect' term $f_{\mathrm{T}}\theta[\Lambda - \tau_{\mathrm{T}} - \kappa]$. We can see the training noise level f_{T} weights between the two terms, while the actual difference in thresholds between the two step-functions is given by the noise strength $2\tau_{\mathrm{T}}$. The choice of regime is also affected by the fractional storage error, because demanding it to be low is achieved by setting the parameter γ to be large, which translates to demanding a strict adherence to the training function. The three plots 2.1–2.3 of the alignment distribution show regimes 1,2 and 3 respectively, for three increasing field strength at a fixed noise level. The same regimes can be qualitatively reproduced by keeping the field strength fixed and decreasing the noise level, with the main difference in the locations of the delta-peaks at $(\kappa - \tau_{\mathrm{T}})$ and/or $(\kappa + \tau_{\mathrm{T}})$.

Figure 2.1 shows the first regime is essentially the original Gardner training function distribution with a non-zero fractional storage error [AEHW90], bar a trivial rescaling in the stability requirement constant ($\kappa \longrightarrow \kappa + \tau_{\mathrm{T}}$). This occurs when the erroneous $f_{\mathrm{T}}\theta[\cdots]$ term in the training function of equation (2.11) dominates over the 'correct' $(1 - f_{\mathrm{T}})\theta[\cdots]$ part. This can happen at low storage errors $\mathcal{F}$, as strictly requiring the erroneous part to be trained automatically ensures adherence of the easier correct part. Indeed, for the strictest zero storage error case, the distribution is the same as the original Gardner regardless of the other parameters and no new behaviour is observed. With the storage error greater than zero, this first regime can still be visited by a low enough external field strength τ_{T} and/or a high field noise level f_{T}. The former because there is then a negligible difference between the two step-functions, and the latter because the training

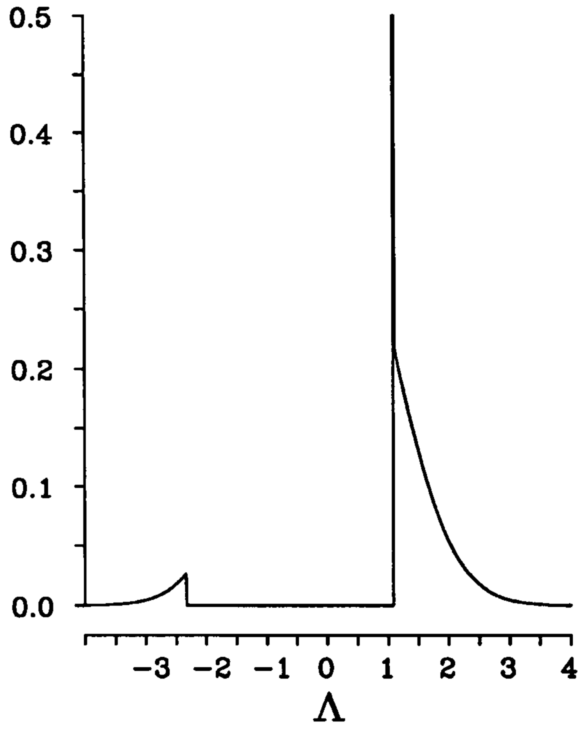

Figure 2.1: Alignment field distribution for a fixed storage error $\mathcal{F}$ =1% and minimum stability constant $\kappa = 1.0$. This is for the first regime in parameter space with a low training field strength $\tau_{\mathrm{T}} = 0.10$, and the noise level at $f_{\mathrm{T}} = 0.24$. The delta peak lies at $(\kappa + \tau_{\mathrm{T}})$. This same regime is qualitatively reproduced with a high noise level $f_{\mathrm{T}} = 0.50$ with a field strength $\tau_{\mathrm{T}} = 0.50$.

function becomes heavily weighted in favour of the erroneous part.

Novel behaviour appears in the second regime illustrated by figure 2.2. This is made by fixing the storage error to a non-zero value while raising the external field strength and/or lowering the noise level from the values in the first regime. This new regime has an additional delta peak at $(\kappa - \tau_{\mathrm{T}})$ and is the manifestation of alignment fields that satisfy the correct term in the training function, but not the tougher erroneous one, suitably aided by the weighting towards the former.

Further raising and/or lowering the external field parameters produces the final

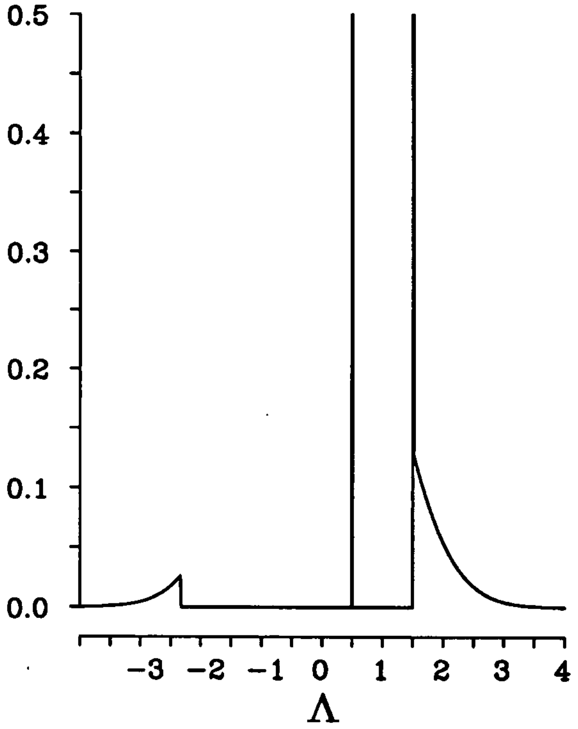

Figure 2.2: Alignment field distribution for the same storage error and stability constant as in the previous figure. This illustrates the second regime in parameter space with a medium training field strength $\tau_{\mathrm{T}} = 0.50$, and the noise level at $f_{\mathrm{T}} = 0.24$. There is now an additional delta peak at $(\kappa - \tau_{\mathrm{T}})$ in addition to the one at $(\kappa + \tau_{\mathrm{T}})$. The same regime is reproduced by a noise level $f_{\mathrm{T}} = 0.30$ at a field strength $\tau_{\mathrm{T}} = 0.50$.

third regime shown in figure 2.3. Here the correct term in the training function continues to grow in importance over the erroneous part, expanding its influence on the distribution at the expense of the delta peak at $(\kappa + \tau_{\mathrm{T}})$. As the external field strength further increases, this diminishing delta peak eventually disappears and the distribution returns to the Gardner case but with the remaining delta peak relocated to $(\kappa \rightarrow \kappa - \tau_{\mathrm{T}})$.

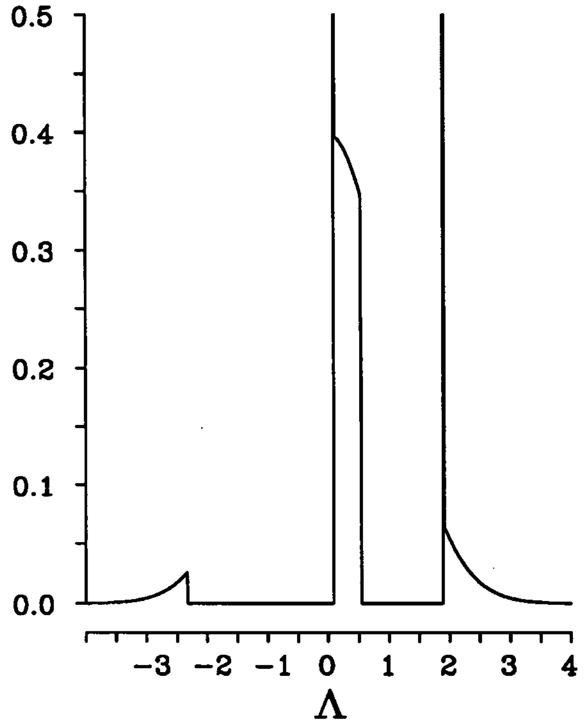

Figure 2.3: Alignment field distribution, again with the same storage error and stability constant as before. This shows the third regime with a high training field strength $\tau_{\mathrm{T}} = 0.90$, and the noise level at $f_{\mathrm{T}} = 0.24$. The same regime is reproduced by a low noise level $f_{\mathrm{T}} = 0.10$ at a field strength $\tau_{\mathrm{T}} = 0.50$.

2.11 Results for the Basins of Attraction

We shall now examine the network's dynamical behaviour by the iterative expression (2.7). The use of diluted connections allows this to be generalised to a time-iterative map of the overlap measure, so that the retrieval of an arbitrary pattern is given by

$$m_{t+1} = \int \mathrm{d}\Lambda \; \rho(\Lambda) \left\{ (1 - f_{\mathrm{R}}) \mathrm{erf} \left[\frac{m_t \Lambda + \tau_{\mathrm{R}}}{\sqrt{2(1 - m_t^2)}} \right] + f_{\mathrm{R}} \mathrm{erf} \left[\frac{m_t \Lambda - \tau_{\mathrm{R}}}{\sqrt{2(1 - m_t^2)}} \right] \right\} \quad (2.45)$$

in terms of the alignment field distribution function.. The additional parameters $(f_{\mathrm{R}}, \tau_{\mathrm{R}})$ are to do with the application of a persistent external field during retrieval, with f_{R} the mean fraction of erroneous bits, and τ_{R} the field strength. Taking the field strengths τ_{T} and τ_{R}, and the fractional storage error $\mathcal{F}$ all to zero restores the retrieval dynamics for the original Gardner model [KA88]. From this dynamical equation we can discover the network's attractor structure of retrieval. The results and analysis of this constitute the remainder of this chapter, beginning first with a discussion on what quantities we intend to measure.

2.11.1 Transitional Points of the Dynamics

The aim here is to discover how the attractor structure of the retrieval dynamics changes as the various network parameters are varied. For a given set of these parameters we start by numerically [PFTV88] finding the fixed-points of the iterative map equation (2.45) for the overlap measure m, where the stable solutions indicate attractor centres and unstable ones define the attractor boundaries. From these two values of the overlap the fidelity of the memory attractor to its training pattern and the size of its basin of attraction are revealed.. The network's iterative dynamics is described by equation (2.45), for an optimal storage capacity α_C given by equation (2.30) and a fractional storage error $\mathcal{F}$ by equation (2.32). The effect of the training and retrieval external fields will be discussed in detail later, but for now we shall focus on the effects of varying the storage capacity and storage error.

The storage capacity α_C is a particularly interesting quantity to examine since it gives a readily accessible indication of a network's performance. We shall refer to figure 2.4 which shows the basin of attraction for an increasing optimal storage capacity in the Gardner case of no external fields and zero error in the storage. The solid lines are attractor boundaries, and the attractors themselves are drawn with dashed lines. In figure 2.4 we can make out two phases in the storage-overlap

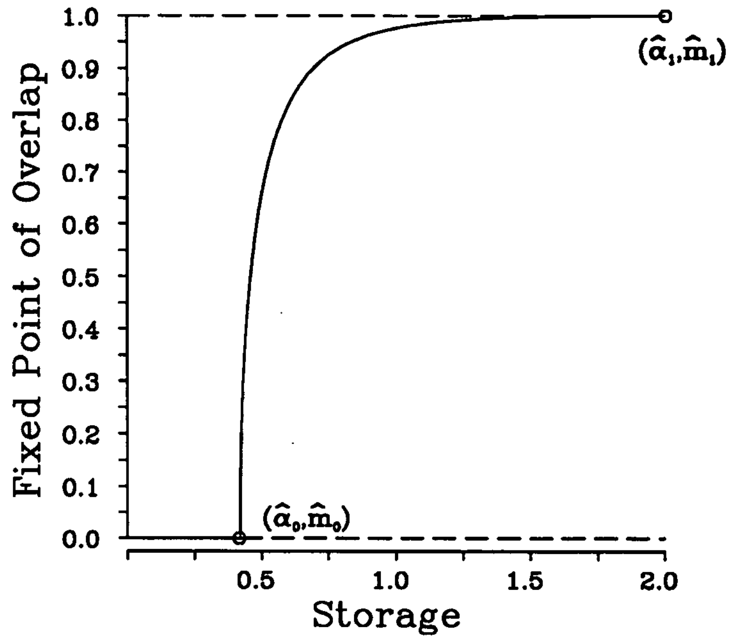

Figure 2.4: Fixed-point map showing the basin of attraction for the Gardner case of no external field and zero storage errors, for increasing optimal storage capacity. The attractor boundaries are drawn in solid lines (———), and the attractor centres by dashed lines (— — —). The circled points are two *transitional points* and mark out the two retrieval regions. From a storage capacity of $\alpha = 0$ to $\alpha = \hat{\alpha}_0$ we have the region of wide retrieval, and beyond that from $\alpha = \hat{\alpha}_0$ to $\alpha = \hat{\alpha}_1$ we have the region of narrow retrieval. Beyond $\alpha_C > \hat{\alpha}_1$ no dynamical retrieval of a pattern is possible.

(α_C, m) space. At low storage capacities the basin boundary has an overlap of essentially zero, implying retrieval of the pattern is guaranteed for any positive, infinitesimal initial overlap. As the number of patterns stored by the network is increased, a level of storage is reached whereby the attractors for the memory patterns can no longer avoid each other, and have to shrink their basin boundaries in response. This storage capacity signals the upper-bound for the region of *wide-retrieval*, beyond which a macroscopic initial overlap is necessary for retrieval.

A further increase in the storage capacity squeezes the basin of attraction until it and the attractor vanish altogether, making the dynamical retrieval of a pattern impossible. From the end of the region of wide-retrieval to this saturation point of the network is the area we call the region of *narrow-retrieval.*

For the cases to be examined, we wish to emphasize how these regions of wide and narrow-retrieval are affected by the external fields. Operationally, these are marked out by points in the storage-overlap space where stable and unstable fixed-points meet. Henceforth they will be referred to as *transitional points*[3] and are denoted by hats: $(\hat{\alpha}, \hat{m})$. Figure 2.4 illustrates where two such points exist for the original Gardner case. The transition from the wide-retrieval to the narrow-retrieval region is given by an optimal storage capacity $\alpha_C = \hat{\alpha}_0$ with an overlap $m = \hat{m}_0$, and the transition from narrow retrieval to zero retrieval is given by the point $(\hat{\alpha}_1, \hat{m}_1)$. For the zero storage error Gardner case, the wide and narrow-retrieval regions are bounded by $(\hat{\alpha}_0 = 0.42, \hat{m}_0 = 0.0)$ and $(\hat{\alpha}_1 = 2.0, \hat{m}_1 = 1.0)$ respectively [Gar89]. From zero storage to the end of the narrow-retrieval region there is a stable fixed-point at $m = 1$ for the memory attractor, but for the storage $\alpha > \hat{\alpha}_0$ there is another spurious attractor with overlap $m = 0$.

We can treat the fraction of erroneous bits per pattern stored ($\mathcal{F}$) as a parameter by solving equation (2.32). The effect of increasing this value is to decrease the quality of the memory attractor, and a corresponding narrowing of the narrow-retrieval region [AEHW90] by a decrease of the $\hat{\alpha}_1$ transitional point. This simple (and somewhat unproductive) response to further increases of the storage error has been verified in this model at the values of 0.1, 1, 5, and 10% of the bits per pattern, where no qualitative changes in the behaviour took place. Nonetheless, §2.10 has shown the necessity of keeping the storage error above zero for novel behaviour so in the results that follow this parameter has been fixed at the somewhat arbitrary value of $\mathcal{F} = 0.01$.

[3] Reference [YW91] called these 'critical-points', a name we now regard as too vague.

Finally this leaves the external field parameters to examine. The cases presented in the following subsections look at the effect on the transitional points with a training field ($f_{\mathrm{T}}, \tau_{\mathrm{T}}$) only, with a retrieval field ($f_{\mathrm{R}}, \tau_{\mathrm{R}}$) only, and with statistically equal training & retrieval fields ($f_{\mathrm{T}} = f_{\mathrm{R}}, \tau_{\mathrm{T}} = \tau_{\mathrm{R}}$). The chosen means of showing their effect is by looking at the resulting regions of wide and narrow retrieval, as given by plotting the transitional points' overlap (– – –) and storage (——).

2.11.2 Transitional Points with Training Field Only

Figures 2.5 and 2.6 show the transitional points as the training noise strength is increased, for two levels of training noise at $f_{\mathrm{T}} = 0.20$ and $f_{\mathrm{T}} = 0.24$ respectively.

The straightforward behaviour shown in figure 2.5 is typical for low noise levels (below $f_{\mathrm{T}}{\sim}0.05$ the network is essentially insensitive to the training field). Given a sufficient noise level the improved content addressability of the system becomes readily apparent as the region of wide-retrieval region bounded by the transitional point ($\hat{\alpha}_0, \hat{m}_0$) increases, peaking at $\hat{\alpha}_0 = 0.52$ for a field strength $\tau_{\mathrm{T}} = 0.52$. As the number of patterns that can be retrieved with a microscopic initial overlap has increased, the basin of attraction can therefore be said to have widened. Unfortunately this improvement is seemingly at the expense of the narrow-retrieval region, whose decrease reduces the number of patterns that can be stored without saturating the network.

Increasing the training noise above $f_{\mathrm{T}}{\sim}$ 0.21 complicates the transitional point plots considerably, as the typical example at $f_{\mathrm{T}} = 0.24$ in figure 2.6 shows.

For low field strengths up to $\tau_{\mathrm{T}}{\sim}$ 0.36 the wide and narrow-retrieval regions increase and decrease respectively, as with the example discussed above. Additional structures indicating new transitional points appear as the field strength is further increased, and are best understood by referring to their fixed-point diagrams. The

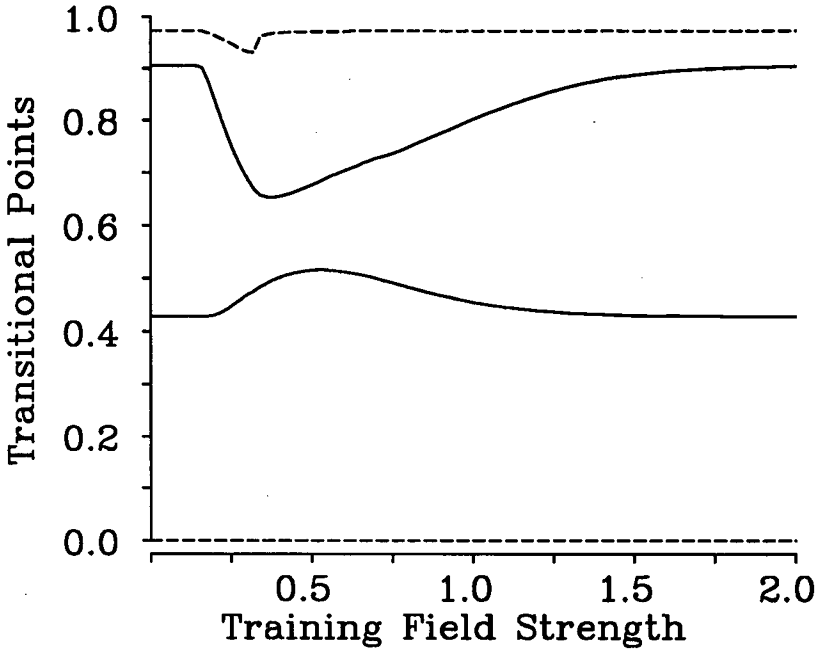

Figure 2.5: Plot of the transitional points with only the training field applied, the storage error is at 1% and the external field noise level is $f_{\mathrm{T}} = 0.20$. The plot is against an increasing field strength τ_{T} with the transitional points' overlaps (– – –) and storage (——) sharing the same ordinate. Here two transitional points are being tracked: the $(\hat{\alpha}_0, \hat{m}_0)$ line with zero overlap throughout and a storage capacity starting at $\alpha_C = 0.42$ which marks the upper-bound of wide-retrieval, and the $(\hat{\alpha}_1, \hat{m}_1)$ line starting with an overlap $m \simeq 0.98$ and storage $\alpha_C \simeq 0.90$ which marks the end of narrow-retrieval. The increase in the wide-retrieval region as indicated by $\hat{\alpha}_0$ shows that content addressability is improved, but the decrease in $\hat{\alpha}_1$ shows a corresponding decrease of the network's saturation limit, and with it the narrow retrieval region.

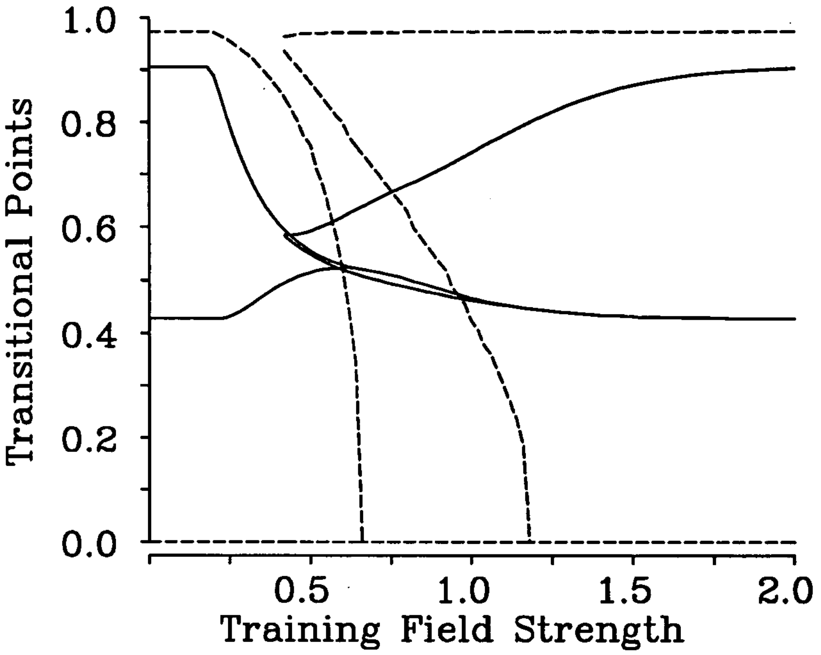

Figure 2.6: Plot of the transitional points with only the training field applied, the storage error is again at 1% and the noise level has increased from the previous plot to $f_{\mathrm{T}} = 0.24$. The meaning of the broken and solid lines remain as before, but this example appears to be far richer. However, this additional structure has a straightforward origin and is due entirely to the creation and later disappearance of an additional stable fixed-point, one of a lower quality than the memory attractor. The evolution of this can be more clearly seen by examining the fixed-point maps from which the transitional points are extracted, which explicitly show this additional attractor.

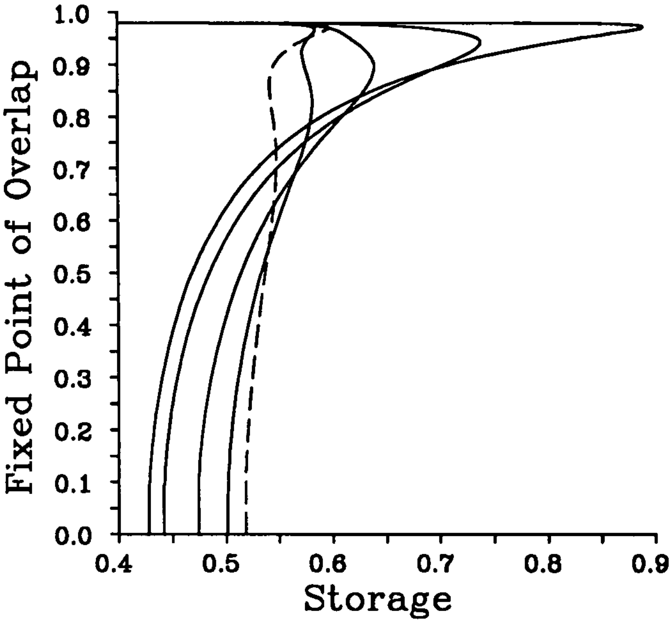

Figure 2.7: The first of three 'snapshots' of the the fixed-point retrieval maps from which the previous transitional points plot is made. We are showing here the fixed-point map's evolution by plotting the fixed-point of overlap against storage capacity for five increasing field strengths, with the largest field strength plot drawn in broken lines. Here the training noise level is $f_T = 0.24$ and the field strengths are τ_T =0.20, 0.28, 0.36, 0.44 and (in broken lines) 0.52. At low field strengths τ_T =0.20, 0.28 & 0.36 the wide and narrow retrieval regions increase and decrease respectively. As the field strength is increased to τ_T =0.44 and 0.52, a new attractor is created and with it two new transitional points with overlaps $m \simeq$ 0.90–0.98, one of which has now assumed the rôle of marking out the upper-limit of narrow retrieval.

important 'snapshots' for this example are given in figures 2.7–2.9, which plot the fixed-points against increasing storage levels, for increasing field strengths.

Figure 2.7 shows that as the field strength is increased the $(\hat{\alpha}_1, \hat{m}_1)$ transitional

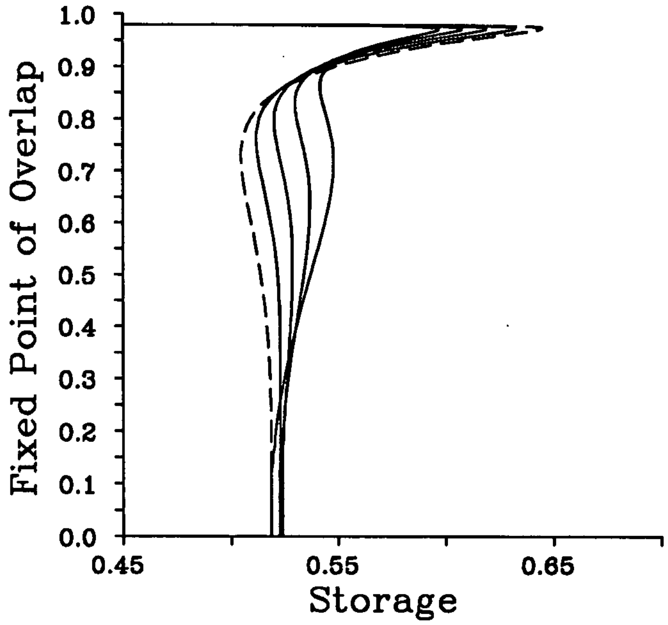

Figure 2.8: The second of the three 'snapshots' of the fixed-point retrieval maps giving the transitional points plot at noise level $f_{\mathrm{T}} = 0.24$. The field strength has now increased to τ_{T} =0.52, 0.56, 0.60, 0.64 & (in broken lines) 0.68. Here we see the original $(\hat{\alpha}_1, \hat{m}_1)$ transitional point merging into the $(\hat{\alpha}_0, \hat{m}_0)$ point at the abscissa at τ_{T} =0.60–0.64. Its rôle marking out the región of narrow retrieval has already been assumed by one of the transitional points associated with the new attractor seen created in the previous figure.

point marking the narrow-retrieval region starts to merge into the wide-retrieval point $(\hat{\alpha}_0, \hat{m}_0)$. Meanwhile two new transitional points appear with an overlap $m \sim 1$, indicating the formation of a stable attractor in addition to the memory pattern's, but of lower fidelity. In figure 2.8 the old $(\hat{\alpha}_1, \hat{m}_1)$ transitional point vanishes but its rôle marking out the region of wide-retrieval is quickly taken over by one of the two new transitional points. Finally, in figure 2.9 the other new transitional point coalesces into the wide-retrieval point, taking with it the extraneous attractor.

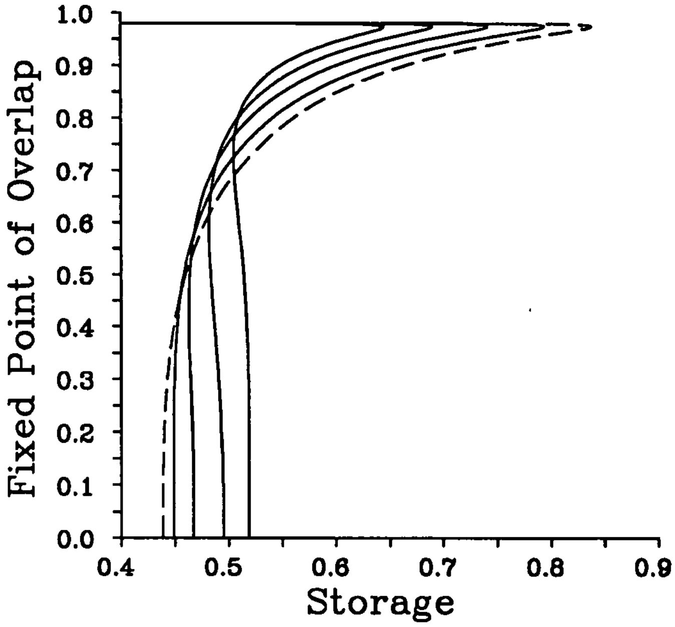

Figure 2.9: The last of the three fixed-point retrieval map snapshots, for the training noise level $f_{\rm T} = 0.24$ case. The training field strength here is further increased by $\tau_{\rm T}$ =0.68, 0.84, 1.00, 1.16 & (in broken lines) 1.32. We can see here the extraneous attractor's stable fixed-point disappearing, merging the final extraneous transitional point into the $(\hat{\alpha}_0, \hat{m}_0)$ point. In the limit of very large training field strengths we are left with the same fixed-point map as for the zero field case.

The extra structure brought along by the appearance of the extra attractor seems to have little practical consequence. Indeed, if anything it appears to degrade the region of wide-retrieval over some range of external field strengths, and hence may be indicative of the maximum training noise level to choose.

This structure is not, however, a result of instability in the replica-symmetric solution and cannot be dismissed on these grounds. The indications from equation (2.43) are that stability of the replica-symmetric solution is only violated beyond the network's limit of storage.

2.11.3 Transitional Points with Retrieval Field Only

As mentioned in §2.1 motivating this model, the effect of an external persistent field on the retrieval dynamics has recently been studied for the zero storage error [EBKS90], but it is still useful to consider the 1% error case for the sake of direct comparisons.

At a low retrieval noise level (figure 2.10) below $f_{\mathrm{R}} \sim 0.20$ the structure is straightforward with no new attractors appearing. The transitional points do differ from the training-field only case because the introduction of the retrieval field breaks the invariance of the dynamical equation (2.45) to ($m \rightarrow -m$) overlap flips, raising the wide retrieval transitional point's overlap $\hat{m}_0$ above zero. Consequently the stable zero-overlap attractor at $\alpha > \hat{\alpha}_0$ is now replaced with a macroscopic one induced by the external field. Meantime the wide-retrieval region $\hat{\alpha}_0$ increases, then eventually coalesces with the falling narrow-retrieval point $\hat{\alpha}_1$. Beyond this the retrieval map has no transitional points, and instead there is just one attractor of steadily decreasing quality with increasing storage, signalling the dominance of the external field in the network's dynamics.

Retrieving with the slightly noisier field shown by figure 2.11 retains the transi-

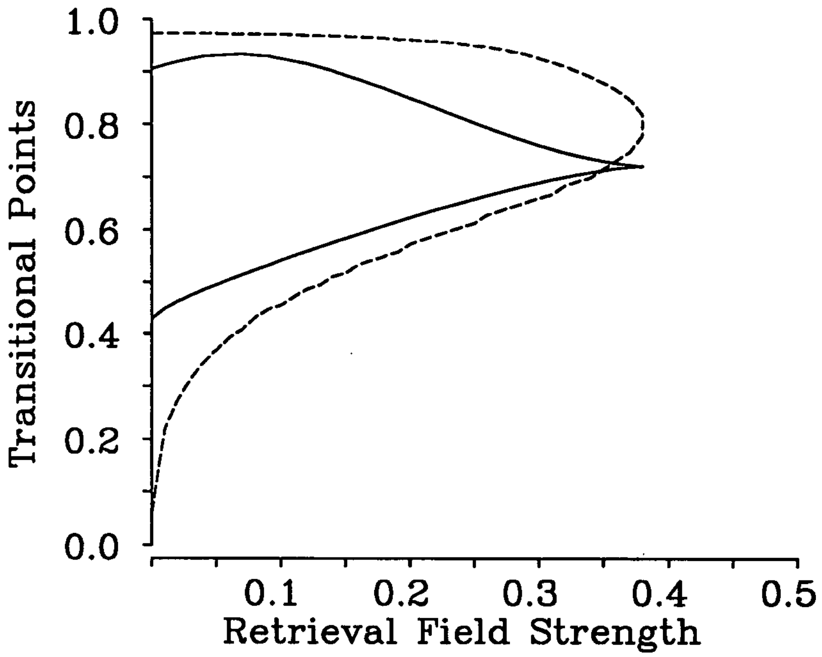

Figure 2.10: Plot of the transitional points using an external field only during retrieval. This one again has the storage error at 1% and the field noise level at $f_R = 0.20$. As with the two plots where only a training field is used, we are plotting on the same axis the transitional points' overlaps (— — —) and storage (———). However, unlike those plots the transitional points of wide and narrow-retrieval eventually come together at $\tau_R = 0.38$ beyond which the network has only a single attractor whose quality decreases with increasing storage. When this happens the lack of any transitional points prevent us from making a demarcation between the wide and narrow retrieval regions. However, examination of the fixed-point maps has shown only a worsening in the basin of attraction with increasing retrieval field strengths. This is corroborated by the following plot for a higher retrieval noise level where the transitional points do survive and we can explicitly see the wide and narrow retrieval regions decreasing.

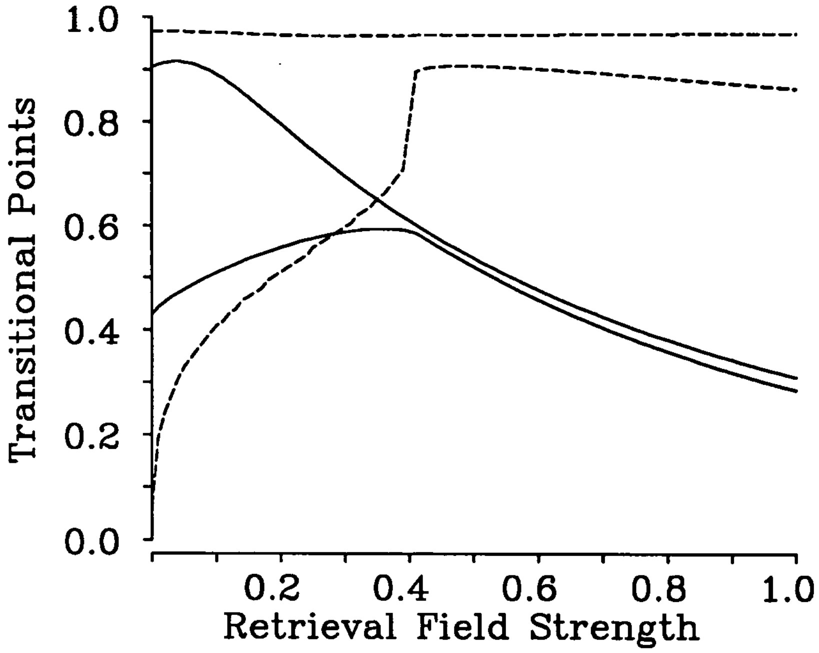

Figure 2.11: The second transitional points plot using an external field only during retrieval. Again the storage error is fixed at 1% but the field noise level is now increased to $f_{\mathrm{R}} = 0.30$. The transitional points begin much like its lower noise level sibling, but do not come together at a given field strength. Instead the region of wide retrieval reaches a maximum at $\tau_{\mathrm{R}} = 0.38$ before it follows the narrow retrieval region on a seemingly steady decline. This detrimental behaviour simply reflects the network's dynamics being dominated by polarisation by the high retrieval field.

tional points throughout the covered field strength range, and consequently it is possible to see how the regions of wide and narrow-retrieval decrease for large field strengths. In comparison to the example of figure 2.10 this is able to show the eventual polarisation of the spin-sites to the external retrieval field.

Lastly, both plots here are within the regime where replica-symmetry is stable.

2.11.4 Transitional Points for Equal Training and Retrieval Fields

Since the motivation for this work is the training of the network with ensembles of external noisy patterns, it seems reasonable to expect the best performance with statistically identical training and retrieval fields. The plots of such an assumption are presented in figures 2.12 and 2.13, for two levels of external field noise. The most immediately apparent observation is their qualitative similarity to the pure retrieval field cases. The quantitative differences are summarised in table 2.1 which shows the maximum wide-retrieval storages $\max\{\hat{\alpha}_0\}$ and the corresponding 'best' field strength, amongst the three cases discussed above for three field noise levels. Nonetheless the improvements are such it may be implied that the retrieval field has a disproportionate effect on the dynamics, and hence an 'optimal' training and retrieval fields combination may involve weakening the latter. This hypothesis is supported by the disparate best external field strengths to use.

Once again the results presented are replica-symmetrically stable.

2.12 Concluding Remarks

We have seen in this chapter a detailed example of a Gardner phase-space of interactions calculation. This model was an extension motivated by the observations that including external fields and noise with the training procedure can improve content addressability.

The first equation calculated was an expression giving an iterative map which describes the dynamics of the neurons' zero-temperature (*i.e.*, non-stochastic and deterministic) update. This map is exact for the first time step from a random start, but can be elevated to all time steps for a network with a highly diluted

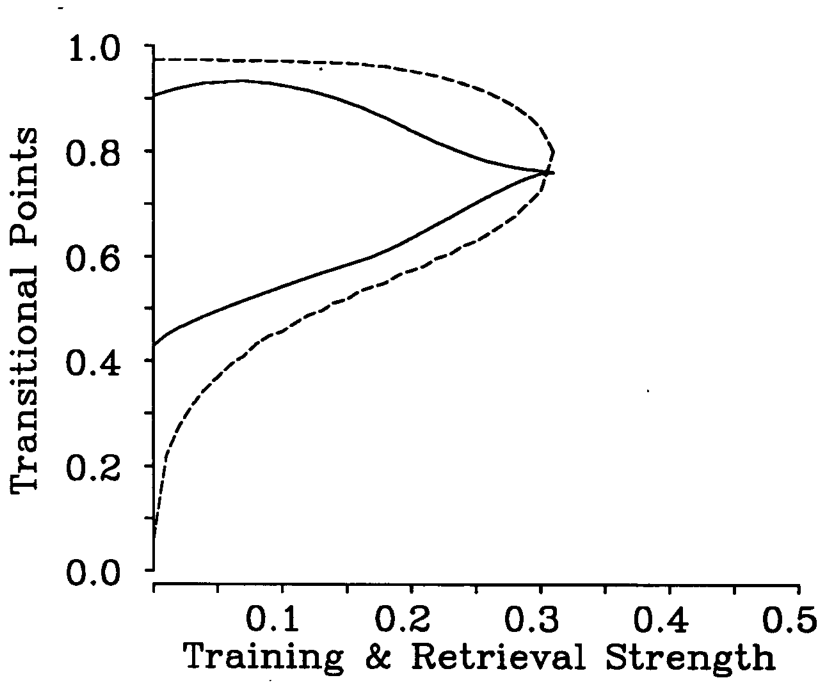

Figure 2.12: This plot shows the transitional points when a statistically equivalent external field is used during training and retrieval. The storage error is at 1%, and the noise levels at $f_{\mathrm{T/R}}$ *i.e.*, $(f_{\mathrm{T}} = f_{\mathrm{R}}) = 0.20$. Qualitatively the behaviour is identical to having only a retrieval field at a low field noise level, with the same merging of the two transitional points.

---

connectivity C much less than the magnitude of the number of sites N.

We then invoked self-averaging for the large system size and wrote the iterative map in terms of a distribution function of the alignment defined in §1.2. This distribution function was then calculated by doing an annealed-optimised search over the phase-space of connections, picking out the solution which satisfied the training function we had. The disorder from the memory patterns we wish to store and the noisiness of the external fields were quenched averaged by use of the replica method, closely followed by taking the replica symmetric ansatz.

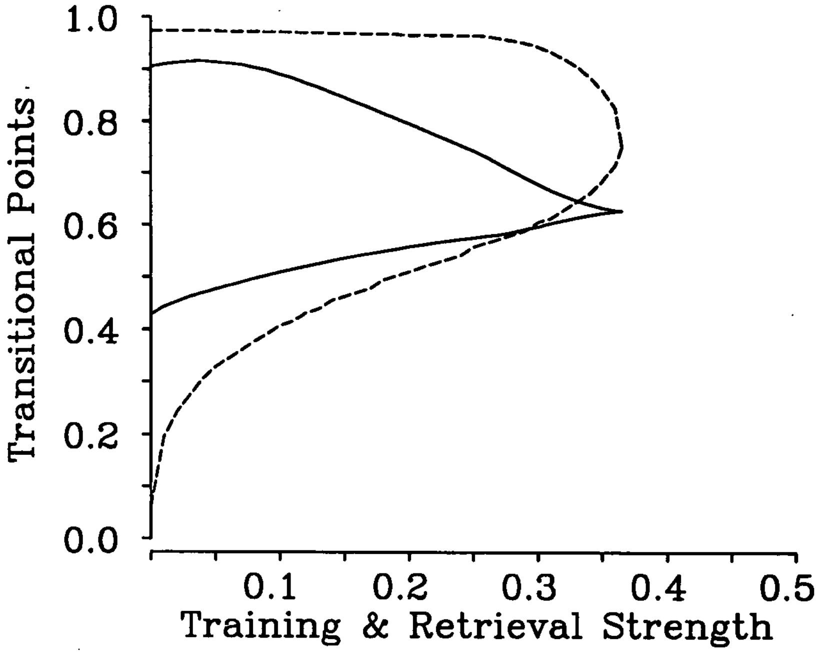

Figure 2.13: The second plot of the transitional points when a statistically equivalent external field is used during training and retrieval. The storage error is at 1%, and the noise levels here are at $f_{T/R} = 0.30$. The behavior is as in the previous figure.

We then found the parameters which control the external field used during the training phase gave three regimes of behaviour, two of which are novel and differ from the familiar Gardner result. These gave three sets of equations for the storage capacity (2.30) α_C, the fraction of storage error (2.32) $\mathcal{F}$, and —via the distribution function (2.44)— the dynamical equation (2.45). The validity of these equations to small local replica-symmetry breaking fluctuations was ensured when condition (2.43) was satisfied.

For the results, we first examined the effects of the three regimes on the alignment field distribution function. We then introduced the idea of a *transitional point* which allowed us to qualitatively measure the basins of attraction as the various

	Mean External Field Noise Level					
	0.20		0.24		0.30	
	$\max\{\hat{\alpha}_0\}$	$\text{best}\{\tau\}$	$\max\{\hat{\alpha}_0\}$	$\text{best}\{\tau\}$	$\max\{\hat{\alpha}_0\}$	$\text{best}\{\tau\}$
Training field	0.52	0.52	0.52	0.60	0.50	0.50
Retrieval field	0.72	0.38	0.66	0.38	0.60	0.38
Training = Retrieval field	0.76	0.30	0.71	0.32	0.63	0.36

Table 2.1: Comparison of the three cases examined at three levels of external field noise. The external field has been applied during training only, during retrieval only, and during both training and retrieval phases. The entries refer to the largest wide-retrieval region obtained, together with the associated field strengths to obtain it. The former is measured by the maximum storage capacity where retrieval of a memory pattern with a microscopic initial overlap is possible, $\hat{\alpha}_0$. By comparison, the region of wide-retrieval in the original Gardner model is bounded by a maximum storage capacity of $\hat{\alpha}_0 = 0.42$.

external field parameters were varied. These points showed that the maximum storage capacity in which a pattern can still be retrieved with a microscopic initial overlap, the region of *wide retrieval*, was increased by the use of external fields. Specifically, improvements in the basins of attraction were looked for in three cases: training field only, retrieval field only, and statistically equal training and retrieval fields. In all three cases the region of wide-retrieval were improved above the original Gardner model's $\hat{\alpha}_0 = 0.42$, with the equal field case marginally highest; *e.g.*, $\max\{\hat{\alpha}_0\} = 0.76$ for training and retrieval fields at strength $f_{\mathrm{T/R}} = 0.30$ and noise level $f_{\mathrm{T/R}} = 0.20$. However, this slight improvement over the corresponding retrieval-field only case ($\max\{\hat{\alpha}_0\} = 0.72$), and the differing value for the best field strength ($\tau_{\mathrm{R}} = 0.38$), perhaps suggests the retrieval field was dominating the dynamics and that a simple equality was not the optimal relationship between the training and retrieval field parameters.

The stability of the replica-symmetric ansatz to small symmetry breaking fluctuations was calculated and it appears to have been respected in all the results presented.

Before we close this chapter we ought to mention that reference [RSW91] has made a survey of retrieval behaviour, comparing the effects of discrete and Gaussian noises in the retrieval field. The authors have pointed out that discrete noise probabilities of the form described by equation (2.2) can be responsible for features in the fixed-point retrieval maps which disappear when using the continuous Gaussian noise. This does not invalidate the results presented here, but it should temper one's enthusiasm for their potential universality. The obvious remedy is to repeat the calculations and numerical analyses in this chapter using Gaussian noise for the retrieval field, and this may be pursued in the future.

Chapter 3

Storage Properties of Sign-Constrained Neural Network Models

3.1 Dual Distribution Functions

The phase-space of interactions is a tremendously powerful technique for elucidating the properties of neural network models with optimal connections. Chapter 2 illustrated this with the specific problem of calculating the attractors in the retrieval dynamics of such a network. We saw that the dynamics is principally determined by the distribution function of the alignment field. From the point of view of understanding the network this distribution is an interesting quantity to calculate in its own right, and we could have similarly calculated the distribution over the synaptic efficacy. Both these distributions allow us insights into the nature of the optimal network without actually having to know the exact value of each individual connection.

In this chapter we shall further develop the phase-space technique and calculate the *dual distribution functions* as introduced by Wong [Won90]. These objects will be used to compare the properties of two different optimal networks given the same sets of patterns to train, with a view of seeing how they cope as the differences between them are varied. These differences can be in the training function as in reference [WRS91]), or they may be in their physical architecture. The example considered here belongs to the latter category, namely in the way the networks' connection weights are constrained to a given sign. This idea of constraining the weights to be a specified sign is our attempt to model the biological observation known as *Dale's rule*, but its consequences here appear to lead to an embarrassing paradox. We shall now spend some time detailing the model and how it leads to this problem.

3.1.1 Dale's Rule and the Sign-Constraint

Dale's rule is an empirical statement which claims that the synapses into a biological neuron are either all excitatory or all inhibitory [Ecc64]. The way this observation is modelled in a neural network model is by restricting the synaptic efficacies into a neuron to be all positive for excitatory connections, or be all negative for inhibitory connections. Previous works have examined the retrieval dynamics [CW90], calculated the storage capacity by geometric arguments [CR91], and given a convergent perceptron-like learning algorithm for generating such optimal synapses [AWC89b].

However it is reference [AWC89a] which motivates this chapter's work. This considers how a sign-constrained network is able to optimally store random, uncorrelated patterns. In particular it hints at a possible paradox whereby the network is seemingly able to store patterns which ought to contradict its constraints.

We shall now detail the model we wish to study, using much of the notations given in §1.2. We require P patterns $\boldsymbol{\xi}^{\mu}, \mu = 1 \ldots P$ to be stored, which is realised when

$$\xi_i^{\mu} \sum_{j \neq i}^{C} \frac{J^{ij}}{|\boldsymbol{J}^i|} \xi_j^{\mu} > \kappa$$

that is, the alignment field is larger than some positive constant κ for all $\mu = 1 \ldots P$ patterns and for all sites $i = 1 \ldots N$. Using the justifications given in §1.2.1, we can simplify the notation by considering a single perceptron and re-writing the patterns as $(\xi_i^\mu \xi_j^\mu \rightarrow \xi_j^\mu)$. This perceptron has C continuous weights $J^j, j = 1 \ldots C$ whose phase space is constrained by equation (1.12) to lie on a unit C-dimensional semi-hypersphere. Together these three items simplify the stability condition to

$$\Lambda^\mu \equiv \sum_{j \neq i}^{C} \frac{J^j}{\sqrt{C}} \xi_j^\mu > \kappa \tag{3.1}$$

where we have taken the liberty of denoting the alignment field by Λ^μ. The final constraint to consider is for enforcing the signs of the interactions. This is realised by introducing a C-dimensional sign vector $\{\mathrm{g}^j\}$ with the possible binary components $\mathrm{g}^j = \pm 1$ such that we require

$$\mathrm{g}^j J^j > 0 \tag{3.2}$$

for all the $j = 1 \ldots C$ weights into the perceptron to be satisfied. This sign vector $\{\mathrm{g}^j\}$ allow us to specify whether each weight should be excitatory or inhibitory, with Dale's rule occupying the two extremes of $\mathrm{g}^j = 1$ or $\mathrm{g}^j = -1$ for all j.

The paradox mentioned above concerns what happens when we flip each component of the sign vector $\{\mathrm{g}^j\}$. Upon doing this, constraint (3.2) will demand a corresponding change in the sign of the weights $(J^j \rightarrow -J^j)$, but this will then violate the stability condition of equation (3.1). Hence it seems two networks with opposing signs in weights are not able to store the same set of memory patterns. Of course, the stability condition can be respected again if we also flip the patterns $(\{\boldsymbol{\xi}^\mu\} \rightarrow \{-\boldsymbol{\xi}^\mu\})$, but this is just telling us a network with opposite weights $\{-J^j\}$ can *only* store the flipped set of patterns.

We arrive at the paradox when we perform the quenched average on the uncorrelated and unbiased memory patterns, *i.e.,* completely random patterns. We can quickly see this calculation is independent of whether we average the set of patterns $\{\boldsymbol{\xi}^\mu\}$ or its opposite $\{-\boldsymbol{\xi}^\mu\}$ [AWC89a], which suggests the opposite network *can* meet the stability requirement without explicitly requiring the patterns to be flipped. We can turn this argument around and say that in effect a sign-constrained network is seemingly able to store a set of patterns $\{\boldsymbol{\xi}^\mu\}$ and its opposite $\{-\boldsymbol{\xi}^\mu\}$, despite the *prima facae* conclusions from the constraints (3.1) and (3.2).

The rest of this chapter is devoted to the task of resolving this paradox. To facilitate this we shall calculate the dual distributions in weight and alignment field distributions, for two networks under the same quenched memory patterns. The two networks have different sign constraints and in effect we are trying to ascertain how the two networks are coping in storing the same patterns. The following section will mathematically write out the calculation, and in the process introduce the notations to be used. Before the presentation of the results there will be a discussion on the stability of the replica-symmetric ansatz, a task necessary to lend justification to the mathematical validity of the calculations.

3.2 Calculating the Dual Distributions

We shall be investigating the storage properties of two networks with differing sign-constraints, by examining the dual distributions of their alignment fields and synaptic efficacy. That is, we shall be calculating the functions $\rho(\Lambda, \tilde{\Lambda})$ and $\rho(w, \tilde{w})$, where we have differentiated between the networks by the tilde symbols. The two networks are required to obey the stability condition (3.1) as well as the sign-constraint (3.2). The latter allows us to vary the difference between the two networks by defining an overlap measure between the two sets of constraints $\{\mathrm{g}^j\}$

and $\{\tilde{\mathrm{g}}^j\}$,

$$m_s \equiv \frac{1}{C}\sum_{j=1}^{C} \mathrm{g}^j \tilde{\mathrm{g}}^j \tag{3.3}$$

and hence for C connections into the neuron, a fraction $(1+m_s)/2$ have the same signs, whilst $(1-m_s)/2$ of them have opposing signs. m_s is now a parameter we can vary, such that $m_s = 1$ gives the two networks identical constraints, and conversely $m_s = -1$ forces them to be antisymmetric to each other. The latter extreme corresponds to the paradox discussed in §3.1.1 above. We shall now detail the calculation for the dual-distribution function of the alignment field.

3.3 The Alignment Field Dual Distribution

We shall take a quenched average over the memory patterns of the dual distribution function, as given by

$$\begin{aligned}
\rho(\Lambda,\tilde{\Lambda}) \;=\; & \left\langle\!\!\left\langle \int \prod_{j}^{C} \left[\theta[\mathrm{g}^j J^j]\mathrm{d}J^j \theta[\tilde{\mathrm{g}}^j \tilde{J}^j]\mathrm{d}\tilde{J}^j\right] \delta[\Lambda-\Lambda^1]\delta[\tilde{\Lambda}-\tilde{\Lambda}^1] \right.\right. \\
& \times \delta[\sum_{j}^{C}(J^j)^2 - C]\delta[\sum_{j}^{C}(\tilde{J}^j)^2 - C]\exp\left(\beta\sum_{\mu}^{P}\left[g(\Lambda^\mu)+g(\tilde{\Lambda}^\mu)\right]\right) \\
& \times \left[\int \prod_{j}^{C}\left[\overset{\theta[\mathrm{g}^j J^j]}{\wedge}\mathrm{d}J^j \overset{\theta[\tilde{\mathrm{g}}^j \tilde{J}^j]}{\wedge}\mathrm{d}\tilde{J}^j\right] \delta[\sum_{j}^{C}(J^j)^2 - C]\delta[\sum_{j}^{C}(\tilde{J}^j)^2 - C]\right. \\
& \left.\left.\left. \times \exp\left(\beta\sum_{\mu}^{P}\left[g(\Lambda^\mu)+g(\tilde{\Lambda}^\mu)\right]\right)\right]^{-1} \right\rangle\!\!\right\rangle_{\xi}
\end{aligned} \tag{3.4}$$

where the first two delta functions sample the alignment field for an arbitrary pattern $\boldsymbol{\xi}^{\mu=1}$. The double-angled brackets denote the two networks are quenched averaged with the same $\mu = 1\ldots P$ set of patterns. Apart from the differing sign-constraints the networks are identical, although this formalism has the potential to allow other differences to be considered, in for example the spherical constraint and the training function $g(\Lambda^\mu)$.

Averaging Over the Memory Patterns

By writing equation (3.4) in terms of $a = 1 \ldots n$ replicas over the connections we can raise the denominator and do the quenched average over the patterns. *i.e.*,

$$\begin{aligned}
\rho(\Lambda, \tilde{\Lambda}) \;=\; & \lim_{n \to 0} \int \prod_{a,\mu}^{n,P} \left[\frac{\mathrm{d}\lambda_a^\mu \mathrm{d}x_a^\mu}{2\pi} \frac{\mathrm{d}\tilde{\lambda}_a^\mu \mathrm{d}\tilde{x}_a^\mu}{2\pi} \right] \int \prod_{a,j}^{n,C} \left[\theta[\mathrm{g}^j J_a^j] \mathrm{d}J_a^j \theta[\tilde{\mathrm{g}}^j \tilde{J}_a^j] \mathrm{d}\tilde{J}_a^j \right] \\
& \delta[\Lambda - \lambda_1^1] \delta[\tilde{\Lambda} - \tilde{\lambda}_1^1] \prod_a^n \left[\delta[\sum_j^C (J_a^j)^2 - C] \delta[\sum_j^C (\tilde{J}_a^j)^2 - C] \right] \\
& \times \prod_{a,\mu}^{n,P} \exp\left(\mathrm{i} x_a^\mu \lambda_a^\mu + \mathrm{i} \tilde{x}_a^\mu \tilde{\lambda}_a^\mu + \beta g(\lambda_a^\mu) + \beta g(\tilde{\lambda}_a^\mu) \right) \\
& \times \left\langle\!\!\left\langle \prod_\mu^P \exp\left(-\mathrm{i} \sum_a^n \left[x_a^\mu \Lambda_a^\mu + \tilde{x}_a^\mu \tilde{\Lambda}_a^\mu \right] \right) \right\rangle\!\!\right\rangle_\xi . \qquad (3.5)
\end{aligned}$$

Using the definition of the alignment field given by equation (3.1), the last term in (3.5) with the dependence on the memory patterns can be averaged. In the large connectivity limit this can be written as an exponential function via

$$\begin{aligned}
& \left\langle\!\!\left\langle \prod_\mu^P \exp\left(-\mathrm{i} \sum_a^n \left[x_a^\mu \Lambda_a^\mu + \tilde{x}_a^\mu \tilde{\Lambda}_a^\mu \right] \right) \right\rangle\!\!\right\rangle_\xi \\
&= \prod_\mu^P \exp\left(-\frac{1}{2C} \sum_j^C \left[\sum_a^n \left(x_a^\mu J_a^j + \tilde{x}_a^\mu \tilde{J}_a^j \right) \right]^2 \right) \\
&= \prod_\mu^P \exp\left(-\frac{1}{2C} \sum_{a,j}^{n,C} \left[\left(x_a^\mu J_a^j \right)^2 + \left(\tilde{x}_a^\mu \tilde{J}_a^j \right)^2 \right] \right. \\
&\qquad\qquad - \frac{1}{C} \sum_j \sum_{a<b} \left[x_a^\mu J_a^j J_b^j x_b^\mu + \tilde{x}_a^\mu \tilde{J}_a^j \tilde{J}_b^j \tilde{x}_b^\mu \right] \\
&\qquad\qquad \left. - \frac{1}{2C} \sum_j \sum_{a,b} \left[x_a^\mu J_a^j \tilde{J}_b^j \tilde{x}_b^\mu + \tilde{x}_a^\mu \tilde{J}_a^j J_b^j x_b^\mu \right] \right)
\end{aligned}$$

which is dealt with by introducing the correlation measures between the replicas

$$\begin{aligned} q_{ab} \equiv \frac{1}{C}\sum_{j}^{C} J_a^j J_b^j, \quad \tilde{q}_{ab} &\equiv \frac{1}{C}\sum_{j}^{C} \tilde{J}_a^j \tilde{J}_b^j, \qquad \forall a, b\,(a<b), \\ r_{ab} &\equiv \frac{1}{C}\sum_{j}^{C} J_a^j \tilde{J}_b^j, \qquad \forall a, b. \end{aligned} \tag{3.6}$$

The quantities q_{ab} and $\tilde{q}_{ab}$ measure the overlap between different replicas within the same network, whilst r_{ab} examines the correlation between the two networks' synapses. By inserting the definitions (3.6) and writing the spherical constraints in integral form, we obtain for the dual distribution of alignment fields

$$\begin{aligned} \rho(\Lambda, \tilde{\Lambda}) \;=\; & \lim_{\substack{n\to 0 \\ C\to\infty}} \int \prod_{a}^{n} \left[\frac{\mathrm{d}E_a}{4\pi}\frac{\mathrm{d}\tilde{E}_a}{4\pi}\right] \prod_{a<b} \left[\frac{\mathrm{d}F_{ab}\mathrm{d}q_{ab}}{2\pi/C}\frac{\mathrm{d}\tilde{F}_{ab}\mathrm{d}\tilde{q}_{ab}}{2\pi/C}\right] \prod_{a,b} \left[\frac{\mathrm{d}K_{ab}\mathrm{d}r_{ab}}{2\pi/C}\right] \\ & \exp\ C\Big\{G_J(\{E_a, \tilde{E}_a, F_{ab}, \tilde{F}_{ab}, K_{ab}\}) + \alpha G_\Lambda(\{q_{ab}, \tilde{q}_{ab}, r_{ab}\}) \\ & \qquad + G_0(\{E_a, \tilde{E}_a, F_{ab}, \tilde{F}_{ab}, q_{ab}, \tilde{q}_{ab}, K_{ab}, r_{ab}\})\Big\} \\ & \times \int \prod_{a}^{n} \left[\frac{\mathrm{d}\lambda_a^1 \mathrm{d}x_a^1}{2\pi}\frac{\mathrm{d}\tilde{\lambda}_a^1 \mathrm{d}\tilde{x}_a^1}{2\pi}\right] \delta[\Lambda - \lambda_1^1]\delta[\tilde{\Lambda} - \tilde{\lambda}_1^1] \\ & \times \exp\Bigg\{-\frac{1}{2}\sum_{a}^{n} \left[(x_a^1)^2 + (\tilde{x}_a^1)^2\right] - \sum_{a<b}\left[x_a^1 q_{ab} x_b^1 + \tilde{x}_a^1 \tilde{q}_{ab} \tilde{x}_b^1\right] \\ & \qquad - \sum_{a,b} x_a^1 r_{ab} \tilde{x}_b^1 + \sum_{a}\left[\mathrm{i}x_a^1\lambda_a^1 + \mathrm{i}\tilde{x}_a^1\tilde{\lambda}_a^1 + \beta g(\lambda_a^1) + \beta g(\tilde{\lambda}_a^1)\right]\Bigg\} \end{aligned} \tag{3.7}$$

which has been explicitly written in a form to be amenable to integration by the method of steepest descent. The functions inside the exponential are

$$\begin{aligned} & G_J(\{E_a, \tilde{E}_a, F_{ab}, \tilde{F}_{ab}, K_{ab}\}) \equiv \\ & \frac{1}{2}(1+m_s)\ln \Bigg[\int_0^\infty\!\!\int_0^\infty \prod_{a}^{n} \mathrm{d}J_a \mathrm{d}\tilde{J}_a \exp\Bigg(-\frac{\mathrm{i}}{2}\sum_{a}^{n}\left[E_a J_a^2 + \tilde{E}_a \tilde{J}_a^2\right] \end{aligned}$$

$$
\begin{aligned}
&\left. +\mathrm{i}\sum_{a<b}\left[J_aF_{ab}J_b+\tilde{J}_a\tilde{F}_{ab}\tilde{J}_b\right]+\mathrm{i}\sum_{a,b}J_aK_{ab}\tilde{J}_b\right)\Bigg] \\
&+\frac{1}{2}(1-m_s)\ln\left[\int_0^{\infty}\int_{-\infty}^{0}\prod_a^n \mathrm{d}J_a\mathrm{d}\tilde{J}_a\exp\left(-\frac{\mathrm{i}}{2}\sum_a^n\left[E_aJ_a^2+\tilde{E}_a\tilde{J}_a^2\right]\right.\right. \\
&\left.\left. +\mathrm{i}\sum_{a<b}\left[J_aF_{ab}J_b+\tilde{J}_a\tilde{F}_{ab}\tilde{J}_b\right]+\mathrm{i}\sum_{a,b}J_aK_{ab}\tilde{J}_b\right)\right], \\
&G_\Lambda(\{q_{ab},\tilde{q}_{ab},r_{ab}\})\equiv \\
&\quad \ln\int\left[\frac{\mathrm{d}\lambda_a\mathrm{d}x_a}{2\pi}\frac{\mathrm{d}\tilde{\lambda}_a\mathrm{d}\tilde{x}_a}{2\pi}\right]\exp\left(-\frac{1}{2}\sum_a^n\left[x_a^2+\tilde{x}_a^2\right]-\sum_{a<b}\left[x_aq_{ab}x_b+\tilde{x}_a\tilde{q}_{ab}\tilde{x}_b\right]\right. \\
&\left.\qquad -\sum_{a,b}x_ar_{ab}\tilde{x}_b+\sum_a^n\left[\mathrm{i}x_a\lambda_a+\mathrm{i}\tilde{x}_a\tilde{\lambda}_a+\beta g(\lambda_a)+\beta g(\tilde{\lambda}_a)\right]\right), \\
&G_0(\{E_a,\tilde{E}_a,F_{ab},\tilde{F}_{ab},q_{ab},\tilde{q}_{ab},K_{ab},r_{ab}\})\equiv \\
&\quad \frac{\mathrm{i}}{2}\sum_a^n\left[E_a+\tilde{E}_a\right]-\mathrm{i}\sum_{a<b}\left[F_{ab}q_{ab}+\tilde{F}_{ab}\tilde{q}_{ab}\right]-\mathrm{i}\sum_{a,b}K_{ab}r_{ab}.
\end{aligned}
\tag{3.8}
$$

The quantities $\{F_{ab},\tilde{F}_{ab},K_{ab}\}$ are the conjugate variables to $\{q_{ab},\tilde{q}_{ab},r_{ab}\}$ respectively, and integrations with respect to them will also be evaluated by the method of steepest descent.

Evaluation at the Saddle Point

In the large connectivity limit the integrations in the first line of equation (3.7) can be performed by the method of steepest descent. This means we need to find the saddle-point which maximises the function $G\equiv G_J+\alpha G_\Lambda+G_0$ by solving the set of equations

$$
\begin{array}{lll}
1=\{J_a^2\}_{G_J}, & 1=\{\tilde{J}_a^2\}_{G_J}, & \forall a, \\
q_{ab}=\{J_aJ_b\}_{G_J}, & \tilde{q}_{ab}=\{\tilde{J}_a\tilde{J}_b\}_{G_J}, & \forall a,b\,(a<b), \\
\mathrm{i}F_{ab}=-\alpha\{x_ax_b\}_{G_\Lambda}, & \mathrm{i}\tilde{F}_{ab}=-\alpha\{\tilde{x}_a\tilde{x}_b\}_{G_\Lambda}, & \forall a,b\,(a<b)
\end{array}
\tag{3.9}
$$

and

$$r_{ab} = \{J_a \tilde{J}_b\}_{G_J}, \quad \mathrm{i}K_{ab} = -\alpha\{x_a \tilde{x}_b\}_{G_J}, \qquad \forall a, b. \tag{3.10}$$

Equations (3.9) are analogous to those already encountered in §2.5, but with different curly-braces operators complicated by having two networks to contend with. Before we evaluate the braces in equations (3.9) and (3.10), we shall first take the replica-symmetric ansatz given in equations (2.17) for the quantities $\{E_a, F_{ab}, q_{ab}\}$ and their dual $\{\tilde{E}_a, \tilde{F}_{ab}, \tilde{q}_{ab}\}$. We shall also employ $r_{ab} = r$, $K_{ab} = K$, $\forall a, b$ for the two remaining replica-dependent quantities. We can then write the two braces operators as

$$\begin{aligned}
\{(\cdots)\}_{G_J} = {} & \frac{1}{2}(1+m_s)\int_0^\infty\int_0^\infty \prod_a^n \left[\mathrm{d}J_a \mathrm{d}\tilde{J}_a\right](\cdots) \\
& \exp\left\{-\frac{\mathrm{i}}{2}(E+F)\sum_a^n J_a^2 - \frac{\mathrm{i}}{2}(\tilde{E}+\tilde{F})\sum_a^n \tilde{J}_a^2\right. \\
& \qquad \left. +\frac{\mathrm{i}}{2}\left(F\left[\sum_a^n J_a\right]^2 + 2K\left[\sum_a^n J_a\right]\left[\sum_a^n \tilde{J}_a\right] + \tilde{F}\left[\sum_a^n \tilde{J}_a\right]^2\right)\right\} \\
& \div \int_0^\infty\int_0^\infty \prod_a^n \left[\mathrm{d}J_a \mathrm{d}\tilde{J}_a\right] \exp\left\{-\frac{\mathrm{i}}{2}(E+F)\sum_a^n J_a^2 - \frac{\mathrm{i}}{2}(\tilde{E}+\tilde{F})\sum_a^n \tilde{J}_a^2\right. \\
& \qquad \left. +\frac{\mathrm{i}}{2}\left(F\left[\sum_a^n J_a\right]^2 + 2K\left[\sum_a^n J_a\right]\left[\sum_a^n \tilde{J}_a\right] + \tilde{F}\left[\sum_a^n \tilde{J}_a\right]^2\right)\right\} \\
+ {} & \frac{1}{2}(1-m_s)\int_0^\infty\int_{-\infty}^0 \prod_a^n \left[\mathrm{d}J_a \mathrm{d}\tilde{J}_a\right](\cdots) \\
& \exp\left\{-\frac{\mathrm{i}}{2}(E+F)\sum_a^n J_a^2 - \frac{\mathrm{i}}{2}(\tilde{E}+\tilde{F})\sum_a^n \tilde{J}_a^2\right. \\
& \qquad \left. +\frac{\mathrm{i}}{2}\left(F\left[\sum_a^n J_a\right]^2 + 2K\left[\sum_a^n J_a\right]\left[\sum_a^n \tilde{J}_a\right] + \tilde{F}\left[\sum_a^n \tilde{J}_a\right]^2\right)\right\} \\
& \div \int_0^\infty\int_{-\infty}^0 \prod_a^n \left[\mathrm{d}J_a \mathrm{d}\tilde{J}_a\right] \exp\left\{-\frac{\mathrm{i}}{2}(E+F)\sum_a^n J_a^2 - \frac{\mathrm{i}}{2}(\tilde{E}+\tilde{F})\sum_a^n \tilde{J}_a^2\right.
\end{aligned}$$

$$+\frac{\mathrm{i}}{2}\left(F\left[\sum_a^n J_a\right]^2 + 2K\left[\sum_a^n J_a\right]\left[\sum_a^n \tilde{J}_a\right] + \tilde{F}\left[\sum_a^n \tilde{J}_a\right]^2\right)\Bigg\} \tag{3.11}$$

and

$$\begin{aligned}\{(\cdots)\}_{G_\Lambda} = &\int\prod_a^n\left[\frac{\mathrm{d}\lambda_a \mathrm{d}x_a}{2\pi}\frac{\mathrm{d}\tilde{\lambda}_a \mathrm{d}\tilde{x}_a}{2\pi}\right](\cdots)\\
&\exp\Bigg\{-\frac{1}{2}(1-q)\sum_a^n x_a^2 - \frac{1}{2}(1-\tilde{q})\sum_a^n \tilde{x}_a^2\\
&\quad -\frac{1}{2}\left(q\left[\sum_a^n x_a\right]^2 + 2r\left[\sum_a^n x_a\right]\left[\sum_a^n \tilde{x}_a\right] + \tilde{q}\left[\sum_a^n \tilde{x}_a\right]^2\right)\\
&\quad +\sum_a^n\left[\mathrm{i}x_a\lambda_a + \mathrm{i}\tilde{x}_a\tilde{\lambda}_a + \beta g(\lambda_a) + \beta g(\tilde{\lambda}_a)\right]\Bigg\}\\
&\div\int\prod_a^n\left[\frac{\mathrm{d}\lambda_a \mathrm{d}x_a}{2\pi}\frac{\mathrm{d}\tilde{\lambda}_a \mathrm{d}\tilde{x}_a}{2\pi}\right]\exp\Bigg\{-\frac{1}{2}(1-q)\sum_a^n x_a^2 - \frac{1}{2}(1-\tilde{q})\sum_a^n \tilde{x}_a^2\\
&\quad -\frac{1}{2}\left(q\left[\sum_a^n x_a\right]^2 + 2r\left[\sum_a^n x_a\right]\left[\sum_a^n \tilde{x}_a\right] + \tilde{q}\left[\sum_a^n \tilde{x}_a\right]^2\right)\\
&\quad +\sum_a^n\left[\mathrm{i}x_a\lambda_a + \mathrm{i}\tilde{x}_a\tilde{\lambda}_a + \beta g(\lambda_a) + \beta g(\tilde{\lambda}_a)\right]\Bigg\}.\end{aligned} \tag{3.12}$$

The next step is to linearise and decouple the terms inside the exponentials of equations (3.11) and (3.12). For the first operator this is done by first re-writing in terms of the quantities

$$\mathrm{A}^2 \equiv \frac{1}{2}\left[1 + \frac{K}{\sqrt{F\tilde{F}}}\right] \quad \text{and} \quad \mathrm{B}^2 \equiv \frac{1}{2}\left[1 - \frac{K}{\sqrt{F\tilde{F}}}\right] \tag{3.13}$$

such that we can write

$$\exp\frac{\mathrm{i}}{2}\left(F\left[\sum_a^n J_a\right]^2 + 2K\left[\sum_a^n J_a\right]\left[\sum_a^n \tilde{J}_a\right] + \tilde{F}\left[\sum_a^n \tilde{J}_a\right]^2\right)$$

$$
\begin{aligned}
&= \exp\left(\frac{\mathrm{i}\mathrm{A}^2}{2}\left[\sqrt{F}\sum_a^n J_a + \sqrt{\tilde{F}}\sum_a^n \tilde{J}_a\right]^2 + \frac{\mathrm{i}\mathrm{B}^2}{2}\left[\sqrt{F}\sum_a^n J_a - \sqrt{\tilde{F}}\sum_a^n \tilde{J}_a\right]^2\right) \\
&= \int \mathcal{D}u\mathcal{D}v \prod_a^n \exp\left(\left[\mathrm{A}u + \mathrm{B}v\right] J_a \sqrt{\mathrm{i}F} + \left[\mathrm{A}u - \mathrm{B}v\right] \tilde{J}_a \sqrt{\mathrm{i}\tilde{F}}\right) \qquad (3.14)
\end{aligned}
$$

by performing two Hubbard-Stratonovich transformations (A.2) which add Gaussian integrals in the variables u and v.

We can treat the second curly braces operators (3.12) in the same manner by introducing

$$
\mathrm{a}^2 \equiv \frac{1}{2}\left[1 + \frac{r}{\sqrt{q\tilde{q}}}\right] \quad \text{and} \quad \mathrm{b}^2 \equiv \frac{1}{2}\left[1 - \frac{r}{\sqrt{q\tilde{q}}}\right] \qquad (3.15)
$$

and perform the linearising transformations by y and z, giving

$$
\begin{aligned}
&\exp -\frac{1}{2}\left(q\left[\sum_a^n x_a\right]^2 + 2r\left[\sum_a^n x_a\right]\left[\sum_a^n \tilde{x}_a\right] + \tilde{q}\left[\sum_a^n \tilde{x}_a\right]^2\right) \\
&= \exp\left(-\frac{\mathrm{a}^2}{2}\left[\sqrt{q}\sum_a^n x_a + \sqrt{\tilde{q}}\sum_a^n \tilde{x}_a\right]^2 - \frac{\mathrm{b}^2}{2}\left[\sqrt{q}\sum_a^n x_a - \sqrt{\tilde{q}}\sum_a^n \tilde{x}_a\right]^2\right) \\
&= \int \mathcal{D}y\mathcal{D}z \prod_a^n \exp\left(-\mathrm{i}\left[\mathrm{a}y + \mathrm{b}z\right] x\sqrt{q} - \mathrm{i}\left[\mathrm{a}y - \mathrm{b}z\right] \tilde{x}\sqrt{\tilde{q}}\right). \qquad (3.16)
\end{aligned}
$$

Upon replacing the above transforms (3.14) and (3.16) back into the operator definitions (3.11) and (3.12) respectively, we find the denominators go to unity in the zero replica limit. This means we can simplify saddle-point equations (3.9) and (3.10) and write them in terms of the following operators

$$
\langle(\cdots)\rangle_J \equiv \int_0^\infty \mathrm{d}J(\cdots)\exp\left(-\frac{\mathrm{i}}{2}(E+F)J^2 + (\mathrm{A}u + \mathrm{B}v)J\sqrt{\mathrm{i}F}\right)
$$

$$
\begin{aligned}
& \div \int_0^\infty \mathrm{d}J \exp\left(-\frac{\mathrm{i}}{2}(E+F)J^2 + (\mathrm{A}u+\mathrm{B}v)J\sqrt{\mathrm{i}F}\right), \\
\langle(\cdots)\rangle_{\tilde{J}+} \quad \equiv \quad & \int_0^\infty \mathrm{d}\tilde{J}(\cdots) \exp\left(-\frac{\mathrm{i}}{2}(\tilde{E}+\tilde{F})\tilde{J}^2 + (\mathrm{A}u-\mathrm{B}v)\tilde{J}\sqrt{\mathrm{i}\tilde{F}}\right) \\
& \div \int_0^\infty \mathrm{d}\tilde{J}(\cdots) \exp\left(-\frac{\mathrm{i}}{2}(\tilde{E}+\tilde{F})\tilde{J}^2 + (\mathrm{A}u-\mathrm{B}v)\tilde{J}\sqrt{\mathrm{i}\tilde{F}}\right), \\
\langle(\cdots)\rangle_{\tilde{J}-} \quad \equiv \quad & \int_{-\infty}^0 \mathrm{d}\tilde{J}(\cdots) \exp(\left(-\frac{\mathrm{i}}{2}(\tilde{E}+\tilde{F})\tilde{J}^2 + (\mathrm{A}u-\mathrm{B}v)\tilde{J}\sqrt{\mathrm{i}\tilde{F}}\right) \\
& \div \int_{-\infty}^0 \mathrm{d}\tilde{J} \exp(\left(-\frac{\mathrm{i}}{2}(\tilde{E}+\tilde{F})\tilde{J}^2 + (\mathrm{A}u-\mathrm{B}v)\tilde{J}\sqrt{\mathrm{i}\tilde{F}}\right), \qquad (3.17)
\end{aligned}
$$

and

$$
\begin{aligned}
& \langle(\cdots)\rangle_{\Lambda} \equiv \int \frac{\mathrm{d}\lambda \mathrm{d}x}{2\pi}(\cdots)\exp\left(-\frac{x^2}{2}(1-q) + \mathrm{i}\left[\lambda - (\mathrm{a}y+\mathrm{b}z)\sqrt{q} + \beta g(\lambda)\right]x\right) \\
& \div \quad \int \frac{\mathrm{d}\lambda \mathrm{d}x}{2\pi}\exp\left(-\frac{x^2}{2}(1-q) + \mathrm{i}\left[\lambda - (\mathrm{a}y+\mathrm{b}z)\sqrt{q} + \beta g(\lambda)\right]x\right), \\
& \langle(\cdots)\rangle_{\tilde{\Lambda}} \equiv \int \frac{\mathrm{d}\tilde{\lambda} \mathrm{d}\tilde{x}}{2\pi}(\cdots)\exp\left(-\frac{\tilde{x}^2}{2}(1-\tilde{q}) + \mathrm{i}\left[\tilde{\lambda} - (\mathrm{a}y-\mathrm{b}z)\sqrt{\tilde{q}} + \beta g(\tilde{\lambda})\right]\tilde{x}\right) \\
& \div \quad \int \frac{\mathrm{d}\tilde{\lambda} \mathrm{d}\tilde{x}}{2\pi}\exp\left(-\frac{\tilde{x}^2}{2}(1-\tilde{q}) + \mathrm{i}\left[\tilde{\lambda} - (\mathrm{a}y-\mathrm{b}z)\sqrt{\tilde{q}} + \beta g(\tilde{\lambda})\right]\tilde{x}\right). \qquad (3.18)
\end{aligned}
$$

We shall now use the above definitions and simplify each of the saddle-point equations (3.9) and (3.10) in turn. We begin with the one obtained by taking the derivative with respect to $\{E_a\}$

$$
\begin{aligned}
1 \quad & = \quad \{J_a^2\}_{G_J} \\
& = \quad \int \mathcal{D}u\mathcal{D}v\left\{\frac{1}{2}(1+m_s)\langle J^2\rangle_J\langle 1\rangle_{\tilde{J}+} + \frac{1}{2}(1-m_s)\langle J^2\rangle_J\langle 1\rangle_{\tilde{J}-}\right\} \\
& = \quad \int \mathcal{D}u\mathcal{D}v\left\{\langle J^2\rangle_J\right\}, \qquad (3.19)
\end{aligned}
$$

and similarly for the derivative with respect to $\{\tilde{E}_a\}$

$$\begin{aligned} 1 &= \{\tilde{J}_a^2\}_{G_J} \\ &= \int \mathcal{D}u\mathcal{D}v \left\{\frac{1}{2}(1+m_s)\langle 1\rangle_J\langle \tilde{J}^2\rangle_{\tilde{J}+} + \frac{1}{2}(1-m_s)\langle 1\rangle_J\langle \tilde{J}^2\rangle_{\tilde{J}-}\right\} \\ &= \int \mathcal{D}u\mathcal{D}v \left\{\frac{1}{2}(1+m_s)\langle \tilde{J}^2\rangle_{\tilde{J}+} + \frac{1}{2}(1-m_s)\langle \tilde{J}^2\rangle_{\tilde{J}-}\right\}. \end{aligned} \tag{3.20}$$

Likewise, the derivatives with respect to $\{F_{ab}\}$ and $\{\tilde{F}_{ab}\}$ are

$$\begin{aligned} q &= \{J_a J_b\}_{G_J} \\ &= \int \mathcal{D}u\mathcal{D}v \left\{\langle J\rangle_J^2\right\}, \\ \tilde{q} &= \{\tilde{J}_a\tilde{J}_b\}_{G_J} \\ &= \int \mathcal{D}u\mathcal{D}v \left\{\frac{1}{2}(1+m_s)\langle \tilde{J}\rangle_{\tilde{J}+}^2 + \frac{1}{2}(1-m_s)\langle \tilde{J}\rangle_{\tilde{J}-}^2\right\}, \end{aligned} \tag{3.21}$$

respectively. Differentiating with respect to $\{q_{ab}\}$ and $\{\tilde{q}_{ab}\}$ gives simply

$$\begin{aligned} \mathrm{i}F &= -\alpha\{x_a x_b\}_{G_\Lambda} \\ &= -\alpha\int \mathcal{D}y\mathcal{D}z \left\{\langle x\rangle_\Lambda^2\right\}, \\ \mathrm{i}\tilde{F} &= -\alpha\{\tilde{x}_a\tilde{x}_b\}_{G_\Lambda} \\ &= -\alpha\int \mathcal{D}y\mathcal{D}z \left\{\langle \tilde{x}\rangle_\Lambda^2\right\}, \end{aligned} \tag{3.22}$$

again respectively. We can see straightaway that the solutions to equations (3.19)–(3.22) can be placed into two groups, each concerned with its own network. Hence we have two sets of *intra-network* quantities $\{E_a, F_{ab}, q_{ab}\}$ and $\{\tilde{E}_a, \tilde{F}_{ab}, \tilde{q}_{ab}\}$, in addition to the *inter-network* quantities $\{r_{ab}, K_{ab}\}$. It is in the saddle-point equations of the inter-network quantities that the inter-network overlap m_s of definition (3.3) arises. Equating the derivative with respect to $\{K_{ab}\}$ equal to zero, we have

$$\begin{aligned} r &= \{J_a\tilde{J}_b\}_{G_J} \\ &= \int \mathcal{D}u\mathcal{D}v \left\{\frac{1}{2}(1+m_s)\langle J\rangle_J\langle\tilde{J}\rangle_{\tilde{J}+} + \frac{1}{2}(1-m_s)\langle J\rangle_J\langle\tilde{J}\rangle_{\tilde{J}-}\right\}. \end{aligned} \tag{3.23}$$

Finally, the differentiation with respect to $\{r_{ab}\}$ gives us

$$\begin{aligned} \mathrm{i}K &= -\alpha\{x_a\tilde{x}_b\}_{G_\Lambda} \\ &= -\alpha\int \mathcal{D}y\mathcal{D}z\,\{\langle x\rangle_\Lambda\langle\tilde{x}\rangle_{\tilde{\Lambda}}\}. \end{aligned} \tag{3.24}$$

We have now expressed the saddle-point equations as Gaussian weighted integrations of moments with respect to the operators (3.17) and (3.18). Appendix E shows how the moments of the former are calculated, but here we shall simply quote the ones we need. For the evaluation of equations (3.19)—(3.21) and (3.23), these are

$$\begin{aligned} \langle J\rangle_J &= \frac{\mathrm{A}u+\mathrm{B}v}{\sqrt{\mathrm{i}E+\mathrm{i}F}}\sqrt{\frac{F}{E+F}}\theta[\mathrm{A}u+\mathrm{B}v], \\ \langle\tilde{J}\rangle_{\tilde{J}+} &= \frac{\mathrm{A}u-\mathrm{B}v}{\sqrt{\mathrm{i}\tilde{E}+\mathrm{i}\tilde{F}}}\sqrt{\frac{\tilde{F}}{\tilde{E}+\tilde{F}}}\theta[\mathrm{A}u-\mathrm{B}v], \\ \langle\tilde{J}\rangle_{\tilde{J}-} &= \frac{\mathrm{A}u-\mathrm{B}v}{\sqrt{\mathrm{i}\tilde{E}+\mathrm{i}\tilde{F}}}\sqrt{\frac{\tilde{F}}{\tilde{E}+\tilde{F}}}\theta[-(\mathrm{A}u-\mathrm{B}v)], \\ \langle J^2\rangle_J &= \frac{1}{\mathrm{i}E+\mathrm{i}F}\left(1+[\mathrm{A}u+\mathrm{B}v]^2\,\frac{F}{E+F}\right)\theta[\mathrm{A}u+\mathrm{B}v], \end{aligned}$$

$$
\begin{aligned}
\langle \tilde{J}^2\rangle_{\tilde{J}+} &= \frac{1}{\mathrm{i}\tilde{E}+\mathrm{i}\tilde{F}}\left(1+[\mathrm{A}u-\mathrm{B}v]^2\,\frac{\tilde{F}}{\tilde{E}+\tilde{F}}\right)\theta[\mathrm{A}u-\mathrm{B}v],\\
\langle \tilde{J}^2\rangle_{\tilde{J}-} &= \frac{1}{\mathrm{i}\tilde{E}+\mathrm{i}\tilde{F}}\left(1+[\mathrm{A}u-\mathrm{B}v]^2\,\frac{\tilde{F}}{\tilde{E}+\tilde{F}}\right)\theta[-(\mathrm{A}u-\mathrm{B}v)].
\end{aligned}
\tag{3.25}
$$

For the remaining three saddle-point equations (3.22) and (3.24) we need the moments with respect to the operator (3.18), which we can write as

$$
\begin{aligned}
\langle x^k\rangle_{\Lambda} &= \int\frac{\mathrm{d}\lambda}{\sqrt{2\pi}}\int\mathcal{D}x\;(1-q)^{-\frac{1}{2}(1+k)}\left[x+\mathrm{i}\left(\frac{\lambda-(\mathrm{a}y+\mathrm{b}z)\sqrt{q}}{\sqrt{1-q}}\right)\right]^k\\
&\quad\times\exp\left(\beta g(\lambda)+\mathrm{i}x\left[\frac{\lambda-(\mathrm{a}y+\mathrm{b}z)\sqrt{q}}{\sqrt{1-q}}\right]\right)\\
&\quad\div\int\frac{\mathrm{d}\lambda}{\sqrt{2\pi}}(1-q)^{-\frac{1}{2}}\exp\beta\left(g(\lambda)-\frac{(\lambda-(\mathrm{a}y+\mathrm{b}z)\sqrt{q})^2}{2\beta(1-q)}\right),\\
\langle \tilde{x}^k\rangle_{\tilde{\Lambda}} &= \int\frac{\mathrm{d}\tilde{\lambda}}{\sqrt{2\pi}}\int\mathcal{D}\tilde{x}\;(1-\tilde{q})^{-\frac{1}{2}(1+k)}\left[\tilde{x}+\mathrm{i}\left(\frac{\tilde{\lambda}-(\mathrm{a}y-\mathrm{b}z)\sqrt{\tilde{q}}}{\sqrt{1-\tilde{q}}}\right)\right]^k\\
&\quad\times\exp\left(\beta g(\tilde{\lambda})+\mathrm{i}\tilde{x}\left[\frac{\tilde{\lambda}-(\mathrm{a}y-\mathrm{b}z)\sqrt{\tilde{q}}}{\sqrt{1-\tilde{q}}}\right]\right)\\
&\quad\div\int\frac{\mathrm{d}\tilde{\lambda}}{\sqrt{2\pi}}(1-\tilde{q})^{-\frac{1}{2}}\exp\beta\left(g(\tilde{\lambda})-\frac{(\tilde{\lambda}-(\mathrm{a}y-\mathrm{b}z)\sqrt{\tilde{q}})^2}{2\beta(1-\tilde{q})}\right).
\end{aligned}
\tag{3.26}
$$

The integrations with respect to the x and $\tilde{x}$ variables are standard moments of a Gaussian distribution. However, the integrations with respect to λ and $\tilde{\lambda}$ require us to take the low annealed-temperature limit of ($\beta \to \infty$), which then allows us to use the method of steepest descent. Consequently, this requires us to find the functions $\hat{\lambda}(y,z)$ and $\hat{\tilde{\lambda}}(y,z)$ which respectively maximise

$$
\begin{aligned}
\mathcal{G}(\lambda) &\equiv g(\lambda)-\frac{1}{\gamma^2}\left[\lambda-(\mathrm{a}y+\mathrm{b}z)\sqrt{q}\right]^2,\\
\tilde{\mathcal{G}}(\tilde{\lambda}) &\equiv g(\tilde{\lambda})-\frac{1}{\tilde{\gamma}^2}\left[\tilde{\lambda}-(\mathrm{a}-\mathrm{b}z)\sqrt{\tilde{q}}\right]^2,
\end{aligned}
\tag{3.27}
$$

where the parameters γ and $\tilde{\gamma}$ are defined as

$$\gamma \equiv 2\beta(1-q) \qquad \text{and} \qquad \tilde{\gamma} \equiv 2\beta(1-\tilde{q}) \tag{3.28}$$

in an identical manner to equation (2.23). Taking the optimal perceptron limits of $(q \to 1)$ and $(\tilde{q} \to 1)$ for both the networks, the moments we seek are

$$\begin{aligned} \langle x \rangle_{\Lambda} &= \frac{\mathrm{i}}{\sqrt{1-q}} \left(\frac{\hat{\lambda}(y,z) - (\mathrm{a}y + \mathrm{b}z)}{\sqrt{1-q}} \right), \\ \langle \tilde{x} \rangle_{\tilde{\Lambda}} &= \frac{\mathrm{i}}{\sqrt{1-\tilde{q}}} \left(\frac{\hat{\tilde{\lambda}}(y,z) - (\mathrm{a}y - \mathrm{b}z)}{\sqrt{1-q}} \right). \end{aligned} \tag{3.29}$$

Upon substituting the moments (3.25) and (3.29) into the saddle-point equations (3.19)–(3.22), we can reduce the double Gaussian integrals into single ones. Hence we obtain the following saddle-point equations

$$\begin{aligned} 1 &= \frac{1}{2(\mathrm{i}E + \mathrm{i}F)} \left(\frac{E + 2F}{E + F} \right), & 1 &= \frac{1}{2(\mathrm{i}\tilde{E} + \mathrm{i}\tilde{F})} \left(\frac{\tilde{E} + 2\tilde{F}}{\tilde{E} + \tilde{F}} \right), \\ q &= \frac{1}{2(\mathrm{i}E + \mathrm{i}F)} \left(\frac{F}{E + F} \right), & \tilde{q} &= \frac{1}{2(\mathrm{i}\tilde{E} + \mathrm{i}\tilde{F})} \left(\frac{\tilde{F}}{\tilde{E} + \tilde{F}} \right), \\ \mathrm{i}F &= \frac{\alpha}{(1-q)^2} \int \mathcal{D}z \left(\hat{\lambda}(z) - z \right)^2, & \mathrm{i}\tilde{F} &= \frac{\alpha}{(1-\tilde{q})^2} \int \mathcal{D}z \left(\hat{\tilde{\lambda}}(z) - z \right)^2, \end{aligned} \tag{3.30}$$

the first four of which we can solve in terms of the inter-replica order parameters q and $\tilde{q}$ in the optimal perceptron limits to give

$$\begin{aligned} \mathrm{i}E + \mathrm{i}F &= \frac{1}{2(1-q)}, & \mathrm{i}\tilde{E} + \mathrm{i}\tilde{F} &= \frac{1}{2(1-\tilde{q})}, \\ \mathrm{i}F &= \frac{1}{2(1-q)^2}, & \mathrm{i}\tilde{F} &= \frac{1}{2(1-\tilde{q})^2}. \end{aligned} \tag{3.31}$$

Upon substituting equations (3.31) back into the last two expressions of equations (3.30), we obtain the optimal storage capacity α_C. We find this quantity is simply half that of the unconstrained network which is given in equation (2.29).

The remaining two saddle-point equations concern the correlation between the networks. We shall deal with them in turn beginning with equation (3.23). Inserting equations (3.25) and (3.31), it becomes

$$\frac{r}{2} = \left[\frac{1}{2}(1+m_s) \int\limits_{\substack{\mathrm{A}u+\mathrm{B}v>0 \\ \mathrm{A}u-\mathrm{B}v>0}} \mathcal{D}u\mathcal{D}v + \frac{1}{2}(1-m_s) \int\limits_{\substack{\mathrm{A}u+\mathrm{B}v>0 \\ \mathrm{A}u-\mathrm{B}v<0}} \mathcal{D}u\mathcal{D}v \right] (\mathrm{A}^2u^2 - \mathrm{B}^2v^2)$$

which is solved by transforming into plane-polar co-ordinates to give

$$\begin{aligned} \frac{r}{2} &= \left[\frac{1}{2}(1+m_s) \int\limits_{-\Phi}^{\Phi} \mathrm{d}\theta + \frac{1}{2}(1-m_s) \int\limits_{\Phi}^{\pi-\Phi} \mathrm{d}\theta \right] (\mathrm{A}^2\cos^2\theta - \mathrm{B}^2\sin^2\theta) \\ &\quad \times \int\limits_0^{\infty} \frac{\mathrm{d}s}{2\pi} s^3 \exp\left(-\frac{1}{2}s^2\right) \\ &= \frac{1}{2}(1-m_s)\pi R + \frac{2}{\pi} m_s R \sin^{-1}\Phi + \frac{m_s}{\pi}\sqrt{1-R^2} \end{aligned} \tag{3.32}$$

where we have defined

$$\begin{aligned} \Phi &\equiv \tan^{-1}(\mathrm{A}/\mathrm{B}), \\ R &\equiv \frac{K}{\sqrt{F\tilde{F}}}, \\ &= \mathrm{A}^2 - \mathrm{B}^2. \end{aligned} \tag{3.33}$$

To solve the final saddle-point equation (3.24) we shall now specify the form of the training function $g(\lambda)$. Namely we shall use the original Gardner theta function [Gar88] for both the networks

$$g(\lambda) = \theta[\lambda - \kappa], \qquad \text{and} \qquad \tilde{g}(\tilde{\lambda}) = \theta[\tilde{\lambda} - \kappa], \tag{3.34}$$

which simply states that we want the alignment field to be larger than some positive *minimum stability constant* κ for all the memory patterns. We shall further simplify matters by going to the limit of zero error in the storage by taking the parameter γ towards infinity —a procedure already discussed in §2.8. The functions $\hat{\lambda}(y,z)$ and $\hat{\hat{\lambda}}(y,z)$ which maximise equations (3.27) and (3.34) are

$$\hat{\lambda}(y,z) = \begin{cases} \kappa & \text{for } (\mathrm{a}y + \mathrm{b}z) = -\infty \ldots \kappa, \\ (\mathrm{a}y + \mathrm{b}z) & \text{otherwise,} \end{cases} \tag{3.35}$$

and

$$\hat{\hat{\lambda}}(y,z) = \begin{cases} \kappa & \text{for } (\mathrm{a}y - \mathrm{b}z) = -\infty \ldots \kappa, \\ (\mathrm{a}y - \mathrm{b}z) & \text{otherwise.} \end{cases} \tag{3.36}$$

Substituting the above equations (3.35) and (3.36) into the moments (3.29) we get from the saddle-point equations (3.30) the optimal storage capacity

$$\alpha_C = \frac{1}{2}\left[\int_{-\infty}^{\kappa} \mathcal{D}z \; (\kappa - z)^2\right]^{-1} \tag{3.37}$$

and for the final saddle-point equation (3.24)

$$\begin{aligned} \mathrm{i}K(1-q)(1-\tilde{q}) &= \alpha \int \mathcal{D}y\mathcal{D}z[\hat{\lambda}(y,z) - (\mathrm{a}y + \mathrm{b}z)][\hat{\hat{\lambda}}(y,z) - (\mathrm{a}y - \mathrm{b}z)] \\ R &= 2\alpha_C \int\limits_{\substack{\mathrm{a}y+\mathrm{b}z<\kappa \\ \mathrm{a}y-\mathrm{b}z<\kappa}} \mathcal{D}y\mathcal{D}z[\kappa - (\mathrm{a}y + \mathrm{b}z)][\kappa - (\mathrm{a}y - \mathrm{b}z)]. \end{aligned} \tag{3.38}$$

using definition (3.33) for the quantity R. We shall make use of equation (3.37) to transform the minimum stability constant κ in preference for the more physically intuitive optimal storage capacity.

The remaining saddle-point equation (3.38) is dealt with by first shifting ($y \rightarrow y' = y - \kappa/\mathrm{a}$), and then moving over to plane-polar co-ordinates to give

$$\begin{aligned} R &= \frac{\alpha_C}{\pi} \int_{\pi-\varphi}^{\pi+\varphi} \mathrm{d}\theta \int_0^\infty \mathrm{d}s\ s^3 \exp\left(-\frac{s^2}{2} - \frac{\kappa s}{\mathrm{a}} \cos\theta - \frac{\kappa^2}{2\mathrm{a}^2}\right) [\mathrm{a}^2 \cos^2\theta - \mathrm{b}^2 \sin^2\theta] \\ &= \frac{\alpha_C}{2\pi} \left\{ (2\pi)^{-1} \exp\left(-\frac{\kappa^2}{2\mathrm{a}^2}\right) \right. \\ &\qquad \times \left[2\varphi r + 2\mathrm{ab} + 2\mathrm{ab}\frac{\kappa^2}{\mathrm{a}^2}\left(\frac{3}{8} - \frac{\mathrm{b}^2}{4}\right) + 2\varphi\frac{\kappa^2}{\mathrm{a}^2}\left(\frac{3}{8} - \frac{\mathrm{b}^2}{2}\right)\right] \\ &\quad \left. + \int_{-\kappa}^{\kappa} \mathcal{D}x \left[\kappa^2 - x^2\right] \left[\frac{\kappa^2}{\mathrm{a}^2} - x^2 + 3\right] \overline{\mathrm{H}}\left[\sqrt{\frac{\kappa^2}{\mathrm{a}^2} - x^2}\right] \right\}, \end{aligned} \tag{3.39}$$

where $\varphi \equiv \tan^{-1}(\mathrm{a}/\mathrm{b})$. Equations (3.32) and (3.39) are coupled with each other, with the solution (r, R) found numerically. There are efficient algorithms for solving these equations [PFTV88] and in practice this is but a minor hindrance to obtaining the results which follow.

Evaluation of the Alignment Field Dual Distribution

Using the angled-brackets notation of equations (3.18), we can evaluate the dual distribution of the alignment field (3.7) as

$$\begin{aligned} \rho(\Lambda, \tilde{\Lambda}) &= \int \mathcal{D}y\mathcal{D}z\ \langle\delta[\Lambda - \lambda_1^1]\rangle_\Lambda \langle\delta[\tilde{\Lambda} - \tilde{\lambda}_1^1]\rangle_{\tilde{\Lambda}} \\ &= \int \mathcal{D}y\mathcal{D}z\ \delta[\Lambda - \hat{\lambda}(y,z)]\delta[\tilde{\Lambda} - \hat{\tilde{\lambda}}(y,z)]. \end{aligned}$$

Using the form of $\hat{\lambda}(y,z)$ and $\hat{\hat{\lambda}}(y,z)$ given by equations (3.35) and (3.36), this splits the double Gaussian integral into four regions

$$\begin{aligned}
\rho(\Lambda,\tilde{\Lambda}) \;=\; & \int\limits_{\substack{\mathrm{a}y+\mathrm{b}z<\kappa \\ \mathrm{a}y-\mathrm{b}z<\kappa}} \mathcal{D}y\mathcal{D}z\; \delta[\Lambda-\kappa]\delta[\tilde{\Lambda}-\kappa] \\
& + \int\limits_{\substack{\mathrm{a}y+\mathrm{b}z<\kappa \\ \mathrm{a}y-\mathrm{b}z>\kappa}} \mathcal{D}y\mathcal{D}z\; \delta[\Lambda-\kappa]\delta[\tilde{\Lambda}-(\mathrm{a}y-\mathrm{b}z)] \\
& + \int\limits_{\substack{\mathrm{a}y+\mathrm{b}z>\kappa \\ \mathrm{a}y-\mathrm{b}z<\kappa}} \mathcal{D}y\mathcal{D}z\; \delta[\Lambda-(\mathrm{a}y+\mathrm{b}z)]\delta[\tilde{\Lambda}-\kappa] \\
& + \int\limits_{\substack{\mathrm{a}y+\mathrm{b}z>\kappa \\ \mathrm{a}y-\mathrm{b}z>\kappa}} \mathcal{D}y\mathcal{D}z\; \delta[\Lambda-(\mathrm{a}y+\mathrm{b}z)]\delta[\tilde{\Lambda}-(\mathrm{a}y-\mathrm{b}z)]
\end{aligned}$$

which evaluates to

$$\begin{aligned}
\rho(\Lambda,\tilde{\Lambda}) \;=\; & \delta[\Lambda-\kappa]\delta[\tilde{\Lambda}-\kappa]\int_{-\infty}^{\kappa}\mathcal{D}x\;\overline{\mathrm{H}}\left[\frac{\kappa-rx}{\sqrt{1-r^2}}\right] \\
& +\delta[\Lambda-\kappa]\theta[\tilde{\Lambda}-\kappa](2\pi)^{-\frac{1}{2}}\overline{\mathrm{H}}\left[\frac{\kappa-r\tilde{\Lambda}}{\sqrt{1-r^2}}\right]\exp\left(-\frac{1}{2}\tilde{\Lambda}^2\right) \\
& +\delta[\tilde{\Lambda}-\kappa]\theta[\Lambda-\kappa](2\pi)^{-\frac{1}{2}}\overline{\mathrm{H}}\left[\frac{\kappa-r\Lambda}{\sqrt{1-r^2}}\right]\exp\left(-\frac{1}{2}\Lambda^2\right) \\
& +\frac{\theta[\Lambda-\kappa]\theta[\tilde{\Lambda}-\kappa]}{2\pi\sqrt{1-r^2}}\exp\left(-\frac{\tilde{\Lambda}^2}{2}-\frac{1}{2}\left[\frac{\Lambda-r\tilde{\Lambda}}{\sqrt{1-r^2}}\right]^2\right). \qquad (3.40)
\end{aligned}$$

Equation (3.40) can be integrated with respect to the alignment field Λ to give the single distribution function $\rho(\tilde{\Lambda})$. As a useful consistency check we find this is the same as the Gardner optimal network, as has been reported in reference [AWC89a]. Furthermore, we can then introduce the conditional probability function $\rho(\Lambda \mid \tilde{\Lambda})$. For $\tilde{\Lambda} > \kappa$, this probability is

$$\begin{aligned}\rho(\Lambda \mid \tilde{\Lambda}) &\equiv \frac{\rho(\Lambda, \tilde{\Lambda})}{\rho(\tilde{\Lambda})} \\ &= \frac{\theta[\Lambda - \kappa]}{\sqrt{2\pi(1-r^2)}} \exp\left(-\frac{1}{2}\left[\frac{\Lambda - r\tilde{\Lambda}}{\sqrt{1-r^2}}\right]^2\right) + \delta[\Lambda - \kappa]\overline{\mathrm{H}}\left[\frac{\kappa - r\tilde{\Lambda}}{\sqrt{1-r^2}}\right]\end{aligned} \tag{3.41}$$

which has an obvious dependency upon the parameter r. The significance and plot of this parameter will be given in the results section, but we shall now turn our attention to the dual distribution function for the synaptic efficacy.

3.4 The Synaptic Field Dual Distribution

The second dual-distribution function we wish to calculate is for the networks' weights. These weights are a direct result of the annealed optimising procedure and the distribution ought to tell us how the networks have coped with their own set of sign constraints. We shall begin by calculating the dual-distribution function $\rho(w, \tilde{w})$, but as in the case of the alignment field in §3.3, this is an essential intermediate step towards finding the more useful conditional probability distribution function $\rho(w \mid \tilde{w})$.

In replica form, we may write the dual-distribution function by sampling in both networks a typical weight index $j = 1$ and replica $a = 1$, giving

$$\begin{aligned}\rho(w, \tilde{w}) = \lim_{\substack{n \to 0 \\ C \to \infty}} \int \prod_a^n \left[\frac{\mathrm{d}E_a}{4\pi}\frac{\mathrm{d}\tilde{E}_a}{4\pi}\right] \prod_{a<b}\left[\frac{\mathrm{d}F_{ab}\mathrm{d}q_{ab}}{2\pi/C}\frac{\mathrm{d}\tilde{F}_{ab}\mathrm{d}\tilde{q}_{ab}}{2\pi/C}\right] \prod_{a,b}\left[\frac{\mathrm{d}K_{ab}\mathrm{d}r_{ab}}{2\pi/C}\right] \\ \exp\ C\Big\{G_J(\{E_a, \tilde{E}_a, F_{ab}, \tilde{F}_{ab}, K_{ab}\}) + \alpha G_\Lambda(\{q_{ab}, \tilde{q}_{ab}, r_{ab}\}) \\ + G_0(\{E_a, \tilde{E}_a, F_{ab}, \tilde{F}_{ab}, K_{ab}, q_{ab}, \tilde{q}_{ab}, r_{ab}\})\Big\}\end{aligned}$$

$$\times \int \prod_{a}^{n} \left[\theta[\mathrm{g}^1 J_a^1]\mathrm{d}J_a^1 \theta[\tilde{\mathrm{g}}^1 \tilde{J}_a^1]\mathrm{d}\tilde{J}_a^1\right] \delta\left[w - J_1^1\right] \delta\left[\tilde{w} - \tilde{J}_1^1\right]$$
$$\times \exp\left(-\frac{\mathrm{i}}{2}\sum_{a}^{n}\left[E_a\left(J_a^1\right)^2 + \tilde{E}_a\left(\tilde{J}_a^1\right)^2\right]\right.$$
$$\left. + \mathrm{i}\sum_{a<b}\left[J_a F_{ab} J_b + \tilde{J}_a \tilde{F}_{ab} \tilde{J}_b\right] + \mathrm{i}\sum_{a,b} J_a K_{ab} \tilde{J}_b\right) \qquad (3.42)$$

where the functions in the exponentials are as given in equations (3.8). Hence the determination of the saddle-point equations also carries over here and we may proceed straightaway with the distribution function itself. By introducing two linearising Hubbard-Stratonovich transformations, we can borrow the angled-brackets notation of equations (3.17) and write the above equation (3.42) free of the superfluous replica and weight indices as

$$\rho(w, \tilde{w}) = \int \mathcal{D}u\mathcal{D}v \left\{ \langle \delta[w - J] \rangle_{J\pm} \langle \delta[\tilde{w} - \tilde{J}] \rangle_{\tilde{J}\pm} \right\} \qquad (3.43)$$

where the operators $\langle(\cdots)\rangle_{J\pm}$ and $\langle(\cdots)\rangle_{\tilde{J}\pm}$ are constructed to handle the weight signs g and $\tilde{\mathrm{g}}$, such that

$$\begin{aligned} \langle \delta[w - J] \rangle_{J\pm} &= \int_0^\infty \mathrm{d}J\ \delta[w - \mathrm{g}J] \exp\left(-\frac{\mathrm{i}}{2}(E + F)J^2 + \mathrm{g}(\mathrm{A}u + \mathrm{B}v)J\sqrt{\mathrm{i}F}\right) \\ &\quad \div \int_0^\infty \mathrm{d}J\ \exp\left(-\frac{\mathrm{i}}{2}(E + F)J^2 + \mathrm{g}(\mathrm{A}u + \mathrm{B}v)J\sqrt{\mathrm{i}F}\right) \qquad (3.44) \end{aligned}$$

and for the conjugate network

$$\begin{aligned} \langle \delta[\tilde{w} - \tilde{J}] \rangle_{\tilde{J}\pm} &= \int_0^\infty \mathrm{d}\tilde{J}\ \delta[\tilde{w} - \tilde{\mathrm{g}}\tilde{J}] \exp\left(-\frac{\mathrm{i}}{2}(\tilde{E} + \tilde{F})\tilde{J}^2 + \tilde{\mathrm{g}}(\mathrm{A}u - \mathrm{B}v)\tilde{J}\sqrt{\mathrm{i}\tilde{F}}\right) \\ &\quad \div \int_0^\infty \mathrm{d}\tilde{J} \exp\left(-\frac{\mathrm{i}}{2}(\tilde{E} + \tilde{F})\tilde{J}^2 + \tilde{\mathrm{g}}(\mathrm{A}u - \mathrm{B}v)\tilde{J}\sqrt{\mathrm{i}\tilde{F}}\right). \qquad (3.45) \end{aligned}$$

Equations (3.44) and (3.45) are sufficiently similar that we shall detail the treatment of only the former. Firstly, the denominator can be written as the function $\overline{\mathrm{H}}[\cdots]$ defined in equation (A.4). Next, the delta function allows the integral in the numerator to be easily evaluated, giving for $\mathrm{g}w \geq 0$

$$\begin{aligned}\langle\delta[w-J]\rangle_{J\pm} &= \exp\left(-\frac{\mathrm{i}}{2}(E+F)w^2+\mathrm{g}(\mathrm{A}u+\mathrm{B}v)w\sqrt{\mathrm{i}F}\right)\\ &\div\sqrt{\frac{2\pi}{\mathrm{i}E+\mathrm{i}F}}\exp\left(\frac{(\mathrm{A}u+\mathrm{B}v)^2F}{2(E+F)}\right)\overline{\mathrm{H}}\left[\frac{\mathrm{g}(\mathrm{A}u+\mathrm{B}v)\sqrt{F}}{\sqrt{E+F}}\right]\end{aligned}$$

with the denominator essentially serving to normalise this function in w. However, we know that in the optimal perceptron limit the saddle-point solutions grow according to equations (3.31). This permits us to expand the function $\overline{\mathrm{H}}[\cdots]$ in the denominator via equation (A.5) for the appropriate range in the argument. It also means the exponential function in the numerator will have the quantity w dominated by either its critical point or by zero. That is, we first acknowledge the restriction $\mathrm{g}w \geq 0$ placed by the sign-constraint, so in the optimal perceptron case w will be dominated either by $\mathrm{g}(\mathrm{A}u+\mathrm{B}v)\sqrt{F}/\sqrt{E+F}$ or by zero, depending whether $\mathrm{g}(\mathrm{A}u+\mathrm{B}v)$ is positive or negative respectively. In either case, since the denominator serves as a normaliser, so the equations are reduced to Dirac delta-functions and we can succinctly write down for equations (3.44) and (3.45)

$$\begin{aligned}\langle\delta[w-J]\rangle_{J\pm} &= \delta\left[w-\frac{\mathrm{g}(\mathrm{A}u+\mathrm{B}v)\sqrt{\mathrm{i}F}}{\mathrm{i}E+\mathrm{i}F}\right]\theta[\mathrm{g}(\mathrm{A}u+\mathrm{B}v)]\\ &\qquad+\delta[w]\ \theta[-\mathrm{g}(\mathrm{A}u+\mathrm{B}v)],\\ \langle\delta[\tilde{w}-\tilde{J}]\rangle_{\tilde{J}\pm} &= \delta\left[\tilde{w}-\frac{\tilde{\mathrm{g}}(\mathrm{A}u-\mathrm{B}v)\sqrt{\mathrm{i}\tilde{F}}}{\mathrm{i}\tilde{E}+\mathrm{i}\tilde{F}}\right]\theta[\tilde{\mathrm{g}}(\mathrm{A}u-\mathrm{B}v)]\\ &\qquad+\delta[\tilde{w}]\ \theta[-\tilde{\mathrm{g}}(\mathrm{A}u-\mathrm{B}v)]. \qquad (3.46)\end{aligned}$$

Substituting equations (3.46) back into expression (3.43) for the dual distribution function, we find the integration region for the double Gaussian weighted integrals are split by the theta functions into four regions and hence

$$\begin{aligned}
\rho(w,\tilde{w}) \;&=\; \delta[w]\,\delta[\tilde{w}] \int_{\mathrm{g}u>0} \mathcal{D}u\;\overline{\mathrm{H}}\left[\frac{\tilde{\mathrm{g}}uR}{\sqrt{1-R^2}}\right] \\
&\quad +(4\pi)^{-1}(1-R^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{4}\left[\frac{w-R\tilde{w}}{\sqrt{1-R^2}}\right]^2-\frac{\tilde{w}^2}{4}\right)\theta\left[\mathrm{g}w\right]\theta\left[\tilde{\mathrm{g}}\tilde{w}\right] \\
&\quad +\delta[w]\,(4\pi)^{-\frac{1}{2}}\exp\left(-\frac{\tilde{w}^2}{4}\right)\theta\left[\tilde{\mathrm{g}}\tilde{w}\right]\overline{\mathrm{H}}\left[\frac{\mathrm{g}\tilde{w}R}{\sqrt{2(1-R^2)}}\right] \\
&\quad +\delta[\tilde{w}]\,(4\pi)^{-\frac{1}{2}}\exp\left(-\frac{\tilde{w}^2}{4}\right)\theta\left[\mathrm{g}w\right]\overline{\mathrm{H}}\left[\frac{\tilde{\mathrm{g}}wR}{\sqrt{2(1-R^2)}}\right], \qquad (3.47)
\end{aligned}$$

which upon integrating with respect to the synaptic field w, gives the single probability distribution in the weights $\tilde{w}$

$$\begin{aligned}
\rho(\tilde{w}) \;&=\; \int \mathrm{d}w\;\rho(w,\tilde{w}) \\
&=\; \frac{1}{2}\delta[\tilde{w}]+(4\pi)^{-\frac{1}{2}}\exp\left(-\frac{\tilde{w}^2}{4}\right)\theta\left[\tilde{\mathrm{g}}\tilde{w}\right]. \qquad (3.48)
\end{aligned}$$

Hence the conditional probability for a synaptic field w given a field $\tilde{w}$ in the conjugate network is

$$\begin{aligned}
\rho(w\mid\tilde{w}) \;&=\; \rho(w,\tilde{w})\div\rho(\tilde{w}) \\
&=\; \delta[w]\;\overline{\mathrm{H}}\left[\frac{\mathrm{g}\tilde{w}R}{\sqrt{2(1-R^2)}}\right]+\frac{\theta[\mathrm{g}w]}{\sqrt{4\pi(1-R^2)}}\exp\left(-\frac{1}{4}\left[\frac{w-R\tilde{w}}{\sqrt{1-R^2}}\right]^2\right) \\
&\qquad (3.49)
\end{aligned}$$

for $\tilde{\mathrm{g}}\tilde{w}>0$.

Now that we have the conditional probabilities (3.41) and (3.49), we are in a position to tackle the paradox in §3.1.1. But first we need to find the conditions under which the calculations are valid, namely whether the replica-symmetric ansatz used is stable to small, local, replica-symmetry breaking fluctuations.

3.5 Stability of Replica Symmetry

The procedure for determining the stability of the replica-symmetric ansatz is the same as that used for the external field model in chapter 2. Once again we are probing the saddle-point solution with symmetric, weakly-asymmetric, and asymmetric fluctuations, and examining the determinants of the simplified stability matrices. We can avoid a lengthy argument over whether the signs of these determinants should be positive or negative definite by checking against any changes in sign, as these indicate a transition to or from replica-symmetric stability depending on our initial condition. The next two sections will give the results for symmetric and asymmetric fluctuations, since appendix §D.2 shows the symmetric and weakly-asymmetric fluctuations yield the same result.

3.5.1 Symmetric Fluctuations

The matrix (D.26) in §D.2.1 shows how the stability matrix simplifies into two non-zero diagonal blocks in the zero replica limit. Each of these blocks correspond to a particular network, hence if symmetric fluctuations are irrelevant in both networks we can say they are also unimportant for the calculation as a whole. The 'sub-determinants' for these two blocks are given by determinant (D.10) and for brevity we shall concentrate on one of these. From the moments (E.2) we can see the elements of this network's sub-determinant have the same quantities as in equations (2.37), bar an additional factor of one-half from the sign-constraint.

Moreover, since the training functions for the sign-constraint network and external field model are the same apart from the external field parameters, the matrix element in equation (2.39) also stands if we set the external field strength τ_{T} to zero.

We can write the analogous equation to expression (2.40) for the sub-determinant as

$$|\mathrm{Q}^{\mathrm{s}}| = 2\alpha_C(1-q)^2 \int\limits_{\hat{\lambda}(z)=\kappa} \mathcal{D}z \left[1 + 2\left(\frac{\hat{\lambda}(z)-z}{\sqrt{1-q}}\right)^2\right] - \frac{1}{4}(1-q)^2\left[2 + \frac{4}{1-q}\right] \quad (3.50)$$

and similarly for the conjugate network. This will not be written out because we can quickly see that in the optimal perceptron limit of (q and $\tilde{q} \to 1$) equation (3.50) goes to zero. However, as in §2.9.1, when the networks are just beneath the optimal limit this determinant evaluates to a constant regardless of the storage capacity. Hence we may conclude the replica-symmetric solution to the sign-constrained network is similarly stable to symmetric and weakly-asymmetric fluctuations.

3.5.2 Asymmetric Fluctuations

The null result from §3.5.1 hints that each of the two networks are stable to replica-symmetry breaking fluctuations as in the original Gardner model [Gar88]. We can then expect that nothing new in the way of replica-symmetry breaking is going to occur unless the inter-replica parameters r and K are included in the discussion, which is indeed the case for asymmetric fluctuations. Moreover, as seen by the resulting matrix (D.31) and its elements (D.32), we find the stability matrix is considerably simplified in the limit of zero replicas into a block diagonal form. Two of these blocks concern themselves with the stability of the intra-network parameters to asymmetric fluctuations, while the central block is concerned solely with the inter-network parameters.

STABILITY OF THE INTRA-NETWORK PARAMETERS

Restricting our attention to the intra-network parameters to one of the networks, we can modify the result from the external field model as in §3.5.1, and write the diagonal elements as

$$\begin{aligned} Q^A_{FF} - 2Q^B_{FF} + Q^C_{FF} &= -\int \mathcal{D}u\mathcal{D}v \left[\langle J\rangle_J^2 - \langle J^2\rangle_J\right]^2 \\ &= -\frac{1}{2(\mathrm{i}E+\mathrm{i}F)^2}, \\ Q^A_{qq} - 2Q^B_{qq} + Q^C_{qq} &= \alpha \int \mathcal{D}z \left[\langle x\rangle_\Lambda^2 - \langle x^2\rangle_\Lambda\right]^2, \\ &= \frac{\alpha}{(1-q)^2} \int_{\hat{\lambda}(z)=\kappa} \mathcal{D}z. \end{aligned} \tag{3.51}$$

In the optimal perceptron limit we can use the saddle-point solutions (3.31) and write the condition for the intra-network parameters to be stable to asymmetric fluctuations as

$$\alpha_C \int_{\hat{\lambda}(z)=\kappa} \mathcal{D}z < \frac{1}{2}, \qquad \text{and} \qquad \alpha_C \int_{\hat{\hat{\lambda}}(z)=\kappa} \mathcal{D}z < \frac{1}{2} \tag{3.52}$$

which since the integrals are bounded above by one-half, are only violated when the maximum storage capacity exceeds its maximum value of $\alpha_C = 1$ [AWC89a, CW90]. We hence conclude the intra-network parameters are stable to small, local replica-symmetry breaking fluctuations when within the maximum storage capacity.

Stability of the Inter-Network Parameters

The relevant sub-determinant for the inter-network parameters K_{ab} and r_{ab} is the central block marked out in matrix (D.31) with the elements given by

$$\begin{aligned} \partial^2 K &= Q^A_{KK} - (Q^B_{KK} + \tilde{Q}^B_{KK}) + Q^C_{KK}, \\ \partial^2 r &= Q^A_{rr} - (Q^B_{rr} + \tilde{Q}^B_{rr}) + Q^C_{rr}. \end{aligned} \tag{3.53}$$

These quantities arise from the permutation symmetries of the elements of the stability matrix and are defined as

$$\begin{aligned} Q^A_{KK} &\equiv \frac{\partial^2 G}{\partial K^2_{ab}}, & Q^A_{rr} &\equiv \frac{\partial^2 G}{\partial r^2_{ab}}, & &\forall a, b, \\ Q^B_{KK} &\equiv \frac{\partial^2 G}{\partial K_{ab} \partial K_{ac}}, & Q^B_{rr} &\equiv \frac{\partial^2 G}{\partial r_{ab} \partial K_{ac}}, & &\forall a, b, c\ (b \neq c), \\ \tilde{Q}^B_{KK} &\equiv \frac{\partial^2 G}{\partial K_{ab} \partial K_{cb}}, & \tilde{Q}^B_{rr} &\equiv \frac{\partial^2 G}{\partial r_{ab} \partial K_{cb}}, & &\forall a, b, c\ (a \neq c), \\ Q^C_{KK} &\equiv \frac{\partial^2 G}{\partial K_{ab} \partial K_{cd}}, & Q^C_{rr} &\equiv \frac{\partial^2 G}{\partial r_{ab} \partial K_{cd}}, & &\forall a, b, c, d\ (a \neq c, b \neq d), \end{aligned} \tag{3.54}$$

where the function $G(\cdots)$ being differentiated is as defined in equation (D.16). Using the above (3.54) and the angled-brackets notation defined in equations (3.17) and (3.18), we can write the elements (3.53) as

$$\begin{aligned} \partial^2 K &= (\mathrm{i})^2 \int \mathcal{D}u \mathcal{D}v \left\{ \frac{1}{2}(1 + m_s) \left[\langle J^2 \rangle_J \langle \tilde{J}^2 \rangle_{\tilde{J}+} - \langle J^2 \rangle_J \langle \tilde{J} \rangle^2_{\tilde{J}+} \right.\right. \\ &\qquad \left. - \langle J \rangle^2_J \langle \tilde{J}^2 \rangle_{\tilde{J}+} + \langle J \rangle^2_J \langle \tilde{J} \rangle^2_{\tilde{J}+} \right] \\ &\qquad + \frac{1}{2}(1 - m_s) \left[\langle J^2 \rangle_J \langle \tilde{J}^2 \rangle_{\tilde{J}-} - \langle J^2 \rangle_J \langle \tilde{J} \rangle^2_{\tilde{J}-} \right. \\ &\qquad \left.\left. - \langle J \rangle^2_J \langle \tilde{J}^2 \rangle_{\tilde{J}-} + \langle J \rangle^2_J \langle \tilde{J} \rangle^2_{\tilde{J}-} \right] \right\} \\ &= (\mathrm{i})^2 \int \mathcal{D}u \mathcal{D}v \left[\langle J^2 \rangle_J - \langle J \rangle^2_J \right] \left\{ \frac{1}{2}(1 + m_s) \left[\langle \tilde{J}^2 \rangle_{\tilde{J}+} - \langle \tilde{J} \rangle^2_{\tilde{J}+} \right] \right. \\ &\qquad \left. + \frac{1}{2}(1 - m_s) \left[\langle \tilde{J}^2 \rangle_{\tilde{J}-} - \langle \tilde{J} \rangle^2_{\tilde{J}-} \right] \right\}, \\ \partial^2 r &= \alpha \int \mathcal{D}y \mathcal{D}z \left[\langle x^2 \rangle_\Lambda - \langle x \rangle^2_\Lambda \right] \left[\langle \tilde{x}^2 \rangle_{\tilde{\Lambda}} - \langle \tilde{x} \rangle^2_\Lambda \right]. \end{aligned} \tag{3.55}$$

The appropriate first and second moments for J and $\tilde{J}$ may be found in equations (E.2) in the appendices, and give

$$
\begin{aligned}
\partial^2 K &= \frac{(\mathrm{i})^2}{\mathrm{i}E+\mathrm{i}F}\left[\frac{1}{2}(1+m_s)\int\limits_{\substack{\mathrm{A}u+\mathrm{B}v>0\\ \mathrm{A}u-\mathrm{B}v>0}} \mathcal{D}u\mathcal{D}v + \frac{1}{2}(1-m_s)\int\limits_{\substack{\mathrm{A}u+\mathrm{B}v>0\\ \mathrm{A}u-\mathrm{B}v<0}} \mathcal{D}u\mathcal{D}v\right]\frac{1}{\mathrm{i}\tilde{E}+\mathrm{i}\tilde{F}} \\
&= \frac{(\mathrm{i})^2}{2\pi(\mathrm{i}E+\mathrm{i}F)(\mathrm{i}\tilde{E}+\mathrm{i}\tilde{F})}\left[2m_s\Phi+\frac{1}{2}(1-m_s)\pi\right], \qquad (3.56)
\end{aligned}
$$

by using the same plane-polar co-ordinates as in the calculation of equations (3.32), the saddle-point solution for the inter-network order parameter r. For the other diagonal element $\partial^2 r$ we require the quantities given by first moments (3.29) and the following second moments

$$
\begin{aligned}
\langle x^2\rangle_\Lambda &= \frac{1}{(1-q)}\left[1-\left(\frac{\hat{\lambda}(y,z)-(\mathrm{a}y+\mathrm{b}z)}{\sqrt{1-q}}\right)-\frac{1}{1-\gamma^2 g''(\hat{\lambda}(y,z))/2}\right], \\
\langle \tilde{x}^2\rangle_{\tilde{\Lambda}} &= \frac{1}{(1-\tilde{q})}\left[1-\left(\frac{\hat{\hat{\lambda}}(y,z)-(\mathrm{a}y-\mathrm{b}z)}{\sqrt{1-\tilde{q}}}\right)-\frac{1}{1-\tilde{\gamma}^2 \tilde{g}''(\hat{\hat{\lambda}}(y,z))/2}\right],
\end{aligned}
\qquad (3.57)
$$

where the training functions $g(\lambda)$ and $\tilde{g}(\lambda)$ are as defined in equations (3.34), and the parameters γ and $\tilde{\gamma}$ in equations (3.28). Substituting the above expressions (3.57) into the equation (3.55) for $\partial^2 r$, we find the integrand is only non-zero over a certain integration range, giving in the zero error storage limit of $(\gamma \to \infty, \tilde{\gamma} \to \infty)$

$$
\partial^2 r = \frac{\alpha}{(1-q)(q-\tilde{q})}\int\limits_{\substack{\mathrm{a}y+\mathrm{b}z<\kappa\\ \mathrm{a}y-\mathrm{b}z<\kappa}} \mathcal{D}y\mathcal{D}z. \qquad (3.58)
$$

Taking the evaluated equations (3.56) and simplifying equation (3.58) by moving over to plane-polar co-ordinates, we find in the limit of the optimal perceptrons the condition for stability of the $K_{ab} = K, r_{ab} = r$ replica-symmetric ansatz is

$$1 > \frac{4\alpha_C}{2\pi} \left\{ 2m_s\Phi + \frac{1}{2}(1 - m_s)\pi \right\} \left\{ \frac{2\phi}{2\pi} \exp\left(-\frac{\kappa^2}{2\mathrm{a}^2} \right) + \int_{-\kappa}^{\kappa} \mathcal{D}x \; \overline{\mathrm{H}} \left[\sqrt{\frac{\kappa^2}{\mathrm{a}^2} - x^2} \right] \right\}, \tag{3.59}$$

that is, equation (3.59) must be satisfied for the central sub-determinant of matrix (D.31) to be positive definite. This condition can be readily shown to be the correct one for zero storage at $\alpha_C = 0$ where there is no source of disorder in the problem and, presumably, replica-symmetry yields the correct result.

We can easily numerically evaluate equation (3.59) to monitor the validity of our results. This was done for the results to be presented, and it transpires that so long as one is within the upper storage capacity of $\alpha_C \leq 1$, then stability of the replica-symmetric ansatz is respected.

3.6 The Inter-Network Order Parameters

In §3.3 and §3.4 we have calculated the conditional probability distributions for the alignment and synaptic fields. A cursory glance at these two distributions indicates the inter-network correlation parameter r defined in equations (3.6) and its conjugate R have an importance elevated above being mathematical artefacts.

The definition of r is as a measure of the correlation in connections between the two networks we wish to compare, so its saddle-point solution given in equation (3.32) is its thermal[1] and quenched averaged observable. Hence, we are justified in

[1] In the sense of a Boltzmann weighted annealed average over the phase-space of interactions.

calling the saddle-point solution of r an actual physical order parameter on the same footing as (for example) the Edwards-Anderson order parameter discussed in appendix B. Another interpretation of the parameter r is given by the conditional probability for the alignment fields in equation (3.41). If we gloss over the rôles of the delta and theta functions, we can say r is the most probable ratio of the two alignment field values $\Lambda/\tilde{\Lambda}$ given the field $\tilde{\Lambda}$. So in this sense r relates the alignment field distributions between the two networks. It is also responsible for the width of the Gaussian contribution, with the Dirac delta function limit reached when $r = \pm 1$. We shall have more to say on this when we return to the paradox.

The origins of the other saddle-point parameter R is in equation (3.33), where we find it is simply a rescaling of K, a quantity conjugate to r arising from its introduction in the calculations. However, as with r, the conditional probability function of equation (3.49) allows us an attempt at an interpretation. This conditional probability tells us that given a conjugate field equal to $\tilde{w}$, the most probable value for the weight w peaks at $R\tilde{w}$ —again upon being somewhat cavalier with the delta and theta functions. This quantity R also determines the width of the Gaussian part of the distribution, which reduces to a delta function in the limit of $R = \pm 1$.

THE PARADOX

We shall now limit ourselves to the case discussed in §3.1.1, namely where the correlation (3.3) between the weight signs is $m_s = -1$. The naïve view is that this results in a complete anti-correlation between the networks, that the inter-network order parameter takes on the value $r = -1$. If this is true, the conditional probability distribution for the alignment field (3.41) will have its Gaussian compressed into a delta function at $\Lambda = -\tilde{\Lambda}$. Hence it is impossible for both networks to have their stability fields larger than zero, that is for both networks to be anything but marginally stable to the same patterns. That then is the case if the two networks are simply related by $J^j = -\tilde{J}^j$ for all j, and with the inter-network order parameter $r = -1$.

Figure 3.1 is a plot of the saddle-point solution of the order parameter r, which measures the correlation between the two networks. The plots are against increasing optimal storage capacities, for nine values of the overlap between the networks' sign constraints. The particular case discussed in the above paradox is highlighted by a broken line, and we can see straight away that the naïve expectation of complete anti-correlation between the networks never occurs. Instead it starts with a negligible correlation at zero storage capacity and decreases to $r = -0.22$ at maximum storage; which by equation (3.37) corresponds to one starting with an infinite minimal stability constant which then decreases to zero. Increasing the weight sign overlap above $m_s = -1$ brings the networks closer to the limit of identical constraints at $m_s = 1$. At $m_s = 1$ the two networks are identical and, as expected, are fully correlated with each other by the inter-network order parameter $r = 1$.

The result that the inter-network order parameter never reaches $r = -1$ even in the $m_s = -1$ limit also affects the conditional probability of the alignment fields. From equation (3.41) we see this provides a finite width to the Gaussian part of the distribution, and hence there is a finite probability the alignment field in one network satisfies the minimum stability requirement $\Lambda > \kappa$, for any corresponding field $\tilde{\Lambda} > \kappa$ in the conjugate network. Balanced against this, we find the first network is not able to achieve the same stability, for the conditional probability does not peak at $\Lambda = \tilde{\Lambda}$, but instead tails off after $\Lambda > \kappa$.

Figure 3.2 gives the corresponding plot for the order parameter conjugate to the inter-network order parameter. Once again we are plotting for increasing optimal storage capacities α_C, for different weight sign overlaps m_s. Unlike its counterpart it is not immediately obvious whether this conjugate order parameter R is also an overlap measure of some quantity between the networks. What we can say is that

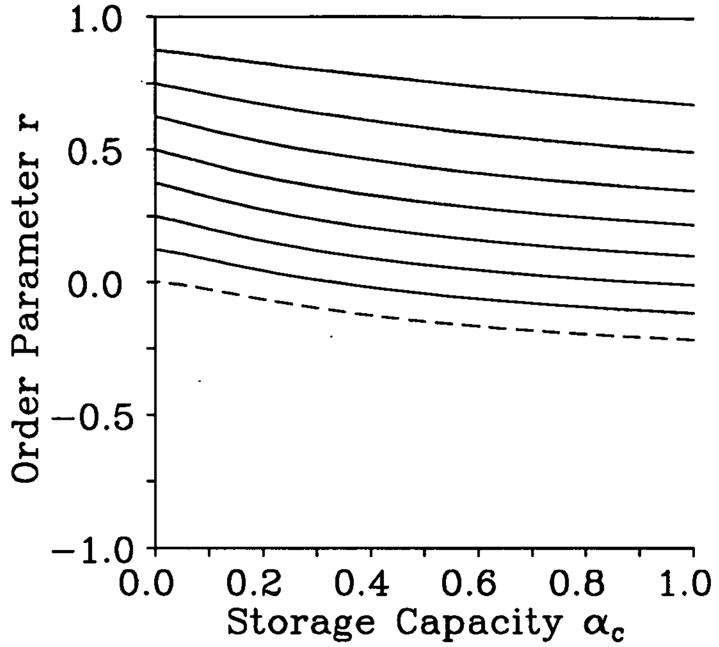

Figure 3.1: Saddle-point solution of the inter-network order parameter r. This order parameter measures the correlation between the two networks, and is plotted here against increasing optimal storage capacity α_C. This is done for different overlaps between the sign-constraints beginning with m_s = -1.0 in dashed lines, up to $m_s = -0.75, -0.5, 0.0, 0.25, 0.5, 0.75$ and 1.00.

since the saddle-point equations never solves to $R = -1$ the Gaussian part to the conditional probability distribution (3.49) for a sign overlap $m_s < 1.0$ will have a finite contribution at $gw > 0$, given $\tilde{g}\tilde{w} > 0$ and $\alpha_C > 0$. Hence for non-zero storage and the paradox's scenario of complete antisymmetry $m_s = -1$ between the networks' sign constraints, the most probable value of weight w given $\tilde{w}$ is not single valued by a delta function at $w = -\tilde{w}$. This bears out the results in figure 3.1 where we have already found the two networks are not related by a simple sign change.

For completeness figures 3.3 and 3.4 plot the inter-network order parameter r and

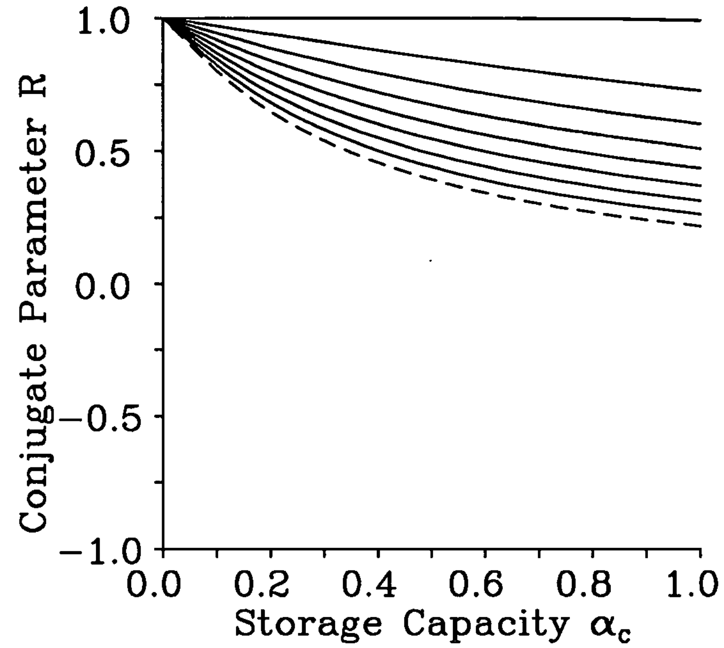

Figure 3.2: Saddle-point solution of the inter-network order parameter R. This order parameter's most obvious physical significance is in determining the most probable weight value w given $\tilde{w}$, as given by the conditional probability distribution for the weights. The plots are for overlaps between the two networks' sign-constraints m_s = -1.0 (in dashed lines) up to $m_s = -0.75, -0.5, 0.0, 0.25, 0.5, 0.75$ and 1.00. These plots are drawn against increasing optimal storage capacity α_C, or equivalently a decreasing minimum stability constant κ.

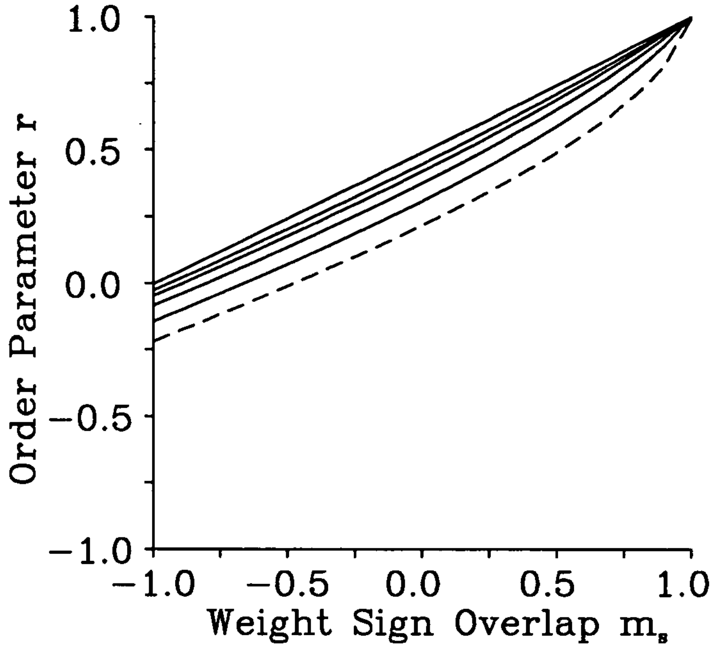

Figure 3.3: Saddle-point solution of the inter-network order parameter r, for overlaps between the sign-constraints $m_s = -1 \ldots 1$. The plots are for increasing minimum stability constant $\kappa =$ and a corresponding decrease in the optimal storage capacity α_C: starting in dashed lines with $(\kappa = 0.0; \alpha_C = 1.0)$, up to $(0.5; 0.48)$, $(1.01; 0.26)$, $(1.51; 0.15)$ and $(5.0; 0.02)$.

its conjugate R against the sign overlap m_s, at different values of the minimum stability constant. The main point of note is that they form a steady convergence to r and $R = 1$ as the two network constraints become identical at $m_s = 1$. The differences between the plots are increased by a low minimum stability, or equivalently, by a high storage capacity.

We can elegantly illustrate the changes in the network's weights as the overlap m_s in the sign-constraints is varied by means of a projection onto a sphere. The idea is to say something about how 'typical' a trained network with a weight vector $\boldsymbol{J}$ is in the phase-space of connections. The inter-network order parameter

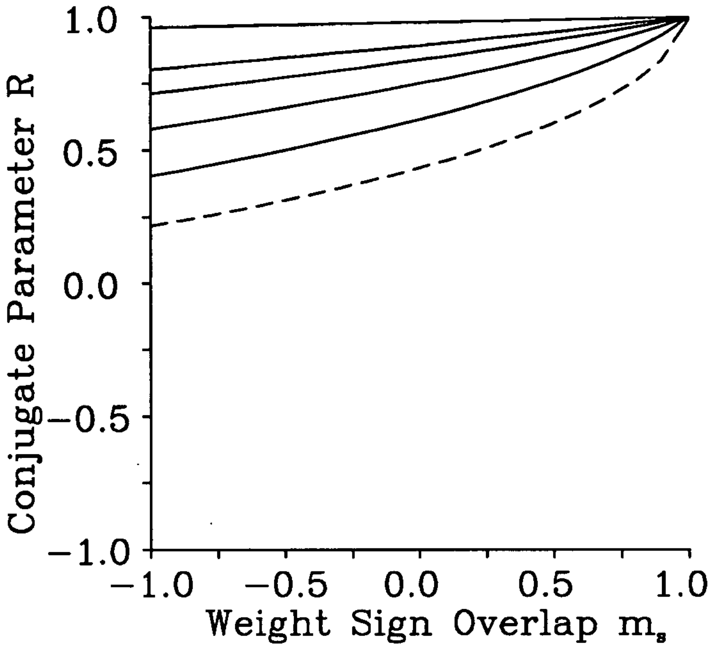

Figure 3.4: Saddle-point solution of the inter-network order parameter R. The plots are against overlaps between the two networks' sign-constraints $m_s = -1.0 \ldots 1$, for $(\kappa = 0.0; \alpha_C = 1.0)$ in dashed lines, up to $(0.5; 0.48)$, $(1.01; 0.26)$, $(1.51; 0.15)$ and $(5.0; 0.02)$.

---

r measures the overlap between two networks whose sign constraints differ by their sign overlap m_s, so from figure 3.3 we see that at $m_s = 1$ the two networks have identical weights $\boldsymbol{J} = \tilde{\boldsymbol{J}}$, whilst differing the most at $m_s = -1$. We can explicitly reveal the dependence of the inter-network order parameter on the sign constraint overlap, by writing $r(m_s)$. We shall also fix the conjugate network's weights and centre our attention on the response of the other network as m_s is varied, explicitly done by writing $\boldsymbol{J}(m_s)$. Hence this network's trained weights are bounded to lie in the phase-space of interactions between the vectors $\boldsymbol{J}(m_s = 1)$ and $\boldsymbol{J}(m_s = -1)$.

We can now elaborate on what we meant earlier by 'typical'. The network is deemed typical if its weight $\boldsymbol{J}(m_s)$ lies along the planar region bounded by the vectors $\boldsymbol{J}(m_s = 1)$ and $\boldsymbol{J}(m_s = -1)$. The quantities $r(m_s)$ and $r(-m_s)$ are the overlap of $\boldsymbol{J}(m_s)$ with $\boldsymbol{J}(m_s = 1)$ and $\boldsymbol{J}(m_s = -1)$ respectively, so $\boldsymbol{J}(m_s)$ will be displaced outside this plane if the angle $[\cos^{-1} r(m_s) + \cos^{-1} r(-m_s)]$ is not equal to the planar angle $\cos^{-1} r(m_s = -1)$ between the bounded region. We can plot how far $\boldsymbol{J}(m_s)$ is displaced outside the plane by mapping it onto a 'world map' with longitude and latitude co-ordinates, as shown in figure 3.5. The points show the trajectories of $r(m_s)$ and $r(-m_s)$ as the sign constraint overlap is varied, for five values of minimum stabilities and storages. We can see that the weight vectors indeed occupy regions of phase-space outside the plane defined by the $m_s = \pm 1$ limits. Moreover these points are uniformly distributed along the longitudinal axis, so we can say the network connections which satisfy the constraints of equations (1.12), (3.1) and (3.2) are evenly placed in the phase-space of interactions.

3.7 Discussion

This chapter motivates the use of the dual distribution functions introduced by Wong [Won90], by investigating a possible paradox in networks where the interaction weights are constrained to a prescribed sign. For this, we consider two networks with differing sign-constraints and calculate the dual distribution functions for their alignment field and synaptic efficacy. This is a replica calculation with quenched disorder in the memory patterns, with the neural connections annealed-optimised in the phase-space of interactions. The stability of these calculations to small, local replica-symmetry breaking fluctuations around the mean field saddle-point is considered. We find that as with the single network case of chapter 2, only the asymmetric fluctuations mode is of any importance. This mode introduces a criterion concerned with the *inter-network* quantities to be satisfied, in addition to the *intra-network* criteria already known. Fortunately none of these

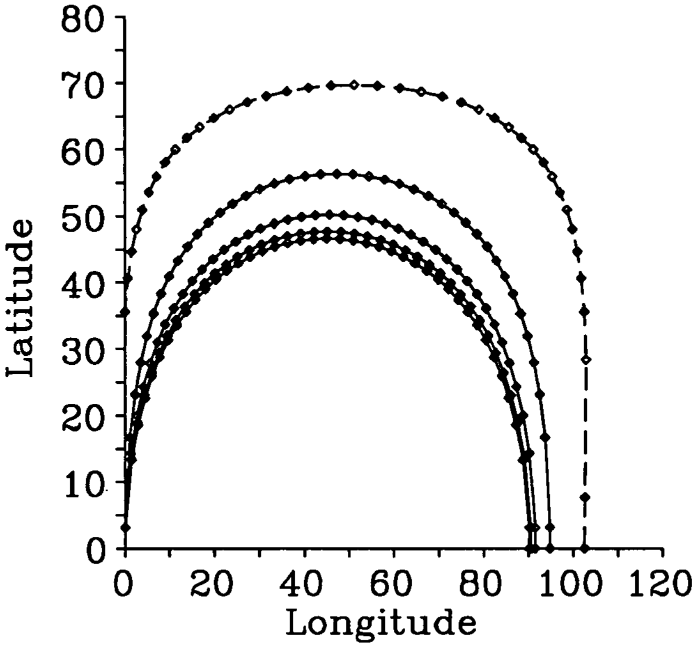

Figure 3.5: Projection of the weight vector $\boldsymbol{J}$ onto a 3-dimensional sphere, as represented by longitude and latitude co-ordinates. The diamonds mark the solutions as the sign overlap m_s is linearly varied from $m_s = 1$ at the graph's western origin to $m_s = -1$ towards the eastern end. The four plots are for the minimum stability constants and optimal storage capacities pairs of: $(\kappa = 0.0; \alpha_C = 1.0)$ (in dashed-lines), (1.0; 0.26) (2.0; 0.10), (3.0; 0.05), and (4.0; 0.03).

criteria are violated for the results resolving the said paradox. In this we find that contrary to the naïve expectations of the paradox, two networks with opposing sign-constraints in their weights but storing the same set of patterns do not have their annealed-optimised synapses simply anti-correlated with each other.

In conclusion, we have given an example of the use of dual distribution functions in elucidating the behaviour of networks optimised in the phase-space of interactions. This is necessary if we cannot analytically tell what the network's synaptic values are, but where a comparison with a 'known' network can still prove useful.

The example here is for two networks with differing sign-constraints. Work has also been done where the networks differ in the noisiness of the training function [WRS91], recovering in the low noise limit the optimal Gardner [Gar88] network, and in the high noise limit the Hopfield-Hebb [Hop82, AGS85] model. One can conceivably extend this programme for networks with other forms of weight constraints such as a Gaussian distribution, or perhaps with different errors in the storage [GD88, AEHW90]. The introduction by Gardner of the phase-space technique has been quickly followed by a proliferation of variations on her model, and the use of dual distribution functions promises a further understanding of their similarities and differences.

Appendix A

Mathematical Notations and Identities

A.1 Gaussian Measure

This is a common notational device to simplify writing down integral measures with a Gaussian weighting term. For an integration with respect to a variable x we define

$$\mathcal{D}x \equiv \frac{\mathrm{d}x}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right). \tag{A.1}$$

A.2 Hubbard-Stratonovich Identity

This identity is often used to 'linearise' the argument of an exponential function. Using the Gaussian measure notation just introduced it is

$$\exp\left(\frac{1}{2}a^2\right) = \int_{-\infty}^{\infty} \mathcal{D}x \exp(\pm ax). \tag{A.2}$$

A.3 Error Functions

The conventional definitions of the error function and its complement are

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{d}t \exp(-t^2), \qquad \mathrm{erfc}(x) \equiv 1 - \mathrm{erf}(x) \tag{A.3}$$

but the functions more commonly encountered here are its Gaussian equivalents defined by

$$\begin{aligned} \mathrm{H}(x) \equiv \int_x^\infty \mathcal{D}t, \qquad \overline{\mathrm{H}}(x) &\equiv 1 - \mathrm{H}(x) \\ &= \mathrm{H}(-x). \end{aligned} \tag{A.4}$$

In the large argument limit these functions can be expanded via

$$\mathrm{H}(x) \approx \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)\left[1 - \frac{1}{x^2} + \frac{3}{x^4} + \cdots\right], \qquad \text{for } x \to \infty. \tag{A.5}$$

A.4 Integral Form of the Delta-function

We can write the Dirac delta functions by its Fourier transform, integrating along either the real number line or equivalently along the imaginary line

$$\delta[a-b] = \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{2\pi} \exp(\pm \mathrm{i}x[a-b]), \qquad = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{\mathrm{d}x}{\mathrm{i}2\pi} \exp(\pm x[a-b]). \tag{A.6}$$

An example where this integral representation is used is in 'extracting' out a known quantity from a function which we wish to keep general, for example

$$\begin{aligned} f(\Lambda) &= \int_{-\infty}^{\infty} \frac{\mathrm{d}\lambda}{2\pi} \delta[\lambda - \Lambda] f(\lambda) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\mathrm{d}\lambda \mathrm{d}x}{2\pi} \exp\left(\mathrm{i}x(\lambda - \Lambda)\right) f(\lambda). \end{aligned} \tag{A.7}$$

Appendix B

The Physical Interpretation of the γ Parameter

This is a discussion on how the parameter defined in equation (2.23) may be given a physical interpretation, as mentioned in reference [WRS91]. The quantity in question was first introduced by Gardner and Derrida [GD88] to allow the simultaneous limits of the inverse annealed temperature β and the inter-replica order parameter q to be sensibly taken. That is we define

$$\gamma^2 \equiv \lim_{\substack{\beta \to \infty \\ q \to 1}} 2\beta(1-q) \tag{B.1}$$

such that γ is always finite.

The order parameter q is the replica-symmetric saddle-point of the quantity

$$q_{ab} \equiv \frac{1}{C} \sum_{j}^{C} J_a^j J_b^j$$

defined in equation (1.11), which measures the correlations between the replicated weights. This quantity q can be seen to be an incarnation of the Edwards-Anderson spin-glass order parameter q_{EA}, and it is with this that the conceptual link of the interpretation depends upon. In the notation of an Ising spin system with $j = 1 \ldots N$ sites $S_j = \pm 1$, the Edwards-Anderson order parameter is defined as

$$q_{EA} \equiv \left\langle\left\langle \frac{1}{N} \sum_j^N \langle S_j \rangle^2 \right\rangle\right\rangle$$

where the double angled brackets denote an average over the quenched disorder, and the single angled brackets a thermal average over the spin configurations.

The local susceptibility for a magnetic spin model can be written in correlation form [Moo84] as

$$\chi_S = \left\langle\left\langle \frac{\beta}{N} \sum_{ij} [\langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle] \right\rangle\right\rangle$$

which for binary spin sites and zero correlations between the sites becomes

$$\chi_S = \left\langle\left\langle \frac{\beta}{N} \sum_j \left[\langle S_j^2 \rangle - \langle S_j \rangle^2\right] \right\rangle\right\rangle, \qquad = \beta(1 - q_{EA}),$$

via self-averaging.

Hence by analogy with equation (B.1), we can write for the phase-space of interactions problem

$$\gamma^2 = \lim_{\substack{\beta \to \infty \\ q \to 1}} \{2\chi_J\} \tag{B.2}$$

where we shall define

$$\chi_J \equiv \left\langle\left\langle \frac{\beta}{N} \sum_j \left[\langle J_j^2 \rangle - \langle J_j \rangle^2\right] \right\rangle\right\rangle \tag{B.3}$$

and loosely call *it* the local susceptibility. This can actually be verified by using the notations defined in (2.18) and (2.19) to show that equation (B.3) evaluates to

$$
\begin{aligned}
\chi_J &= \beta\left[\left\{J_a^2\right\}_{G_J} - \left\{J_a J_b\right\}_{G_J}\right] \\
&= \beta \int \mathcal{D}u \left[\langle J^2\rangle_J - \langle J\rangle_J^2\right] \\
&= \frac{\beta}{\mathrm{i}E + \mathrm{i}F} \int \mathcal{D}u \\
&= \beta(1-q)
\end{aligned}
$$

where the first and second moments of J are given by equations (2.21), and $(\mathrm{i}E + \mathrm{i}F)^{-1} = (1-q)$ at the saddle-point.

To conclude, we have seen how the parameter γ can be expressed in terms of the phase-space of interactions' analogue of the local susceptibility of magnetic spin systems. The conceptual bridge comes through both models having an Edwards-Anderson order parameter, which we can use in the correlation form of the local susceptibility. The somewhat implicit line of reasoning involved leaves open the question of whether we are justified in calling χ_J the local susceptibility in the traditional spin-system sense. Whatever our reservations, equation (B.3) does give an insight into the correlations amongst the connections which satisfy all the constraints. That is, χ_J is the variance of the region of interaction phase-space which satisfies the training function, weighted by the inverse annealed temperature.

In the zero storage error case of $\mathcal{F} = 0$, γ diverges, which means the variance of the weight-space is shrinking at a slower rate than the annealed temperature going to zero. Reference [WRS91] builds on this to interpret the solution of interactions as having a finite —indeed wide— variance slightly above the ground state. With a storage error $\mathcal{F} > 0$ we find γ is finite [GD88], so the width of solutions in weight-space now tends to zero proportionally with the temperature.

Appendix C

Maximising the Function $\mathcal{G}(\lambda)$ for the External Field Model

This appendix provides more detailed treatments to some of the calculations described in §2.6 on maximising the training function. We shall make use of the quantities of the external field model defined throughout chapter 2.

C.1 Ordering the Transition Points

By solving the function $\mathcal{G}(\lambda)$ defined in equation (2.27) at the three possible values of $\hat{\lambda}(z) = \{(\kappa - \tau_{\mathrm{T}}), (\kappa + \tau_{\mathrm{T}}), z\}$, we find there are six possible points along z where $\hat{\lambda}(z)$ may change value, namely at

$$\begin{aligned} z \;=\; & \kappa - \tau_{\mathrm{T}},\; \kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}},\; \kappa + \tau_{\mathrm{T}}, \\ & \kappa + \tau_{\mathrm{T}} - \gamma,\; \kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}},\; \text{and } \kappa - \frac{\gamma^2 f}{4\tau_{\mathrm{T}}}. \end{aligned} \qquad \text{(C.1)}$$

To determine in what order these *transition points* occur at as z is varied requires $(5 + 4 + 3 + 2 + 1) = 15$ comparisons to be made, which give

Order of Transition Points	Condition
$\kappa + \tau_{\mathrm{T}} > \kappa - \tau_{\mathrm{T}}$	always,
$\kappa + \tau_{\mathrm{T}} > \kappa + \tau_{\mathrm{T}} - \gamma$	always,
$\kappa + \tau_{\mathrm{T}} > \kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	always,
$\kappa + \tau_{\mathrm{T}} > \kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}}$	always,
$\kappa + \tau_{\mathrm{T}} > \kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	always,
$\kappa - \tau_{\mathrm{T}} > \kappa + \tau_{\mathrm{T}} - \gamma$	$2\tau_{\mathrm{T}} < \gamma$,
$\kappa - \tau_{\mathrm{T}} > \kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	$2\tau_{\mathrm{T}} < \gamma\sqrt{f_{\mathrm{T}}}$,
$\kappa - \tau_{\mathrm{T}} > \kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}}$	never,
$\kappa - \tau_{\mathrm{T}} > \kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$2\tau_{\mathrm{T}} < \gamma\sqrt{f_{\mathrm{T}}}$,
$\kappa + \tau_{\mathrm{T}} - \gamma > \kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	never,
$\kappa + \tau_{\mathrm{T}} - \gamma > \kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}}$	$2\tau_{\mathrm{T}} > \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right]$,
$\kappa + \tau_{\mathrm{T}} - \gamma > \kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$2\tau_{\mathrm{T}} < \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right]$ and for $2\tau_{\mathrm{T}} > \gamma\left[1 + \sqrt{1 - f_{\mathrm{T}}}\right]$,
$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}} > \kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}}$	$2\tau_{\mathrm{T}} > \gamma\left[\sqrt{f_{\mathrm{T}}} - \sqrt{1 - f_{\mathrm{T}}}\right]$,
$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}} > \kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	always,
$\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}} > \kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$2\tau_{\mathrm{T}} < \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right]$.

(C.2)

The conditions in table (C.2) clearly show the transition points are uniquely ordered as the training field strength $2\tau_{\mathrm{T}}$ lies in one of the six ranges

$$\begin{aligned} -\infty < \gamma\left[\sqrt{f_{\mathrm{T}}} - \sqrt{1 - f_{\mathrm{T}}}\right] &< \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right] \\ &< \gamma\sqrt{f_{\mathrm{T}}} < \gamma < \gamma\left[1 + \sqrt{1 - f_{\mathrm{T}}}\right] < \infty. \end{aligned} \tag{C.3}$$

That is, as the field strength is varied through the ranges (C.3), $\hat{\lambda}(z)$ may take on one of six forms for us to investigate.

C.2 Value of $\mathcal{G}(\lambda)$ at the Transition Points

The function $\hat{\lambda}(z)$ maximises the function $\mathcal{G}(\lambda)$ defined in equation (2.27), and may take on one of three values: $\hat{\lambda}(z) = \{(\kappa - \tau_{\mathrm{T}}), (\kappa + \tau_{\mathrm{T}})$ or $z\}$. We now consider which of these three is the maximum at the six transition points of z given in equation (C.1). This is answered by comparing the value of the function $\mathcal{G}(\lambda)$ at the three values of $\hat{\lambda}(z)$, for each of the transition points. Once again we find a dependence on the parameters which can be concisely expressed around the field strength τ_{T}. For z at the following transition points, the correct values of $\hat{\lambda}(z)$ are

Transition	$\hat{\lambda}(z)$
$\kappa - \tau_{\mathrm{T}}$	$\kappa - \tau_{\mathrm{T}}$ & z when $2\tau_{\mathrm{T}} > \gamma\sqrt{f_{\mathrm{T}}}$, $\kappa + \tau_{\mathrm{T}}$ when $2\tau_{\mathrm{T}} < \gamma\sqrt{f_{\mathrm{T}}}$.
$\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}}$	$\kappa - \tau_{\mathrm{T}}$ & z when $2\tau_{\mathrm{T}} > \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right]$, $\kappa + \tau_{\mathrm{T}}$ when $2\tau_{\mathrm{T}} < \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right]$.
$\kappa + \tau_{\mathrm{T}}$	$\kappa + \tau_{\mathrm{T}}$ & z always.
$\kappa + \tau_{\mathrm{T}} - \gamma$	$\kappa + \tau_{\mathrm{T}}$ & z when $2\tau_{\mathrm{T}} < \gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right]$, $\kappa - \tau_{\mathrm{T}}$ when $\gamma\left[1 - \sqrt{1 - f_{\mathrm{T}}}\right] < 2\tau_{\mathrm{T}} < \gamma$, z when $2\tau_{\mathrm{T}} > \gamma$.
$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}}$ when $2\tau_{\mathrm{T}} < \gamma\sqrt{f_{\mathrm{T}}}$, $\kappa + \tau_{\mathrm{T}}$ & z when $2\tau_{\mathrm{T}} > \gamma\sqrt{f_{\mathrm{T}}}$.
$\kappa - \frac{\gamma^2 f}{4\tau_{\mathrm{T}}}$	z when $2\tau_{\mathrm{T}} < \gamma\left[1 - \sqrt{f_{\mathrm{T}}}\right]$, $\kappa - \tau_{\mathrm{T}}$ & $\kappa + \tau_{\mathrm{T}}$ when $\gamma\left[1 - \sqrt{f_{\mathrm{T}}}\right] < 2\tau_{\mathrm{T}} < \gamma\sqrt{f_{\mathrm{T}}}$, z when $2\tau_{\mathrm{T}} > \gamma\sqrt{f_{\mathrm{T}}}$.

(C.4)

C.3 Reduction to Three Regimes

The final task is to combine all the above results. That is, for each of the six parameter ranges given in equation (C.3) we wish to write down the function $\hat{\lambda}(z)$ itself. This function changes at the six possible transition points given by expression (C.1), with its ordering along the z line given by table (C.2). Hence, with table (C.4) which lists $\hat{\lambda}(z)$ at the transition points, we can finally map out the maximising function $\hat{\lambda}(z)$ throughout the entire z range, for the six identified ranges of the external field strength.

$$2\tau_{\mathrm{T}} < \gamma\left[\sqrt{f_{\mathrm{T}}} - \sqrt{1-f_{\mathrm{T}}}\right]$$

z range		$\hat{\lambda}(z)$
$-\infty$	$\kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	z
$\kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}} - \gamma$	z
$\kappa + \tau_{\mathrm{T}} - \gamma$	$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	$\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1-f_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1-f_{\mathrm{T}}}$	$\kappa - \tau_{\mathrm{T}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa - \tau_{\mathrm{T}}$	$\kappa + \tau_{\mathrm{T}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa + \tau_{\mathrm{T}}$	∞	z

(C.5)

$$\gamma\left[\sqrt{f_{\mathrm{T}}} - \sqrt{1-f_{\mathrm{T}}}\right] < 2\tau_{\mathrm{T}} < \gamma\left[1 - \sqrt{1-f_{\mathrm{T}}}\right]$$

z range		$\hat{\lambda}(z)$
$-\infty$	$\kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	z
$\kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}} - \gamma$	z
$\kappa + \tau_{\mathrm{T}} - \gamma$	$\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1-f_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1-f_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	$\kappa - \tau_{\mathrm{T}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa - \tau_{\mathrm{T}}$	$\kappa + \tau_{\mathrm{T}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa + \tau_{\mathrm{T}}$	∞	z

(C.6)

$$\gamma\left[1-\sqrt{1-f_{\mathrm{T}}}\right] < 2\tau_{\mathrm{T}} < \gamma\sqrt{f_{\mathrm{T}}}$$

z range		$\hat{\lambda}(z)$
$-\infty$	$\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}$	z
$\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}$	$\kappa+\tau_{\mathrm{T}}-\gamma$	$\kappa-\tau_{\mathrm{T}}$
$\kappa+\tau_{\mathrm{T}}-\gamma$	$\kappa-\frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$\kappa-\tau_{\mathrm{T}}$
$\kappa-\frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$\kappa+\tau_{\mathrm{T}}-\gamma\sqrt{f_{\mathrm{T}}}$	$\kappa+\tau_{\mathrm{T}}$
$\kappa+\tau_{\mathrm{T}}-\gamma\sqrt{f_{\mathrm{T}}}$	$\kappa-\tau_{\mathrm{T}}$	$\kappa+\tau_{\mathrm{T}}$
$\kappa-\tau_{\mathrm{T}}$	$\kappa+\tau_{\mathrm{T}}$	$\kappa+\tau_{\mathrm{T}}$
$\kappa+\tau_{\mathrm{T}}$	∞	z

(C.7)

$$\gamma\sqrt{f_{\mathrm{T}}} < 2\tau_{\mathrm{T}} < \gamma$$

z range		$\hat{\lambda}(z)$
$-\infty$	$\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}$	z
$\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}$	$\kappa+\tau_{\mathrm{T}}-\gamma$	$\kappa-\tau_{\mathrm{T}}$
$\kappa+\tau_{\mathrm{T}}-\gamma$	$\kappa-\tau_{\mathrm{T}}$	$\kappa-\tau_{\mathrm{T}}$
$\kappa-\tau_{\mathrm{T}}$	$\kappa-\frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	z
$\kappa-\frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$\kappa+\tau_{\mathrm{T}}-\gamma\sqrt{f_{\mathrm{T}}}$	z
$\kappa+\tau_{\mathrm{T}}-\gamma\sqrt{f_{\mathrm{T}}}$	$\kappa+\tau_{\mathrm{T}}$	$\kappa+\tau_{\mathrm{T}}$
$\kappa+\tau_{\mathrm{T}}$	∞	z

(C.8)

$$\gamma < 2\tau_{\mathrm{T}} < \gamma\left[1+\sqrt{1-f_{\mathrm{T}}}\right]$$

z range		$\hat{\lambda}(z)$
$-\infty$	$\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}$	z
$\kappa-\tau_{\mathrm{T}}-\gamma\sqrt{1-f_{\mathrm{T}}}$	$\kappa-\tau_{\mathrm{T}}$	$\kappa-\tau_{\mathrm{T}}$
$\kappa-\tau_{\mathrm{T}}$	$\kappa+\tau_{\mathrm{T}}-\gamma$	z
$\kappa+\tau_{\mathrm{T}}-\gamma$	$\kappa-\frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	z
$\kappa-\frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$\kappa+\tau_{\mathrm{T}}\gamma\sqrt{f_{\mathrm{T}}}$	z
$\kappa+\tau_{\mathrm{T}}\gamma\sqrt{f_{\mathrm{T}}}$	$\kappa+\tau_{\mathrm{T}}\kappa+\tau_{\mathrm{T}}$	$\kappa+\tau_{\mathrm{T}}$
$\kappa+\tau_{\mathrm{T}}$	∞	z

(C.9)

and

$$2\tau_{\mathrm{T}} > \gamma\left[1 + \sqrt{1 - f_{\mathrm{T}}}\right]$$

z range		$\hat{\lambda}(z)$
$-\infty$	$\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}}$	z
$\kappa - \tau_{\mathrm{T}} - \gamma\sqrt{1 - f_{\mathrm{T}}}$	$\kappa - \tau_{\mathrm{T}}$	$\kappa - \tau_{\mathrm{T}}$
$\kappa - \tau_{\mathrm{T}}$	$\kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	z
$\kappa - \frac{\gamma^2 f_{\mathrm{T}}}{4\tau_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}} - \gamma$	z
$\kappa + \tau_{\mathrm{T}} - \gamma$	$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	z
$\kappa + \tau_{\mathrm{T}} - \gamma\sqrt{f_{\mathrm{T}}}$	$\kappa + \tau_{\mathrm{T}}$	$\kappa + \tau_{\mathrm{T}}$
$\kappa + \tau_{\mathrm{T}}$	∞	z

(C.10)

From tables (C.5)-(C.10) we can see that some of them give the same $\hat{\lambda}(z)$ functions. In fact there are only three unique regimes of parameter space we need to be concerned with, and these are the ones given in expression (2.28).

Appendix D

Stability Analysis for the Replica Symmetric Ansatz

This appendix derives the expressions with which we judge whether the replica-symmetric ansätze employed in the preceding chapters are valid. In particular we are interested in checking against the effects of small, local replica-symmetry breaking fluctuations. These fluctuations are categorised by the permutation properties of the replica indices into three types: symmetric, weakly-asymmetric, and asymmetric fluctuations.

There follows two main sections, one concerned with the external field model of chapter 2, and the other a more involved treatment of the dual sign-constrained network of chapter 3. The starting point in both cases is a Taylor expansion of the saddle-point equation around the stationary point up to the second order. The resulting Hessian matrix of the second order derivatives tells us whether the stationary point indeed maximises the saddle-point equation. The daunting task of evaluating the Hessian is made tractable by examining the effects of the three types of fluctuations in turn, each of which gives a vastly simplified matrix. This appendix will finish by showing how in the zero replica limit the weakly-asymmetric and symmetric fluctuations produce the same Hessian, and we need only concern ourselves with stability to symmetric and asymmetric fluctuations.

The relevant chapters will ascertain replica-symmetry stability by examining the signs of the determinants of these Hessian matrices. Namely, whether the signs change as a parameter is varied, indicating a transition to or from instability depending on the model's initial stability. Hence any unimportant constant factors can be divided out from the determinants of the Hessians.

D.1 Stability of Replica Symmetry for the External Field Model

We are interested in ascertaining the stability of the stationary point of equations (2.20), which is supposed to maximise the equation

$$G(\{E_a, F_{ab}, q_{ab}\}) \equiv G_J(\{E_a, F_{ab}\}) + \alpha G_\Lambda(\{q_{ab}\}) + G_0(\{E_a, F_{ab}, q_{ab}\}) \quad \text{(D.1)}$$

where the functions $G_J(\cdots)$, $G_\Lambda(\cdots)$ and $G_0(\cdots)$ are given in equations (2.15).

We shall probe the stationary point with replica-symmetry breaking fluctuations of the form

$$\begin{aligned} E_a &= E + \delta E_a, \quad \forall a, \\ F_{ab} &= F + \delta F_{ab}, \quad \forall a, b\,(a < b), \\ q_{ab} &= q + \delta q_{ab}, \quad \forall a, b\,(a < b), \end{aligned}$$

which give the second order term of the Taylor expansion of the function (D.1) as

$$\delta^2 G = \frac{1}{2!}\left[\sum_{a,b} \frac{\partial^2 G}{\partial E_a \partial E_b}\delta E_a \delta E_b + \sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial F_{ab}\partial F_{cd}}\delta F_{ab}\delta F_{cd}\right.$$

$$+\sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial q_{ab}\partial q_{cd}} \delta q_{ab}\delta q_{cd} + 2\sum_{a(b<c)} \frac{\partial^2 G}{\partial E_a \partial F_{bc}} \delta E_a \delta F_{bc}$$
$$+2\sum_{a(b<c)} \frac{\partial^2 G}{\partial E_a \partial q_{bc}} \delta E_a \delta q_{bc} + 2\sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial F_{ab}\partial q_{cd}} \delta F_{ab}\delta q_{cd}\Bigg] \qquad \text{(D.2)}$$

where the restricted summations enforce enumerations over only the defined replicated quantities. A superficial glance at equations (D.1) and (2.15) allows us to immediately simplify the above equation (D.2), by observing

$$\begin{aligned} \frac{\partial^2 G}{\partial E_a \partial q_{bc}} &= 0, \\ \frac{\partial^2 G}{\partial F_{ab}\partial q_{cd}} &= -\mathrm{i}\delta_{a,c}\delta_{b,d}, \qquad \text{for all } (a<b) \text{ and } (c<d). \end{aligned}$$

For the other terms we shall make use of any symmetry properties in the replica indices. This is done by calculating the double derivatives of equation (D.1) and identifying all the unique terms. The following four subsections will examine the permutation symmetries of these derivatives.

Fluctuations in $\delta E_a \delta E_b$

The permutation symmetries of the double derivatives mean not all of them are unique. We shall make use of this by first dealing with the term containing fluctuations in $\delta E_a \delta E_b$. We can re-write the summations in equation (D.2) to reflect the two possibilities given by the replica indices by defining

$$\begin{aligned} Q^A_{EE} &\equiv \frac{\partial^2 G}{\partial E_a^2}, \quad \forall a, \\ Q^B_{EE} &\equiv \frac{\partial^2 G}{\partial E_a E_b}, \quad \forall a,b\,(a \neq b), \end{aligned} \qquad \text{(D.3)}$$

and hence the term in question can be written as

$$\begin{aligned}\sum_{a,b}\frac{\partial^2 G}{\partial E_a \partial E_b}\delta E_a \delta E_b &= Q^A_{EE}\sum_a \delta E_a^2 + Q^B_{EE}\sum_{a\neq b}\delta E_a \delta E_b \\ &= (Q^A_{EE}-Q^B_{EE})\sum_a \delta E_a^2 + Q^B_{EE}\sum_{a,b}\delta E_a \delta E_b. \qquad \text{(D.4)}\end{aligned}$$

Fluctuations in $\delta E_a \delta F_{bc}$

We can do a similar analysis for the term with the fluctuations $\delta E_a F_{bc}$, and partition the summation into the three groups

$$\begin{aligned}&\sum_{a(b<c)}\frac{\partial^2 G}{\partial E_a \partial F_{bc}}\delta E_a \delta F_{bc} \\ &= Q^A_{EF}\sum_{a(a<b)}\delta E_a \delta F_{ab} + Q^A_{EF}\sum_{a(b<a)}\delta E_a \delta F_{ba} + Q^B_{EF}\sum_{\substack{a(b<c)\\ a\neq b,c}}\delta E_a \delta F_{bc} \\ &= Q^A_{EF}\left[\sum_{a(a<b)}\delta E_a \delta F_{ab} + \sum_{a(b<a)}\delta E_a \delta F_{ba}\right] \\ &\quad + Q^B_{EF}\left[\sum_{a(b<c)}\delta E_a \delta F_{bc} - \left(\sum_{a(a<b)}\delta E_a \delta F_{ab} + \sum_{a(b<a)}\delta E_a \delta F_{ba}\right)\right]\end{aligned}$$

which because the restricted sums over the indices $a(a<b)$ and $a(b<a)$ are the same as that over $(a<b)$, simplifies into

$$\sum_{a(b<c)}\frac{\partial^2 G}{\partial E_a \partial F_{bc}}\delta E_a \delta F_{bc} = 2(Q^A_{EF}-Q^B_{EF})\sum_{(a<b)}\delta E_a \delta F_{ab} + Q^B_{EF}\sum_{a(b<c)}\delta E_a \delta F_{ab}. \qquad \text{(D.5)}$$

Fluctuations in $\delta F_{ab}\delta F_{cd}$

The term with the fluctuations in $\delta F_{ab}\delta F_{cd}$ is a little bit more complicated. We start by writing

$$\begin{aligned}
&\sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial F_{ab}\partial F_{cd}}\delta F_{ab}\delta F_{cd} \\
&= Q^A_{FF}\sum_{(a<b)}\delta F^2_{ab} + Q^B_{FF}\left[\sum_{\substack{(a<b)(a<c)\\ b\neq c}}\delta F_{ab}\delta F_{ac} + \sum_{(b<a)(a<c)}\delta F_{ba}\delta F_{ac}\right. \\
&\qquad \left. + \sum_{\substack{(b<a)(c<a)\\ b\neq c}}\delta F_{ba}\delta F_{ca} + \sum_{(a<b)(c<a)}\delta F_{ab}\delta F_{ca}\right] + Q^C_{FF}\sum_{\substack{(a<b)(c<d)\\ a\neq c,b\neq d}}\delta F_{ab}\delta F_{cd}.
\end{aligned}$$

The summations with the restrictions $\left\{{}^{(a<b)(a<c)}_{\;\;\;b\neq c}\right\}$ and $\{(b<a)(a<c)\}$ give the same indices as $\left\{{}^{(b<a)(c<a)}_{\;\;\;b\neq c}\right\}$ and $\{(a<b)(c<a)\}$ respectively, so we can write

$$\begin{aligned}
&\sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial F_{ab}\partial F_{cd}}\delta F_{ab}\delta F_{cd} \\
&= Q^A_{FF}\sum_{(a<b)}\delta F^2_{ab} + 2Q^B_{FF}\left[\sum_{\substack{(a<b)(a<c)\\ b\neq c}}\delta F_{ab}\delta F_{ac} + \sum_{(b<a)(a<c)}\delta F_{ba}\delta F_{ac}\right] \\
&\qquad +Q^C_{FF}\left[\sum_{(a<b)(c<d)}\delta F_{ab}\delta F_{cd} - 2\left(\sum_{\substack{(a<b)(a<c)\\ b\neq c}}\delta F_{ab}\delta F_{ac} + \sum_{(b<a)(a<c)}\delta F_{ba}\delta F_{ac}\right)\right. \\
&\qquad\qquad \left. - \sum_{(a<b)}\delta F^2_{ab}\right] \\
&= (Q^A_{FF} - Q^B_{FF})\sum_{(a<b)}\delta F^2_{ab} + Q^C_{FF}\sum_{(a<b)(c<d)}\delta F_{ab}\delta F_{cd} \\
&\qquad +2(Q^B_{FF} - Q^C_{FF})\left[\sum_{\substack{(a<b)(a<c)\\ b\neq c}}\delta F_{ab}\delta F_{ac} + \sum_{(b<a)(a<c)}\delta F_{ba}\delta F_{ac}\right].
\end{aligned}$$

This can be further simplified by realising the summations

$$\begin{aligned}\sum_{\substack{(a<b)(a<c)\\ b\neq c}} \delta F_{ab}\delta F_{ac} + \sum_{(b<a)(a<c)} \delta F_{ba}\delta F_{ac} &= \sum_{(a<b)(a<c)} \delta F_{ab}\delta F_{ac} + \sum_{(b<a)(a<c)} \delta F_{ba}\delta F_{ac} - \sum_{(a<b)} \delta F_{ab}^2 \\ &= 2\sum_{(a<b)(a<c)} \delta F_{ab}\delta F_{ac} - \sum_{(a<b)} \delta F_{ab}^2\end{aligned}$$

hence allowing us to write

$$\begin{aligned}\sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial F_{ab}\partial F_{cd}}\delta F_{ab}\delta F_{cd} &= (Q_{FF}^A - 2Q_{FF}^B + Q_{FF}^C)\sum_{(a<b)} \delta F_{ab}^2 \\ &\quad +4(Q_{FF}^B - Q_{FF}^C)\sum_{(a<b)(a<c)} \delta F_{ab}\delta F_{ac} \\ &\quad +Q_{FF}^C \sum_{(a<b)(c<d)} \delta F_{ab}\delta F_{cd}. \qquad \text{(D.6)}\end{aligned}$$

FLUCTUATIONS IN $\delta q_{ab}\delta q_{cd}$

Finally, summations over fluctuations in $\delta q_{ab}\delta q_{cd}$ follow an identical treatment to give

$$\begin{aligned}\sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial q_{ab}\partial q_{cd}}\delta q_{ab}\delta q_{cd} &= (Q_{qq}^A - 2Q_{qq}^B + Q_{qq}^C)\sum_{(a<b)} \delta q_{ab}^2 \\ &\quad +4(Q_{qq}^B - Q_{qq}^C)\sum_{(a<b)(a<c)} \delta q_{ab}\delta q_{ac} \\ &\quad +Q_{qq}^C \sum_{(a<b)(c<d)} \delta q_{ab}\delta q_{cd}. \qquad \text{(D.7)}\end{aligned}$$

Collecting equations (D.4)–(D.7) into equation (D.2), we can remove all the summation restrictions by setting the diagonal terms $\delta F_{aa} = \delta q_{aa} = 0$ to give

$$\begin{aligned}
2\delta^2 G = {} & (Q^A_{EE} - Q^B_{EE})\sum_a \delta E_a^2 + Q^B_{EE}\left[\sum_a \delta E_a\right]^2 \\
& +2(Q^A_{EF} - Q^B_{EF})\sum_{a,b} \delta E_a \delta F_{ab} + Q^B_{EF}\sum_a \delta E_a \sum_{b,c} \delta F_{bc} \\
& +\frac{1}{2}(Q^A_{FF} - 2Q^B_{FF} + Q^C_{FF})\sum_{a,b} \delta F_{ab}^2 + (Q^B_{FF} - Q^C_{FF})\sum_a \left[\sum_b \delta F_{ab}\right]^2 \\
& +\frac{Q^C_{FF}}{4}\left[\sum_{a,b} \delta F_{ab}\right]^2 \\
& +\frac{1}{2}(Q^A_{qq} - 2Q^B_{qq} + Q^C_{qq})\sum_{a,b} \delta q_{ab}^2 + (Q^B_{qq} - Q^C_{qq})\sum_a \left[\sum_b \delta q_{ab}\right]^2 \\
& +\frac{Q^C_{qq}}{4}\left[\sum_{a,b} \delta q_{ab}\right]^2 \\
& -\mathrm{i}\sum_{a,b} \delta F_{ab} \delta q_{ab}. \qquad \text{(D.8)}
\end{aligned}$$

Equation (D.8) allows us to explicitly see which fluctuations mode to choose, avoiding the problem of directly diagonalising the Hessian. The following subsections examine how this is done by choosing three types of fluctuations, whose mutual orthogonality and spanning of the $[n + 2n(n-1)/2] = n^2$ space must be ensured if they are to be a complete description.

D.1.1 Symmetric Fluctuations

We shall begin by inserting fluctuations which are symmetric in the replica indices, that is

$$\begin{aligned}
E_a &= E + \delta E, \quad \forall a, \\
F_{ab} &= F + \delta F, \quad \forall a, b\, (a \neq b), \qquad \text{(D.9)} \\
q_{ab} &= q + \delta q, \quad \forall a, b\, (a \neq b),
\end{aligned}$$

with the diagonal fluctuations F_{aa} and q_{aa} set to zero for all a. This vector spans a subspace of 3-dimensions and reduces the Hessian matrix to

$$\begin{pmatrix} \partial^2 E & \partial E \partial F & 0 \\ \partial E \partial F & \partial^2 F & -\frac{in}{2}(n-1) \\ 0 & -\frac{in}{2}(n-1) & \partial^2 q \end{pmatrix}$$

where the elements are given by

$$\begin{aligned} \partial^2 E &= n(Q^A_{EE} + (n-1)Q^B_{EE}), \\ \partial E \partial F &= \frac{n}{2}(n-1)(2Q^A_{EF} + (n-2)Q^B_{EF}), \\ \partial^2 F &= \frac{n}{2}(n-1)\left[Q^A_{FF} + 2(n-2)Q^B_{FF} + \frac{1}{2}(n-2)(n-3)Q^C_{FF}\right], \\ \partial^2 q &= \frac{n}{2}(n-1)\left[Q^A_{qq} + 2(n-2)Q^B_{qq} + \frac{1}{2}(n-2)(n-3)Q^C_{qq}\right]. \end{aligned}$$

Since we shall be primarily interested in sign changes in the eigenvalues of this matrix, we can simplify the above by considering their product. Upon dividing out common factors and taking the $(n \to 0)$ zero replica limit the determinant is:

$$\begin{vmatrix} (Q^A_{EE} - Q^B_{EE}) & -(Q^A_{EF} - Q^B_{EF}) & 0 \\ 2(Q^A_{EF} - Q^B_{EF}) & (Q^A_{FF} - 4Q^B_{FF} + 3Q^C_{FF}) & -\mathrm{i} \\ 0 & -\mathrm{i} & (Q^A_{qq} - 4Q^B_{qq} + 3Q^C_{qq}) \end{vmatrix}. \qquad \text{(D.10)}$$

D.1.2 Weakly-Asymmetric Fluctuations

We next consider weakly-asymmetric fluctuations as defined by

$$\begin{aligned} E_a &= E + \delta E_a, & &\forall a, \\ F_{ab} &= F + \delta F_a + \delta F_b, & &\forall a, b\ (a \neq b), \\ q_{ab} &= q + \delta q_a + \delta F_b, & &\forall a, b\ (a \neq b), \end{aligned} \qquad \text{(D.11)}$$

which when required to be orthogonal to the symmetric fluctuations (D.9) results in the constraints

$$\sum_a \delta E_a = \sum_a \delta F_a = \sum_a q_a = 0. \tag{D.12}$$

Hence weakly-symmetric fluctuations span $3 \times (n-1)$ dimensions and reduces the Hessian to

$$\begin{pmatrix} \partial^2 E & \partial E \partial F & 0 \\ \partial E \partial F & \partial^2 F & -\mathrm{i}(n-2) \\ 0 & -\mathrm{i}(n-2) & \partial^2 q \end{pmatrix}$$

where the matrix elements are now given by

$$\begin{aligned} \partial^2 E &= (Q^A_{EE} - Q^B_{EE}), \\ \partial E \partial F &= (n-2)(Q^A_{EF} - Q^B_{EF}), \\ \partial^2 F &= (n-2)\left[Q^A_{FF} - 2Q^B_{FF} + Q^C_{FF} + (n-2)(Q^B_{FF} - Q^C_{FF})\right], \\ \partial^2 q &= (n-2)\left[Q^A_{qq} - 2Q^B_{qq} + Q^C_{qq} + (n-2)(Q^B_{qq} - Q^C_{qq})\right]. \end{aligned}$$

In the zero replica limit the determinant for this matrix is the same as that for the symmetrical fluctuations matrix (D.10).

D.1.3 Asymmetric Fluctuations

Finally, we shall look at asymmetric fluctuations as defined by

$$\begin{aligned} E_a &= E + \delta E_a, \quad \forall a, \\ F_{ab} &= F + \delta F_{ab}, \quad \forall a, b \, (a \neq b), \\ q_{ab} &= q + \delta q_{ab}, \quad \forall a, b \, (a \neq b), \end{aligned} \tag{D.13}$$

which has the constraints

$$\delta E_a = \sum_b \delta F_{ab} = \sum_b \delta q_{ab} = 0 \tag{D.14}$$

in order to be orthogonal to the weakly-asymmetric fluctuations (D.11). These constraints reduce the dimensions spanned by δF_{ab} and δq_{ab} from $[2 \times n(n-1)/2]$ to $[2 \times n(n-3)/2]$. This give the total dimensions spanned by all three types of fluctuations as $[3 + 3(n-1) + n(n-3)] = n^2$, which is as stipulated at the end of §D.1.

The Hessian for these fluctuations is just

$$\begin{pmatrix} \frac{1}{2}(Q^A_{FF} - 2Q^B_{FF} + Q^C_{FF}) & -\frac{\mathrm{i}}{2} \\ -\frac{\mathrm{i}}{2} & \frac{1}{2}(Q^A_{qq} - 2Q^B_{qq} + Q^C_{qq}) \end{pmatrix} \tag{D.15}$$

and since we shall be interested in the sign of the product of eigenvalues, we can again divide the determinant of matrix (D.15) by any constant factors.

D.2 Stability of Replica Symmetry for the Dual Network Distributions

We now turn our attention to the dual distribution calculations of chapter 3. We shall attempt to follow the same procedure as for the external field model in §D.1, and produce a set of determinants whose signs will determine the stability of the analytical results to small replica-symmetry breaking fluctuations. By having to consider two networks each with their own sets of fluctuations, the dimensionality of the stability matrix is drastically increased and with it an explosion in the number of terms to the second order Taylor expansion around the stationary point.

Fortunately this turns out to be but an irritating inconvenience, and the real source of novel behaviour actually arises from the inter-network order parameter r and its conjugate K.

The equation whose saddle-point we wish to examine is

$$\begin{aligned} G(\{E_a, \tilde{E}_a, F_{ab}, \tilde{F}_{ab}, q_{ab}, \tilde{q}_{ab}, K_{ab}, r_{ab}\}) &\equiv G_J(\{E_a, \tilde{E}_a, F_{ab}, \tilde{F}_{ab}, K_{ab}\}) \\ &+\alpha G_\Lambda(\{q_{ab}, \tilde{q}_{ab}, r_{ab}\}) + G_0(\{E_a, \tilde{E}_a, F_{ab}, \tilde{F}_{ab}, q_{ab}, \tilde{q}_{ab}, K_{ab}, r_{ab}\}) \end{aligned} \tag{D.16}$$

with the functions on the right hand side defined in equations (3.8). The fluctuations we wish to probe are of the form

$$\begin{array}{rclrcll} E_a &=& E + \delta E_a, & \tilde{E}_a &=& \tilde{E} + \delta \tilde{E}_a, & \forall a, \\ F_{ab} &=& F + \delta F_{ab}, & \tilde{F}_{ab} &=& \tilde{F} + \delta \tilde{F}_{ab}, & \forall a, b \, (a < b), \\ q_{ab} &=& q + \delta q_{ab}, & \tilde{q}_{ab} &=& \tilde{q} + \delta \tilde{q}_{ab}, & \forall a, b \, (a < b), \\ K_{ab} &=& K + \delta K_{ab}, & r_{ab} &=& r + \delta r_{ab}, & \forall a, b \end{array} \tag{D.17}$$

which presents a total space of $(2[n + 2 \times \frac{1}{2}n(n-1)] + 2n^2) = 4n^2$ dimensions for the fluctuations to cover. Rather than write the analogous second order Taylor expansion term of equation (D.2) we can better show the mixing between the eight sets of quantities with a representation of the second order derivatives of equation (D.16) in matrix form

$$\left(\begin{array}{ccc|cc|ccc}
\partial^2 E & \partial E \partial F & 0 & \partial K \partial E & 0 & \partial E \partial \tilde{E} & \partial E \partial \tilde{F} & 0 \\
\partial E \partial F & \partial^2 F & \partial F \partial q & \partial F \partial K & 0 & \partial \tilde{E} \partial F & \partial F \partial \tilde{F} & 0 \\
0 & \partial F \partial q & \partial^2 q & 0 & \partial q \partial r & 0 & 0 & \partial q \partial \tilde{q} \\
\hline
\partial K \partial E & \partial F \partial K & 0 & \partial^2 K & \partial K \partial r & \partial K \partial \tilde{E} & \partial \tilde{F} \partial K & 0 \\
0 & 0 & \partial q \partial r & \partial K \partial r & \partial^2 r & 0 & 0 & \partial \tilde{q} \partial r \\
\hline
\partial E \partial \tilde{E} & \partial \tilde{E} \partial F & 0 & \partial K \partial \tilde{E} & 0 & \partial^2 \tilde{E} & \partial \tilde{E} \partial \tilde{F} & 0 \\
\partial E \partial \tilde{F} & \partial F \partial \tilde{F} & 0 & \partial \tilde{F} \partial K & 0 & \partial \tilde{E} \partial \tilde{F} & \partial^2 \tilde{F} & \partial \tilde{F} \partial \tilde{q} \\
0 & 0 & \partial q \partial \tilde{q} & 0 & \partial \tilde{q} \partial r & 0 & \partial \tilde{F} \partial \tilde{q} & \partial^2 \tilde{q}
\end{array}\right) \tag{D.18}$$

where we have immediately written down any zero elements. Before we can substitute in the three modes of replica-symmetry breaking fluctuations we must elucidate the permutation symmetries in the replica indices of the elements in matrix (D.18). This is, of course, done by studying the double derivatives to equation (D.16) but as we shall see, commonality in symmetries mean we need not treat each of the 24 elements on an individual basis.

The Intra-network Matrix Elements

We can see the two 3 × 3 diagonal blocks of matrix (D.18) are just realisations of networks which have already been dealt with previously in §D.1. We can hence borrow equation (D.8) for the symmetry properties of these two blocks; bar a simple re-labelling in the tilde notation for the second network.

The Elements with Simple Symmetries

The other unique 3 × 3 block in the top-righthand corner has elements which are essentially independent of their replica indices. This gives the six elements

$$\begin{aligned}
\sum_{a,b} \frac{\partial^2 G}{\partial E_a \partial \tilde{E}_b} \delta E_a \delta \tilde{E}_b &= Q^B_{E\tilde{E}} \sum_{a,b} \delta E_a \delta \tilde{E}_b, \\
\sum_{a(b<c)} \frac{\partial^2 G}{\partial E_a \partial \tilde{F}_{bc}} \delta E_a \delta \tilde{F}_{bc} &= \frac{Q_{E\tilde{F}}}{2} \sum_{a,b,c} \delta E_a \delta \tilde{F}_{bc}, \\
\sum_{a(b<c)} \frac{\partial^2 G}{\partial \tilde{E}_a \partial F_{bc}} \delta \tilde{E}_a \delta F_{bc} &= \frac{Q_{\tilde{E}F}}{2} \sum_{a,b,c} \delta \tilde{E}_a \delta F_{bc}, \\
\sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial F_{ab} \partial \tilde{F}_{cd}} \delta F_{ab} \delta \tilde{F}_{cd} &= \frac{Q_{F\tilde{F}}}{4} \sum_{a,b,c,d} \delta F_{ab} \delta \tilde{F}_{cd}, \\
\sum_{(a<b)(c<d)} \frac{\partial^2 G}{\partial q_{ab} \partial \tilde{q}_{cd}} \delta q_{ab} \delta \tilde{q}_{cd} &= \frac{Q_{q\tilde{q}}}{4} \sum_{a,b,c,d} \delta q_{ab} \delta \tilde{q}_{cd},
\end{aligned} \tag{D.19}$$

where we have set the diagonal fluctuations $\delta F_{aa} = \delta \tilde{F}_{aa} = \delta q_{aa} = \delta \tilde{q}_{aa}, \forall a$ to zero in order to write the summations in an unrestricted form.

Another element with a simple symmetry is the off-diagonal elements in the central block. From equation (3.8) we can see straightaway that

$$\sum_{a,b,c,d} \frac{\partial^2 G}{\partial K_{ab} \partial r_{cd}} \delta K_{ab} \delta r_{cd} = -\mathrm{i} \sum_{a,b} \delta K_{ab} \delta r_{ab}. \tag{D.20}$$

The preceding two subsections have reduced our task to studying the symmetries of just eight elements. These will be appropriately grouped in the next four subsections.

The Elements $\partial K_{ab} \partial E_c$ and $\partial K_{ab} \partial \tilde{E}_c$

The symmetries for these terms are straightforward with the indices allowing only two possibilities. We can express this for the $\partial K_{ab} \partial E_c$ term as

$$\sum_{a,b,c} \frac{\partial^2 G}{\partial K_{ab} \partial E_c} \delta K_{ab} \delta E_c = Q^A_{KE} \sum_{a,b} \delta K_{ab} \delta E_a + Q^B_{KE} \sum_{\substack{a,b,c \\ a \neq c}} \delta K_{ab} \delta E_c$$

and similarly for the $\partial K_{ab}\partial \tilde{E}_c$ term. In unrestricted forms these terms are

$$\begin{aligned}
\sum_{a,b,c} \frac{\partial^2 G}{\partial K_{ab}\partial E_c}\delta K_{ab}\delta E_c &= (Q^A_{KE} - Q^B_{KE})\sum_{a,b}\delta K_{ab}\delta E_a + Q^B_{KE}\sum_{a,b,c}\delta K_{ab}\delta E_c, \\
\sum_{a,b,c} \frac{\partial^2 G}{\partial K_{ab}\partial \tilde{E}_c}\delta K_{ab}\delta \tilde{E}_c &= (Q^A_{K\tilde{E}} - Q^B_{K\tilde{E}})\sum_{a,b}\delta K_{ab}\delta \tilde{E}_b + Q^B_{K\tilde{E}}\sum_{a,b,c}\delta K_{ab}\delta \tilde{E}_c.
\end{aligned} \tag{D.21}$$

THE ELEMENTS $\partial F_{ab}\partial K_{cd}$ AND $\partial \tilde{F}_{ab}\partial K_{cd}$

The next term we shall consider arises from taking the derivatives with respect to $\partial F_{ab}\partial K_{cd}$, whose permutation symmetry can be grouped into two parts

$$\begin{aligned}
&\sum_{(a<b)c,d} \frac{\partial^2 G}{\partial F_{ab}\partial K_{cd}}\delta F_{ab}\delta K_{cd} \\
&= Q^A_{FK}\sum_{(a<b)c}\delta F_{ab}\delta K_{ac} + Q^A_{FK}\sum_{(a<b)c}\delta F_{ab}\delta K_{bc} + Q^B_{FK}\sum_{\substack{(a<b)c,d \\ a,b\neq c}}\delta F_{ab}\delta K_{cd} \\
&= \left(Q^A_{FK} - Q^B_{FK}\right)\left[\sum_{(a<b)c}\delta F_{ab}\delta K_{ac} + \sum_{(a<b)c}\delta F_{ab}\delta K_{bc}\right] + Q^B_{FK}\sum_{(a<b)c,d}\delta F_{ab}\delta K_{cd},
\end{aligned}$$

and for the $\partial \tilde{F}_{ab}\partial K_{cd}$ term

$$\begin{aligned}
\sum_{(a<b)c,d} \frac{\partial^2 G}{\partial \tilde{F}_{ab}\partial K_{cd}}\delta \tilde{F}_{ab}\delta K_{cd} &= \left(Q^A_{FK} - Q^B_{FK}\right)\left[\sum_{(a<b)c}\delta \tilde{F}_{ab}\delta K_{ca} + \sum_{(a<b)c}\delta \tilde{F}_{ab}\delta K_{cb}\right] \\
&\quad + Q^B_{FK}\sum_{(a<b)c,d}\delta \tilde{F}_{ab}\delta K_{cd}.
\end{aligned}$$

These can be written in a simplified form as

$$
\begin{aligned}
\sum_{(a<b)c,d} \frac{\partial^2 G}{\partial F_{ab} \partial K_{cd}} \delta F_{ab} \delta K_{cd} &= \left(Q^A_{FK} - Q^B_{FK}\right) \sum_a \left[\sum_b \delta F_{ab}\right] \left[\sum_c \delta K_{ac}\right] \\
&\quad + \frac{Q^B_{FK}}{2} \sum_{a,b,c,d} \delta F_{ab} \delta K_{cd}, \\
\sum_{(a<b)c,d} \frac{\partial^2 G}{\partial \tilde{F}_{ab} \partial K_{cd}} \delta \tilde{F}_{ab} \delta K_{cd} &= \left(Q^A_{FK} - Q^B_{FK}\right) \sum_a \left[\sum_b \delta \tilde{F}_{ab}\right] \left[\sum_c \delta K_{ca}\right] \\
&\quad + \frac{Q^B_{\tilde{F}K}}{2} \sum_{a,b,c,d} \delta \tilde{F}_{ab} \delta K_{cd}.
\end{aligned} \tag{D.22}
$$

THE ELEMENTS $\partial q_{ab} \partial r_{cd}$ AND $\partial \tilde{q}_{ab} \partial r_{cd}$

These two terms possess exactly the symmetries as for $\partial F_{ab} \partial K_{cd}$ and $\partial \tilde{F}_{ab} \partial K_{cd}$ in the previous subsection. Hence we can simply write

$$
\begin{aligned}
\sum_{(a<b)c,d} \frac{\partial^2 G}{\partial q_{ab} \partial r_{cd}} \delta q_{ab} \delta r_{cd} &= \left(Q^A_{qr} - Q^B_{qr}\right) \sum_a \left[\sum_b \delta q_{ab}\right] \left[\sum_c \delta r_{ac}\right] \\
&\quad + \frac{Q^B_{qr}}{2} \sum_{a,b,c,d} \delta q_{ab} \delta r_{cd}, \\
\sum_{(a<b)c,d} \frac{\partial^2 G}{\partial \tilde{q}_{ab} \partial r_{cd}} \delta \tilde{q}_{ab} \delta r_{cd} &= \left(Q^A_{qr} - Q^B_{qr}\right) \sum_a \left[\sum_b \delta \tilde{q}_{ab}\right] \left[\sum_c \delta r_{ca}\right] \\
&\quad + \frac{Q^B_{\tilde{q}r}}{2} \sum_{a,b,c,d} \delta \tilde{q}_{ab} \delta r_{cd}.
\end{aligned} \tag{D.23}
$$

THE ELEMENTS $\partial K_{ab} \partial K_{cd}$ AND $\partial r_{ab} \partial r_{cd}$

Finally we shall consider the two diagonal terms of the central block in the stability matrix (D.18). For the derivative with respect to $\partial K_{ab} \partial K_{cd}$ we find there are four possible symmetries

$$\begin{aligned}\sum_{a,b,c,d} \frac{\partial^2 G}{\partial K_{ab} \partial K_{cd}} \delta K_{ab} \delta K_{cd} &= Q^A_{KK} \sum_{a,b} \delta K^2_{ab} + Q^B_{KK} \sum_{\substack{a,b,c \\ b \neq c}} \delta K_{ab} \delta K_{ac} \\ &+ \tilde{Q}^B_{KK} \sum_{\substack{a,b,c \\ a \neq c}} \delta K_{ab} \delta K_{cb} + Q^C_{KK} \sum_{\substack{a,b,c,d \\ a \neq c, b \neq d}} \delta K_{ab} \delta K_{cd}\end{aligned}$$

which can be rewritten in a form without restrictions on the summations. Since the derivatives with respect to $\partial r_{ab} \partial r_{cd}$ follow the same symmetries, we shall write these two terms together as

$$\begin{aligned}\sum_{a,b,c,d} \frac{\partial^2 G}{\partial K_{ab} \partial K_{cd}} \delta K_{ab} \delta K_{cd} &= \left(Q^A_{KK} - [Q^B_{KK} + \tilde{Q}^B_{KK}] + Q^C_{KK}\right) \sum_{a,b} \delta K^2_{ab} \\ &+ \left(Q^B_{KK} - Q^C_{KK}\right) \sum_a \left[\sum_b \delta K_{ab}\right] \left[\sum_c \delta K_{ac}\right] \\ &+ \left(\tilde{Q}^B_{KK} - Q^C_{KK}\right) \sum_a \left[\sum_b \delta K_{ba}\right] \left[\sum_c \delta K_{ca}\right] + Q^C_{KK} \sum_{a,b,c,d} \delta K_{ab} \delta_{cd}, \\ \sum_{a,b,c,d} \frac{\partial^2 G}{\partial r_{ab} \partial r_{cd}} \delta r_{ab} \delta r_{cd} &= \left(Q^A_{rr} - [Q^B_{rr} + \tilde{Q}^B_{rr}] + Q^C_{rr}\right) \sum_{a,b} \delta r^2_{ab} \\ &+ \left(Q^B_{rr} - Q^C_{rr}\right) \sum_a \left[\sum_b \delta r_{ab}\right] \left[\sum_c \delta r_{ac}\right] \\ &+ \left(\tilde{Q}^B_{rr} - Q^C_{rr}\right) \sum_a \left[\sum_b \delta r_{ba}\right] \left[\sum_c \delta r_{ca}\right] + Q^C_{rr} \sum_{a,b,c,d} \delta r_{ab} \delta_{cd}. \qquad \text{(D.24)}\end{aligned}$$

This completes the symmetry analysis for the terms in the stability matrix (D.18). We can now substitute in the three possible replica-symmetry breaking fluctuations and examine their effects on the matrix determinant.

D.2.1 Symmetric Fluctuations

The fluctuations symmetric in the replica indices are defined as

$$\begin{aligned}
E_a &= E+\delta E, & \tilde{E}_a &= \tilde{E}+\delta\tilde{E}, & &\forall a,\\
F_{ab} &= F+\delta F, & \tilde{F}_{ab} &= \tilde{F}+\delta\tilde{F}, & &\forall a,b\,(a\neq b),\\
q_{ab} &= q+\delta q, & \tilde{q}_{ab} &= \tilde{q}+\delta\tilde{q}, & &\forall a,b\,(a\neq b),\\
K_{ab} &= K+\delta K, & r_{ab} &= r+\delta r, & &\forall a,b,
\end{aligned} \tag{D.25}$$

with the diagonal contributions $F_{aa}, \tilde{F}_{aa}, q_{aa}$, and $\tilde{q}_{aa}$ set to zero for all $a = 1\ldots n$. Inserting these into the stability matrix (D.18) we find the two intra-network diagonal blocks are the same as in matrix (D.10), with the elements of order n and $n(n-1)$. In contrast, the elements in the rest of the stability matrix are of order $n^{\geq 2}$. Hence in the zero replica ($n \to 0$) limit only the two intra-network blocks survive and the stability matrix reduces to

$$\left(\begin{array}{ccc|ccc}
\partial^2 E & \partial E\partial F & 0 & 0 & 0 & 0\\
\partial E\partial F & \partial^2 F & -\mathrm{i} & 0 & 0 & 0\\
0 & -\mathrm{i} & \partial^2 q & 0 & 0 & 0\\
\hline
0 & 0 & 0 & \partial^2\tilde{E} & \partial\tilde{E}\partial\tilde{F} & 0\\
0 & 0 & 0 & \partial\tilde{E}\partial\tilde{F} & \partial^2\tilde{F} & -\mathrm{i}\\
0 & 0 & 0 & 0 & -\mathrm{i} & \partial^2\tilde{q}
\end{array}\right) \tag{D.26}$$

with the elements in each block of the same form as those in matrix (D.10). Hence we can say that if each network is uninfluenced by symmetric fluctuations, then the dual distribution calculation as a whole is also unaffected.

D.2.2 Weakly-Asymmetric Fluctuations

The weakly-asymmetric fluctuations are given by

$$\begin{array}{rclrcll} E_a & = & E+\delta E_a, & \tilde{E}_a & = & \tilde{E}+\delta\tilde{E}_a, & \forall a, \\ F_{ab} & = & F+\delta F_a+\delta F_b, & \tilde{F}_{ab} & = & \tilde{F}+\delta\tilde{F}+\delta\tilde{F}_b, & \forall a,b\,(a\neq b), \\ q_{ab} & = & q+\delta q_a+\delta q_b, & \tilde{q}_{ab} & = & \tilde{q}+\delta\tilde{q}_a+\delta\tilde{q}_b, & \forall a,b\,(a\neq b), \\ K_{ab} & = & K+\delta K_a+\delta\tilde{K}_b, & r_{ab} & = & r+\delta r_a+\delta\tilde{r}_b, & \forall a,b, \end{array} \qquad \text{(D.27)}$$

with the orthogonality constraints

$$\begin{array}{rcl} \sum_a \delta E_a = \sum_a \delta F_a = \sum_a \delta q_a & = & 0, \\ \sum_a \delta\tilde{E}_a = \sum_a \delta\tilde{F}_a = \sum_a \delta\tilde{q}_a & = & 0, \\ \sum_a \delta r_a = \sum_a \delta\tilde{r}_a = \sum_a \delta K_a = \sum_a \delta\tilde{K}_a & = & 0, \end{array} \qquad \text{(D.28)}$$

which results in the space spanned by the fluctuations (D.27) to $10(n-1)$ dimensions . Substituting in these fluctuations, we find the two intra-network diagonal blocks are of order $(n-2)$, whilst the other terms are of order $n^{\geq 1}$. So as with symmetric fluctuations, in the zero replica limit the stability matrix is reduced to two diagonal blocks with zero elements elsewhere. Furthermore, this matrix will have the same determinant as in the symmetric case, as in the case for the single network (external field) model of §D.1.

D.2.3 Asymmetric Fluctuations

Finally we shall now examine fluctuations of the type

$$\begin{aligned}
E_a &= E + \delta E_a, & \tilde{E}_a &= \tilde{E} + \delta \tilde{E}_a, & \forall a, \\
F_{ab} &= F + \delta F_{ab}, & \tilde{F}_{ab} &= \tilde{F} + \delta \tilde{F}_{ab}, & \forall a, b\,(a \neq b), \\
q_{ab} &= q + \delta q_{ab}, & \tilde{q}_{ab} &= \tilde{q} + \delta \tilde{q}_{ab}, & \forall a, b\,(a \neq b), \\
K_{ab} &= K + \delta K_{ab}, & r_{ab} &= r + \delta r_{ab}, & \forall a, b.
\end{aligned} \qquad \text{(D.29)}$$

Demanding orthogonality with the weakly-asymmetric fluctuations leads to the constraints

$$\begin{aligned}
\delta E_a = \sum_b \delta F_{ab} = \sum_b \delta q_{ab} &= 0, \\
\delta \tilde{E}_a = \sum_b \delta \tilde{F}_{ab} = \sum_b \delta \tilde{q}_{ab} &= 0, \\
\sum_a \delta r_{ab} = \sum_b \delta r_{ab} = \sum_a \delta K_{ab} = \sum_b \delta K_{ab} &= 0,
\end{aligned} \qquad \text{(D.30)}$$

which restricts the space spanned by these fluctuations to $2[n(n-3)+(n-1)^2]$ dimensions. Hence the total number of dimensions covered by the three fluctuation modes is $[8+10(n-1)+2n(n-3)+2(n-1)^2] = 4n^2$, as demanded in §D.2.

Substituting in the asymmetric fluctuations, we find the stability matrix is also greatly simplified in the zero replica limit. Unlike the previous two types of fluctuations however, we do obtain a non-trivial result in that we now have three decoupled blocks making up the matrix, that is

$$\left(\begin{array}{cc|cc|cc} \partial^2 F & -\frac{\mathrm{i}}{2} & 0 & 0 & 0 & 0 \\ -\frac{\mathrm{i}}{2} & \partial^2 q & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \partial^2 K & -\mathrm{i} & 0 & 0 \\ 0 & 0 & -\mathrm{i} & \partial^2 r & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \partial^2 \tilde{F} & -\frac{\mathrm{i}}{2} \\ 0 & 0 & 0 & 0 & -\frac{\mathrm{i}}{2} & \partial^2 \tilde{q} \end{array}\right). \tag{D.31}$$

The elements for matrix (D.31) are given by

$$\begin{aligned} \partial^2 F &= \frac{1}{2}\left(Q^A_{FF} - 2Q^B_{FF} + Q^C_{FF}\right), \\ \partial^2 \tilde{F} &= \frac{1}{2}\left(Q^A_{\tilde{F}\tilde{F}} - 2Q^B_{\tilde{F}\tilde{F}} + Q^C_{\tilde{F}\tilde{F}}\right), \\ \partial^2 q &= \frac{1}{2}\left(Q^A_{qq} - 2Q^B_{qq} + Q^C_{qq}\right), \\ \partial^2 \tilde{q} &= \frac{1}{2}\left(Q^A_{\tilde{q}\tilde{q}} - 2Q^B_{\tilde{q}\tilde{q}} + Q^C_{\tilde{q}\tilde{q}}\right), \\ \partial^2 K &= Q^A_{KK} - (Q^B_{KK} + \tilde{Q}^B_{KK}) + Q^C_{KK}, \\ \partial^2 r &= Q^A_{rr} - (Q^B_{rr} + \tilde{Q}^B_{rr}) + Q^C_{rr}. \end{aligned} \tag{D.32}$$

Appendix E

Mathematics for the Sign-Constrained Network

In this appendix we shall deal with the angled-bracket operators defined in equations (3.11), and in particular arbitrary k^{th} moments of them. We can spare ourselves the chore of looking at all three operators separately by recognising that they may be dealt with in essentially the same manner. Using the function defined in the mathematics appendix (A.4), we can write down

$$\begin{aligned}
\langle J^k\rangle_J \;&\equiv\; \int_0^\infty \mathrm{d}J(\mathrm{g}J)^k \exp\left(-\frac{1}{2}\beta^2 J^2 + \alpha \mathrm{g} x J\right) \\
&\quad \div \int_0^\infty \mathrm{d}J \exp\left(-\frac{1}{2}\beta^2 J^2 + \alpha \mathrm{g} x J\right) \\
&= \int_{-\frac{\alpha \mathrm{g} x}{\beta}}^{\infty} \mathcal{D}J \left[\frac{\mathrm{g}}{\beta}\right]^k \left[J + \frac{\alpha \mathrm{g} x}{\beta}\right]^k \div \overline{\mathrm{H}}\left[\frac{\alpha \mathrm{g} x}{\beta}\right] \qquad \text{(E.1)}
\end{aligned}$$

as representing some generalised form of equations (3.11) for either of the two networks, and for positive or negative weight constraints. Depending which of the two networks is being considered, β^2 is hence either $\mathrm{i}(E+F)$ or $\mathrm{i}(\tilde{E}+\tilde{F})$, α is either $\sqrt{\mathrm{i}F}$ or $\sqrt{\mathrm{i}\tilde{F}}$, and x is either $(\mathrm{A}u + \mathrm{B}v)$ or $(\mathrm{A}u - \mathrm{B}v)$. Similarly, the flag g is set depending whether we have positive ($\mathrm{g} = 1$) or negative ($\mathrm{g} = -1$) weights.

The key assumption in calculating the moments $\langle J^k\rangle_J$ of equations (E.1) is that the variables α and β^2 are of order $(1-q)^{-1}$ (likewise of order $(1-\tilde{q})^{-1}$ for the conjugate network), and hence in the optimal perceptron limit of (q and $\tilde{q} \to 1$), the quantity α/β becomes large. The *a-priori* justification for this comes from considering non sign-constrained networks where this is indeed the case, as can be seen by solving equations (2.25) which is a non sign-constrained model. Alternatively, and somewhat more rigorously, we can show that the assumption of α/β being large is consistent with the solutions (3.31) of equations (3.19)–(3.21). The consequence of this is that we can expand the appropriate functions $\overline{\mathrm{H}}[\cdots]$ for large arguments via equation (A.5). Hence up to the fourth moment we find

$$\begin{aligned}
\langle J\rangle_J &= \beta^{-1}\left[\frac{\mathrm{g}}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2x^2}{2\beta^2}\right)\div\overline{\mathrm{H}}\left[\frac{\alpha\mathrm{g}x}{\beta}\right]+\frac{\alpha x}{\beta}\right]\\
&= \frac{\alpha x}{\beta^2}\theta\left[\frac{\alpha\mathrm{g}x}{\beta}\right],\\
\langle J^2\rangle_J &= \beta^{-2}\left[\frac{\alpha x}{\beta}\frac{\mathrm{g}}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2x^2}{2\beta^2}\right)\div\overline{\mathrm{H}}\left[\frac{\alpha\mathrm{g}x}{\beta}\right]+1+\frac{\alpha^2x^2}{\beta^2}\right]\\
&= \beta^{-2}\left[1+\frac{\alpha^2x^2}{\beta^2}\right]\theta\left[\frac{\alpha\mathrm{g}x}{\beta}\right],\\
\langle J^3\rangle_J &= \beta^{-3}\left[\left(2+\frac{\alpha^2x^2}{\beta^2}\right)\frac{\mathrm{g}}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2x^2}{2\beta^2}\right)\div\overline{\mathrm{H}}\left[\frac{\alpha\mathrm{g}x}{\beta}\right]+\frac{3\alpha x}{\beta}+\frac{\alpha^3x^3}{\beta^3}\right]\\
&= \beta^{-3}\left[\frac{3\alpha x}{\beta}+\frac{\alpha^3x^3}{\beta^3}\right]\theta\left[\frac{\alpha\mathrm{g}x}{\beta}\right],\\
\langle J^4\rangle_J &= \beta^{-4}\left[\left(\frac{5\alpha x}{\beta}+\frac{\alpha^3x^3}{\beta^3}\right)\frac{\mathrm{g}}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2x^2}{2\beta^2}\right)\div\overline{\mathrm{H}}\left[\frac{\alpha\mathrm{g}x}{\beta}\right]\right.\\
&\qquad\left.+\frac{\alpha^4x^4}{\beta^4}+6x^2\frac{\alpha^2}{\beta^2}+3\right]\\
&= \beta^{-4}\left[x^4\frac{\alpha^4}{\beta^4}+\frac{6\alpha^2x^2}{\beta^2}+3\right]\theta\left[\frac{\alpha\mathrm{g}x}{\beta}\right]. \qquad \text{(E.2)}
\end{aligned}$$

The first two moments are needed for solving the saddle-point equations (3.19)–(3.21), and the latter two will appear when calculating the stability matrix of the replica-symmetric ansatz.

Bibliography

[AEHW90] D J Amit, M R Evans, H Horner, and K Y M Wong. Retrieval phase diagrams for attractor neural networks with optimal interactions. *Journal of Physics A: Maths and General*, 23:3361–3381, 1990.

[AGS85] D J Amit, H Gutfreund, and H Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 79:1007–1018, 1985.

[AGS86] D J Amit, H Gutfreund, and H Sompolinsky. Statistical mechanics of neural network near saturation. *Annals of Physics*, 173:30–67, 1986.

[Ami89] D J Amit. Modelling brain function: The world of attractor neural networks. *Cambridge University Press*, 1989.

[Arf85] G Arfken. Mathematical methods for physicists (3rd edition). *Academic Press Incorporated*, 1985.

[AWC89a] D J Amit, K Y M Wong, and C Campbell. The interaction space of neural networks with sign-constrained synapses. *Journal of Physics A: Maths and General*, 22:4687–4693, 1989.

[AWC89b] D J Amit, K Y M Wong, and C Campbell. Perceptron learning with sign-constrained weights. *Journal of Physics A: Maths and General*, 22(12):2039–2045, 1989.

[BDvM91] D Bollé, P Dupont, and J van Mourik. The optimal storage capacity for a neural network with multi-state neurons. *Europhysics Letters*, 15:893–898, 1991.

[Cop65] E T Copson. Asymptotic expansions. *Cambridge University Press*, 1965.

[CR91] C Campbell and A Robinson. On the storage capacity of neural network models with sign-constrained weights. *Journal of Physics A: Maths and General*, 24:L93–L95, 1991.

[CW90] C Campbell and K Y M Wong. Dynamics and storage capacity of neural networks with sign-constrained weights. *Proceedings of the* 11th *Sitges Conference on Statistical Mechanics of Neural Networks, L. Garrido (editor) (Springer-Verlag)*, 368:237–252, 1990.

[dAT78] J R L de Almeida and D J Thouless. Stability of the Sherrington-Kirkpatrick solution of a spin glass model. *Journal of Physics A: Maths and General*, 11:983–990, 1978.

[DGZ87] B Derrida, E J Gardner, and A Zippelius. An exactly solvable asymmetric neural network model. *Europhysics Letters*, 4:167, 1987.

[EA75] S F Edwards and P W Anderson. Theory of spin-glasses. *Journal of Physics F: Metal Physics*, 5:965–974, 1975.

[EBKS90] A Engel, M Bouten, A Komoda, and R Serneels. Enlarged basin of attraction in neural networks with persistent stimuli. *Physical Review A*, 42:4998–5005, 1990.

[Ecc64] J C Eccles. The physiology of synapses. *Berlin: Springer-Verlag*, 1964.

[For88] B M Forrest. Content-addressability in neural networks. *Journal of Physics A: Maths and General*, 21:245–256, 1988.

[Fra91] Silvia Franz. Private communications. 1991.

[Gar88] E J Gardner. The space of interactions in neural network models. *Journal of Physics A: Maths and General*, 21:257–270, 1988.

[Gar89] E J Gardner. Optimal basins of attraction in randomly-sparse neural network models. *Journal of Physics A: Maths and General*, 22(12):1969–1974, 1989.

[GD88] E J Gardner and B Derrida. Optimal storage of neural network models. *Journal of Physics A: Maths and General*, 21:270–284, 1988.

[GSW89] E J Gardner, N Stroud, and D J Wallace. Training with noise and the storage of correlated patterns in a neural network model. *Journal of Physics A: Maths and General*, 22(12):2019–2030, 1989.

[Heb49] D O Hebb. The organisation of behaviour: A neuropsychological theory. *New York: Wiley*, 1949.

[Hen91] N Hendrich. Associative memory in damaged neural networks. *Journal of Physics A: Maths and General*, 24:2877–2887, 1991.

[HKP91] J Hertz, A Krogh, and R G Palmer. Introduction to the theory of neural computing. *Lectures in the Science of Complexity, Santa Fe Institute Studies in the Sciences of Complexity, lectures volume 1 (Addison-Wesley Longman)*, 1991.

[Hop82] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science USA*, 79:2554–2558, 1982.

[KA88] T B Kepler and L F Abbott. Domains of attraction in neural networks. *Journal de Physique France*, 49:1657–1662, 1988.

[KS75] S Kirkpatrick and D Sherrington. Solvable models of a spin-glass. *Physical Review Letters*, 17:1792–1796, 1975.

[KS78] S Kirkpatrick and D Sherrington. Infinite-ranged models of spin-glasses. *Physical Review B*, 17:4384–4403, 1978.

[KZ] R Kree and A Zippelius. Asymmetrically diluted neural networks. *Preprint.*

[Lau88] B Lautrup. The theory of the Hopfield model. *Lecture Notes*, 1988.

[Lau91] B Lautrup. Uniqueness of Parisi's scheme for replica symmetry breaking. *Preprint*, 1991.

[Moo84] M A Moore. Random systems in condensed matter physics. *Statistical and Particle Physics: Common Problems and Techniques. Proceedings of the* 26th *Scottish Universities Summer School in Physics 1983, edited by K C Bowler and A J McKane (SUSSP Publications)*, pages 303–357, 1984.

[MP88] M L Minsky and S A Papert. Perceptrons (expanded edition). *Massachusetts Institute of Technology Press*, 1988.

[MPV87] M Mezard, G Parisi, and M V Virasoro. Spin glass theory and beyond. *World Scientific Lecture Notes in Physics Volume 9*, 1987.

[Pal89] R Palmer. Broken ergodicity. *Lectures in the Science of Complexity, Santa Fe Institute Studies in the Sciences of Complexity, lectures volume 1 edited by D Stein (Addison-Wesley Longman)*, pages 275–300, 1989.

[PFTV88] W H Press, B P Flannery, S A Teukolsky, and W T Vetterling. Numerical recipes in C: The art of scientific computing. *Cambridge University Press*, 1988.

[RSW91] A Rau, D Sherrington, and K Y M Wong. External fields in attractor neural networks with different learning rules. *Journal of Physics A: Maths and General*, 24:313–326, 1991.

[She90] D Sherrington. Complexity due to disorder and frustration. *Lectures in the Science of Complexity, Santa Fe Institute Studies in the Sciences of Complexity, lectures volume 2 edited by E Jen (Addison-Wesley Longman)*, pages 415–453, 1990.

[Ste89] D Stein. Disordered systems. *Lectures in the Science of Complexity, Santa Fe Institute Studies in the Sciences of Complexity, lectures volume 1 edited by D Stein (Addison-Wesley Longman)*, pages 301–353, 1989.

[TR90] A Treves and E T Rolls. Neuronal networks in the hippocampus involved in memory. *Proceedings of the* 11th *Sitges Conference on Statistical Mechanics of Neural Networks, L. Garrido (editor) (Springer-Verlag)*, 368:81–95, 1990.

[Wid61] D V Widder. Advanced calculus (2nd edition). *Prentice-Hall*, 1961.

[Won90] K Y M Wong. Private communications. 1990.

[WRS91] K Y M Wong, A Rau, and D Sherrington. Weight space organisation of optimised neural networks. *Oxford preprint, ref: OUTP–91–38S*, 1991.

[WS90a] K Y M Wong and D Sherrington. Optimally adapted attractor neural networks in the presence of noise. *Journal of Physics A: Maths and General*, 23:4659–4672, 1990.

[WS90b] K Y M Wong and D Sherrington. Training noise adaptation in attractor neural networks. *Journal of Physics A: Maths and General*, 23:L175–L182, 1990.

[YW91] H W Yau and D J Wallace. Enlarging the attractor basins of neural networks with noisy external fields. *Journal of Physics A: Maths and General*, 24:5639–5650, 1991.