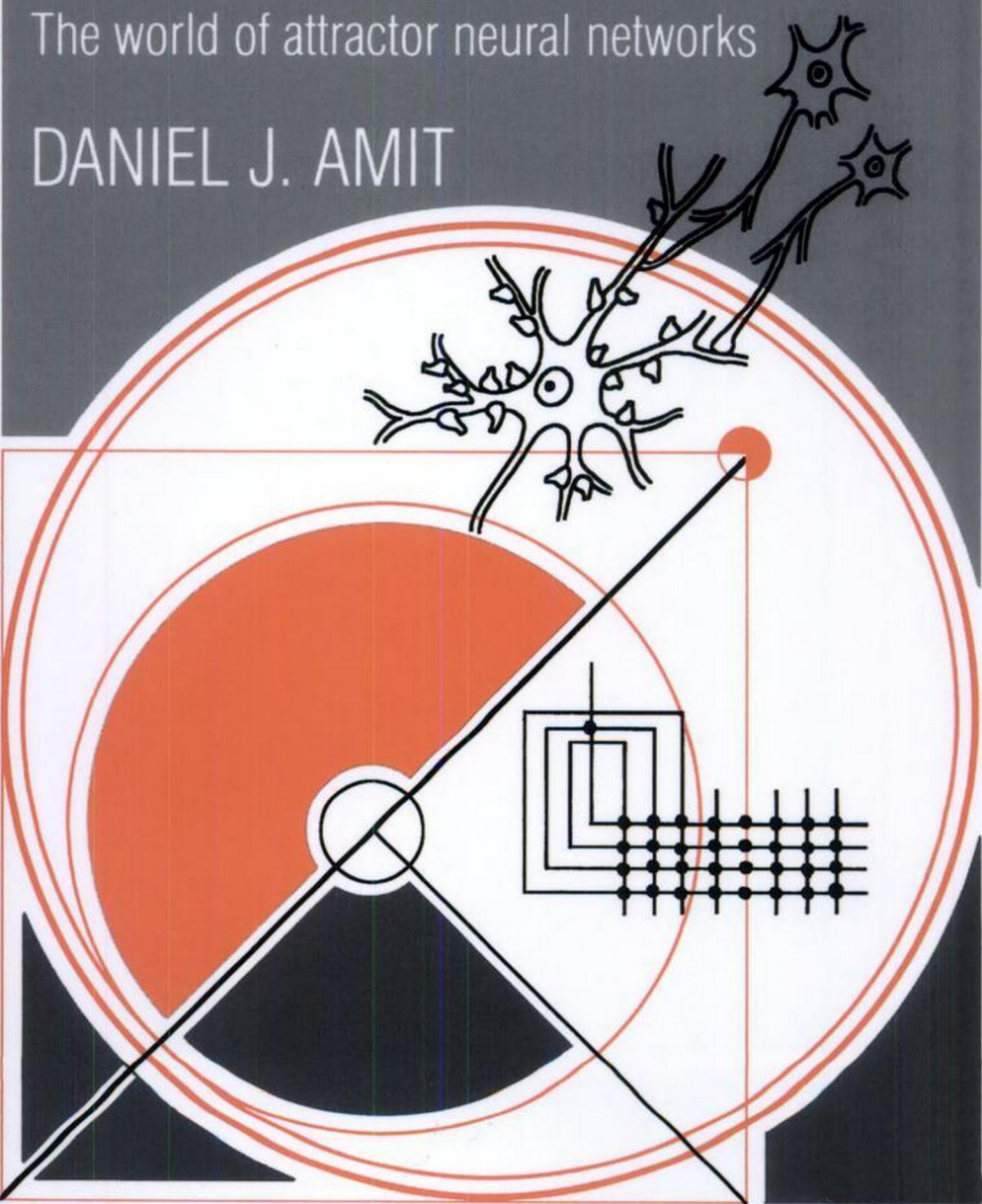
Modeling Brain Function

The world of attractor neural networks

DANIEL J. AMIT

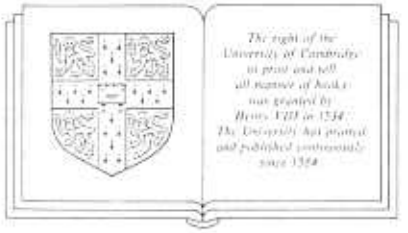


Modeling brain function

The world of attractor neural networks

DANIEL J. AMIT

Racah Institute of Physics



Cambridge University Press

Cambridge

New York Port Chester Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1989

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Amit, D. J., 1938-
Modeling brain function: the world of attractor neural networks
Daniel J. Amit
p. cm.
Includes bibliographies and index.
ISBN 0-521-36100-1
1. Brain – Computer simulation. 2. Neural circuitry. 3. Neural computers. I. Title.
[DNLM: 1. Models, Neurological. 2. Nervous System – physiology. WL 102 A517m]
QP376.A427 1989
591.1′88 – dc20
DNLM/DLC
for Library of Congress
89-15741
CIP

British Library Cataloging-in-Publication Data

Amit, Daniel J.
Modeling brain function: the world of attractor neural networks
1. Artificial intelligence
I. Title
006.3

ISBN 0-521-36100-1 hard covers

Source of epigraph (p. xviii): Shakespeare, *Hamlet* 2.2.436

Contents

Preface

This book summarizes in some detail the ideas, techniques and results developed in the last 5-6 years in the physics community about the collective properties of large assemblies of neurons. The subject has been, and still is, a source of great excitement among physicists the world over and new original ideas are generated incessantly. This enthusiasm has produced a wealth of new concepts and new detailed results which has not gone unnoticed outside physics departments. Biologists have begun to ask themselves whether the properties that physics anticipates in neural networks can indeed be observed and whether they provide useful theoretical guides for the empirical investigation of brain activity; computer scientists would not rule out these ideas as candidates for coherent parallel processing; psychologists and neurologists have been expecting some new useful metaphors for interpreting behavioral disfunction; cognitive scientists study the new concepts in their continued struggle with the elusiveness of processes of mind, even on the most elementary levels; and technologists have added, of course, Attractor Neural Networks to the list of future industries for sale.

One explanation for this impact of the study of neural networks seems to be in the type of new concepts that have been generated. They appear plausible upon introspection and they are based on elements with biological flavor. Another attraction is the clarity, the wealth and the detail provided by the quantitative analysis of the properties of such networks. It would have been easy and uncontroversial to write a book restricted to the technical details and the results. Physicists

would have liked it better and others would have ignored it without regret or complaint. I have set myself a more challenging task, of combining the presentation of the quantitative results, with their full technical beauty, with an attempt to communicate across disciplines. It has created multiple tensions. First, it has significantly expanded the exposition of every subject, adding a non-technical description of each result as well as an attempt to connect the result to topics in other disciplines, even if by way of speculation or metaphor. Second, it has required whole chapters, e.g., Chapter 3, to summarize for the diligent non-physicist, who may like to confront the full theoretical apparatus, an entire culture of physics. Thirdly, it implies that no uniform language could be used throughout the monograph and often the text is in English. I have tried to make it manifestly clear when words attain restricted formal meanings.

I have tried to inform every directly concerned discipline, i.e., biology and cognitive science, that we are conscious of some of our limitations. I had to inform the biologist that we know we have not done justice to his real neurons, and this had to be done without embarking on a full course in neuro-physiology. I had to admit that we have not solved any important outstanding problem in cognitive science, yet insist on the fact that we are introducing novel relevant concepts. This position crosses boundaries of theoretical arguments in cognitive science. I had to indicate that I am conscious of them, react to them, yet I could not expose the full background. The need to address so many concerned audiences and remain brief was perhaps the hardest part of the task, leaving behind many unresolved tensions.

These issues have led to the writing of the first chapter. It might have more appropriately been a chapter of conclusion. Had I had my ideal reader, he would have read this chapter superficially on first reading. Then, if the rest of the book would have kept up his interest, he would have come back to reread the introduction. A slightly less ideal reader would skip the introduction on first reading and at best read it at the end. This should partially explain why quite a few of the concepts mentioned in the Introduction are not fully clarified within this chapter, i.e., Chapter 1, and preserve some of their colloquial sense. The main part of the book is dedicated to the clarification of the new concepts which these models put forth.

The introduction is intended for three classes of imaginary readers: the biologist who is convinced that physicists do not know that the

world is complicated and therefore cannot possibly bring out of neurons anything of interest; the expert in artificial inteligence who has to be convinced that the Attractor Neural Networks are at least something new (if not something important); and the cognitive scientist-philosopher to whom I feel I owe an explicit statement of my commitments on several basic issues.

Chapter 1 is, therefore, an essay on the rest of the book, trying to identify the commitments implied by the theory. The course of events has, of course, been the reverse. During the whole period of the development of the subject, no commitments were being made while results were being derived. Success breeds responsibility, and if physics intends its results to be taken seriously, it has to commit itself. The stands expressed in this essay may be at odds with some accepted theories. This may be the result either of the fact that new concepts can be unambiguously introduced via the new approach, or because the epistemological approach is inconsistent, which is the only way, I can conceive, that it could be wrong. Such an adverse eventuality does not, of course, reflect on the correctness of the technical developments, which have withstood many theoretical and practical tests. Inasmuch as the outlook argued for in the Introduction may appear controversial, it should be considered as an opening of a discussion, rather than a conclusive state of affairs.

In order to facilitate the navigation in the maze, I have provided a very detailed table of contents which may indicate how to exercise the hopscotch (à la Julio Cortazar). **Chapter 2** should be read by all non-physicists since it provides a detailed explanation of what *attractors* are. Physicists may choose to read only the first two sections which describe how network dynamics follows from simplified neural interaction. **Chapter 3** is a pedagogical chapter (I hope), with little direct connection to neural networks. It has the following objectives:

- To the non-physicist, it should convey the idea of non-ergodicity in a noisy system, which makes *cooperative* or *collective* behavior non-trivial.

- It should make clear the connection between the treatment of free-energies and that of dynamical phenomena such as neural networks.

- It should give a hint of the context from which physicists have been drawing their intuitions.
- Some physicists may find Sections 3 and 4 useful as a different perspective on the connection between dynamics, *order-parameters*, *free-energies* etc.

Chapter 3, which has been the hardest to write, can be skipped by everyone.

Chapter 4 is part of the main story. This is a good point to start reading the book, looking back at previous chapters when certain arguments become mysterious. The first three sections are intended for general consumption, giving simple technical rule-of-thumb tools for the identification of attractors. Sections 4 and 5 are a gradual introduction to the wonderful world of *mean-field* theories. They can be skipped at no risk, unless one intends to analyze one's own new model. **Chapter 5** is a step beyond simple single pattern attractors, on one possible road toward structured cognitive processes. It is almost all carried out on the technical level of the rule-of-thumb tools of Chapter 4, and could and should be followed by all readers who could master Section 3 in Chapter 4. Readers with no interest in any technicalities should read only Sections 5.1, 5.3.1, 5.4.1–5.4.4. **Chapter 6** is about limits of performance. It again can be read on three levels. The totally non-technical reader can do with Sections 6.1, and 6.5 and the prose in Section 6.6. The rule-of-thumb reader should add Sections 6.2 and the full 6.6. Sections 6.3, 6.4, together with the appendix, are the pinnacle of the technical side of the book. One should have a very good motivation before embarking on these sections.

Chapter 7 describes the extent of robustness of the results of the preceding three chapters. This chapter is mostly non-technical and the few formulae spread around are for purposes of definition of the type of variation under which robustness is investigated, rather than for purposes of derivation. Yet one may get the main points by skipping over Sections 7.2.3, 7.2.4 and 7.3.4 and the appendix. **Chapter 8** deals with the storage of hierarchically organized data structures. Its main points are summarized in Sections 8.1, 8.2.1, 8.2.4, 8.3.3–8.3.4, 8.4. **Chapter 9** is a description of the nascent state of a theory of learning, in the context of ANN's. The main ideas can be found in Sections 9.1 and 9.3.1. One level of technical material deals with proofs of the convergence of learning algorithms – Sections 9.2.1, 9.2.3. Another level

is involved in the definition of some learning scenarios – Sections 9.3.2–9.3.4. Finally, **Chapter 10**, about hardware, is purely descriptive.

In conclusion, I should stress that this book does not represent an effort toward impartiality. It is the outgrowth of a special intense experience that I was fortunate to have in the collaboration with H. Gutfreund and H. Sompolinsky. They should have coauthored this book with me, since much of what it contains, that is solid, was born in our collective effort. Without them this book would never have come into existence. Other priorities have left me alone in the field, and I can only pray that my precious collaborators would not object too much to the context into which I have inserted our joint technical work. They should not share in the blame for speculation, metaphor and polemic.

There is a long list of credits. First to students – I. Kanter, A. Crisanti, A. Treves, M. Aharoni, Y. Stein – who worked with us at various stages. There is a major intellectual debt I owe John Hopfield and Gerard Toulouse, who have also been a source of encouragement in the process of writing the book. My ideas have also been much affected by M. Abeles, V. Braitenberg, G. Parisi, M. Virasoro, the late E. Gardner, B. Shanon, D. Andler and H. Atlan. I owe a special debt to Profs. R. Gavison, of the Law School, and D. Lehmann, of the department of Computer Science, who have read substantial parts of the manuscript and have helped me with comments and criticism.

Rome
November, 1988

If it live in your memory, begin at this line —
let me see, let me see

1

Introduction

1.1 Philosophy and Methodology

1.1.1 Reduction to physics and physics modeling analogues

When physics ventures to describe biological or cognitive phenomena it provokes a fair amount of suspicion. The attempt is sometimes interpreted as the expression of an epistemological dogma which asserts that all natural phenomena are reducible to physical laws; that there is an intrinsic unity of science; that there are no *independent* levels or languages of description, only more or less detailed ones. The intent of this section is to allay such concern with regard to the present monograph, which remains neutral on the issue of reductionism. Yet, before explaining the conceptual alternative, of analogies to physical concepts, which has informed the work of physicists in the field of neural networks, it is hard to resist a few comments on the general issue of reductionism, as well as an expression of our own commitment.

It should be pointed out that the misgivings about reductionism cast many shadows. Biologists often still harbor traces of vitalism and feel quite uncomfortable at the thought that *life*, *evolution* or *selection* could be described by laws of physics and chemistry. Cognitive scientists resent both the reduction of cognitive phenomena to neurobiology[1,2] as well as to computer language[3]. A physicist who reads Fodor's proof of the impossibility of reduction between different levels of description should be troubled about the connection that was so ingeniously erected by Boltzmann and Gibbs between the

macroscopic phenomena of thermodynamics and the underlying microscopic dynamics of Newton, Maxwell and Planck. It is rather curious that Fodor, who is aware of the successful reduction of the theory of heat to molecular theory, does not submit it to the acid test of his proof. In particular, the *kind*[1] 'heat transfer' is *ab initio* no more expressible in terms of molecular *kinds* than the kinds appearing in Gersham's law (a law in economic theory used by Fodor) are in terms of psychological or biological kinds. Instead a physicist would typically conclude, more pragmatically, that a reduction can be so defined as to be proved out of existence – *à la* Fodor. Alternatively, reduction can be given a very intuitive sense in which it not only exists but is extremely useful and productive.

Admittedly the physicist's reduction may be considered somewhat weak, in that it is neither necessary nor unique (in that it cannot be logically deduced and there can, in principle, be more than one consistent reduction). It remained undisturbed, for example, when Newtonian mechanics was supplemented by special relativity and not even when quantum mechanics revolutionized the description of the microscopic world. Yet weak as such reductionism may seem, and we will not even engage in articulating it, one can hardly imagine the discovery of the wonderful world of universality classes in the thermodynamics of phase transitions, or the exceptional phenomena of spin-glasses, without it.

It appears that the question of reductionism, much like that of determinism, is here to stay. Our commitment to reductionism stems not from a theorem that it necessarily governs all of science, but rather because 'it has been the scientifically productive idea in...this century, and the present essay is an attempt to take that line a step further'[4]. To be somewhat more speculative, it may constitute one of the main defining characteristics of scientific activity.

Hopfield[5], the instigator of the present wave of research in neural networks in the physics community, states that 'the brain is a physical system.' This may indeed sound like a call for a reduction of thought processes – memory as the particular case in point – to physics. Yet the enterprise is innocent on this count. In fact, as will become entirely clear throughout this volume, all concepts originating in physics are used as analogues. This is the fate of energy, field, relaxation etc. These analogies originate in the mathematical similarity between

[1] A concept which is natural in a language of description at some given level.

models of schematized neurophysiological data and models of certain physical systems, notably magnets. See e.g., Section 3.2. The physicists' effort is to affect a reduction of some simple mental processes to modeled biology. The resulting mathematical *form* of the models is quite familiar in statistical physics, and consequently many deep results as well as detailed techniques can be transported over to save effort and provide insight. There is of course a second level of possible reduction, that which expresses the elements of the modeled neurophysiology in term of physics and chemistry. This appears to be a simpler task, perhaps because some of it has already been convincingly accomplished. The best example being the Hodgkin-Huxley equations for the dynamics of the ionic concentration gradients across neural membranes[6,7]. It will not be discussed here.

1.1.2 Methods for mind and matter

The commitment to reductionism implies *inter alia* a commitment to a standard method. The alternative is to expect that methods of inquiry and reasoning may depend on the subject matter, especially when it touches upon the living or the thinking. We make this commitment as well. This again is an issue which has raised much polemic over the ages. As Hebb puts it[4]

> All science, from physics to physiology, is a function of its philosophic presuppositions, but psychology is more vulnerable than others to the effect of misconception in fundamental matters because the object of its study is after all the human mind and the nature of human thought, and it is very easy for philosophic ideas about the soul,..., or about determinism and free will, to affect the main lines of theory. As long as the ideas are implicit they are dangerous; make them explicit and perhaps they can be defused.

While this also is an issue that cannot be resolved one way or the other, the only way of making systematic progress is to adopt the Aristotelian point of view that

> ...if there is not some single common method for the investigation of particulars, then putting our inquiry into practice becomes still more difficult. For *we will have to grasp in*

> *each case what the method of inquiry should be*[8]. (Emphasis added).

All this ties in nicely with yet another methodological maxim expressed in a home-made poster of a mathematician friend: 'If there is something you do not understand about a problem, there surely is a simpler problem you do not understand something about.' Physics has been traditionally faithful to this idea, but has been applying it bottom-up. Complex situations are preferentially studied on top of simple, well-understood solutions. This is our attitude here as well.

The particular simplifications used in constructing the physicists' foray into the study of mind are the subject of the next section. But even on the more abstract level, where issues pertaining to the general nature of constraints on cognitive processes are discussed, a physicist is likely to feel that natural language, for example, is much too complicated a subject-matter for a starting point. He is likely to try to construct a structure of increasing complexity consisting of definite realizations of simple processes possessing cognitive flavor. *One* of the main criteria for the selection of these stages is their analizability. Being definite and specific, their properties and constraints can be studied in a non-abstract manner, avoiding mysterious conclusions which are brought about by the opaqueness of complexity.

An expression of the resulting effect has been succinctly summarized in eulogizing the perceptron[9]:

> ...Although we do not have an equally elaborated theory of 'learning,' we can at least demonstrate that in cases where 'learning' or 'adaptation' or 'self-organization' does occur, its occurrence can be thoroughly elucidated and carries no suggestion of mysterious little-understood principles of complex systems. Whether there are such principles we cannot know. But the perceptron provides no evidence; and our success in analyzing it adds another piece of circumstantial evidence for the thesis that cybernetic processes that work can be understood, and those that cannot be understood are suspect.

It may very well be the case that the resulting structure will be disappointing, like the perceptron, in that it will leave no mystery to satisfy our awe at our own mental activity. Still, some new insights may

surface with implications for cognitive psychology, neurobiology, computer science or technology. The theoretical construction may become very long and elaborate and it may account for an increasing variety of phenomena of *cognitive flavor* and yet persist in falling short of satisfying the great expectations accompanying questions of mind. This may then be considered as supporting the lingering sense, hinted at by the quotation from Minsky and Papert[9] above and strongly advocated by Turing[10], that mental phenomena are but an expression of a very complex structure operating on relatively simple principles. Moreover, it appears that whenever a body of data of controllable scope is considered, it is *simple* principles that are brought to the surface, rather than *magic*.

1.1.3 Some methodological questions

Many methodological assumptions permeate this book. I am neither aware of all of them nor do I intend to devote excessive space to those I am aware of. A few will be discussed; those that have led to some difficulties of communication across disciplinary lines.

The first assumption to be discussed is the attitude to verification and/or falsification. It is a noble desideratum that a theory make contact with experiment. Physicists usually search for confirming evidence. Biologists, influenced by Popper[11] via Eccles[12], require falsifiable predictions. The theory of *attractor neural networks* (ANN) in all its recent elaborate developments has engaged in providing a minimal amount of propositions which can be confronted with experiment. Given that both neurobiology and psychophysics are inundated by empirical data, the pressures exerted on the emerging theoretical framework must be fended off by an explanation.

There is a sense in which the theory has been amply confirmed. In many instances systems have been constructed, either as hardware implementations or as computer simulations, in such a way that the underlying dynamical mechanisms of the individual component 'neurons and synapses' have been the same as those described in the models, see e.g., Chapter 10. These artificial networks can be considered as experimental setups. Their *cooperative emergent* properties have been compared with those predicted by the analysis of the models. The agreement has been truly impressive, as this book will testify. Yet this important fact will please no experimenter who records, using very

ingenious techniques, the electrical activities in the cortex of cats or monkeys, nor the psychologist trying to account for data on perception, dislections, attention, etc.

It should be first of all emphasized that a major task of any theory of neural networks is to produce *exceptional* input-output relations. They have to be exceptional in that they should correspond, even if initially only in a metaphorical sense, to our intuitions about cognitive processes. Attractive features are *biological plausibility; associativity; parallel processing; emergent behavior (cooperativity); freedom of homunculi; potential for abstraction.* Then, if any of those features is captured by the model, it has to prove robust to the type of disorder, fluctuations, disruptions that we imagine the brain to be operating under.

All the big italicized words in the previous paragraph are used here in their small sense. The clarification of their meaning and its realization are what this entire book is all about. Telegraphically:

- *biological plausibility* is simply the requirement that the elements composing the network not be outlandish from a physiological point of view;
- *associativity* is the impression that many similar inputs are basically collapsed on a prototype for purposes of cognition and manipulation – a picture viewed from many angles and in different light and shading represents a single individual;
- *parallel processing* is the articulation of the tension between the slow basic cycle-time of neuronal processes and the impressive speed with which the system as a whole reacts to tasks that are prohibitive to high-speed serial computers;
- *emergent behavior* stands for an input-output relation which is rather unlikely (non-generic) given the system's elements. This, it seems to me, is a significant component of the mystery of mind;
- *freedom from homunculi* is the eternal goal of eliminating little external observers which should assign ultimate meaning to outputs;
- *potential for abstraction* is the ability of a network to operate similarly on a variety of inputs that are not simply associated

in form but are classified together only for the purpose of the particular operation.

It is a theoretical construction which encompasses these features that one is after. Large parts of it have been successfully erected. But the extreme difficulty involved in the specification of the input to a cortical area, the connection to the (motor) output, and the relative importance of various physiological constraints all advise that formulations keep as many options open as is conceivably possible. At the present stage, the broad spectrum of options is narrowed only under the pressure of biological information or presumed cognitive input-output relations. Thus neither verification nor falsification provide useful methodologies, except *vis a vis* artificial networks, as mentioned above.

To the extent that the program is successful, there will emerge a narrow class of model networks which exhibit a broad spectrum of properties. Moreover, these models will be largely analytically lucid. It is as likely as not that empirical reality about the central nervous system will be temporarily redefined, by the theoretical framework. For example, electro-physiological activity in a region of cortex associated with some sensory modality (visual, auditory, somatic etc.) may not show the type of behavior identified as the emergent dynamics. This may be interpreted as a refutation of the theoretical construction. Alternatively, one may consider it more beneficial, provided the model is attractive enough, to use a rather common physicist's stratagem: to argue that the experiment has missed the theory, rather than the other way around. It could be, for example, that the cortical area where cognitive processing of a particular sensory input or inter-modal association had been mistakenly identified by the experiment which is presumed to refute the model. With 10^{11} or more neurons or 10^7 elementary networks, of some 10^4 neurons each, to choose from, precious time can be gained, but not eternity of course. One must keep in mind that the advocacy of such a stratagem is by no means a depreciation of the experimental effort. Rather it is a natural component in the organic dialectical relationship between theory and experiment. As such it has served as a respectable and productive methodology.

And another, minor, methodological point. Given that the model hovers somewhere between the biological, the psycho-physical and the cognitive, there is a tendency to oscillate between the inclusion of additional biological constraints and the development of increasingly elaborate structures and computational features. As the biological substrate

becomes more realistic the maneuvering with structure becomes more cumbersome. Elaborate structures are required for richer input-output relations to surface. We will therefore make it common practice that once a substrate complication is well understood, it will be removed to facilitate the structural (and possibly the cognitive) end of the description.

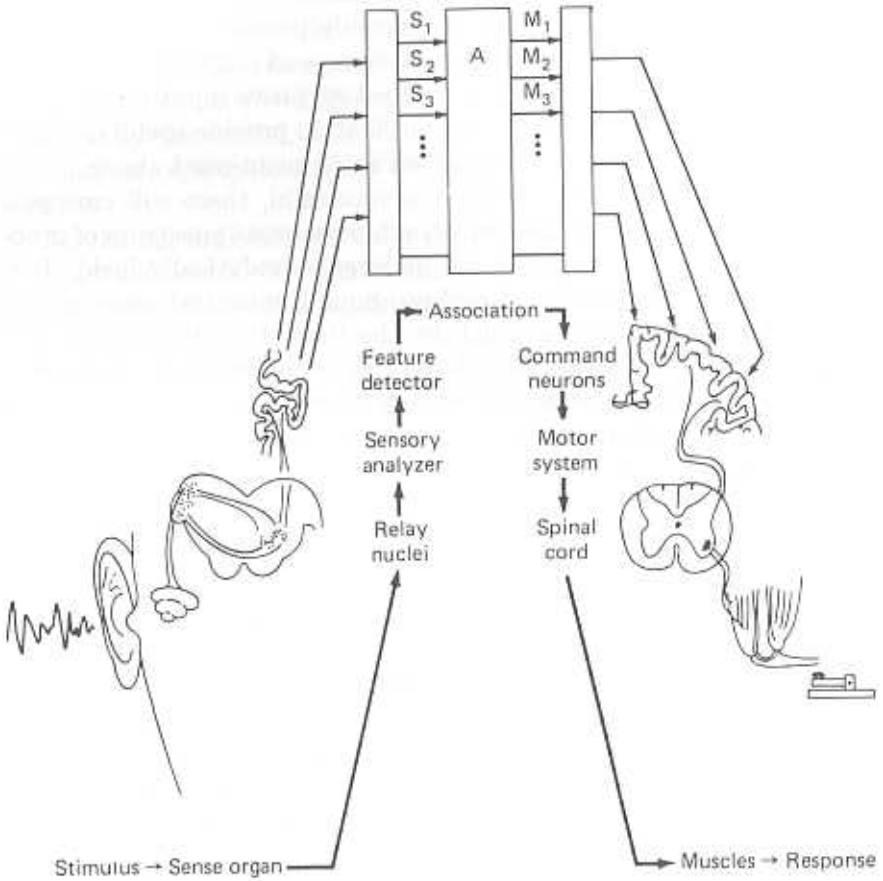


Figure 1.1: A schematic representation of the brain as a system of sequential stations – input (S), associative central processing (A) and motor output (M). (From ref. [15], by permission.)

1.2 Neurophysiological Background

1.2.1 Building blocks for neural networks

The intention of this book is to contribute to a perspective of mind which is basically biological. Hence a description of the biological situation is called for. But for someone who does not actually perform neurophysiological experiments it will not be possible, for a long time to come, to surpass Patricia Churchland's[13] description of the state of the art in the neural sciences. The reader is, therefore, referred to this outstanding text for a general overview of the field, as well as for an extensive list of more specialized references. Only those features which are essential for the construction of the model will be summarized here.

The basic elements are, naturally, *neurons* and *synapses*. There is a fairly large variety of types of neurons in the human nervous system – variations are found in size, in structure and in function. For present purposes it will be assumed that deviations from the 'canonical' type are related to specialized systemic function, such as sensory input or motor output. This is not a statement of fact. Rather, it is a choice of context. We subscribe to a coarse division of the nervous system into three parts – *input, central processing, output*, as is expressed in the drawing in Figure 1.1. This is a division into neural systems, each rather complex in structure, with different functions and computational roles. Here the emphasis is on the central part – evolutionarily the latest[14], where we expect the maximal amount of universality and hence simplicity. If the underlying principles of the workings of the central nervous system depend on details of the structure of individual neurons, it is unlikely that physics will contribute much to their clarification. Nor would it be a great tribute to the distance travelled by our species along the evolutionary tree. Beyond a certain level, complex function must be a result of the interaction of large numbers of simple elements, each chosen from a small variety.

The Attractor Neural Networks (ANN's) will be constructed from neurons which have the canonical division into an input part (the *dendritic arbor*) , a processing part (*soma*) and a signal transmission part (*axon*). Such a neuron is depicted in Figure 1.2A, alongside its schematic representation B. At this schematic level such neurons can represent any of the several types of cortical neurons such as the Golgi, pyramidal, stellate or interneuron cells. The *neurons* communicate via

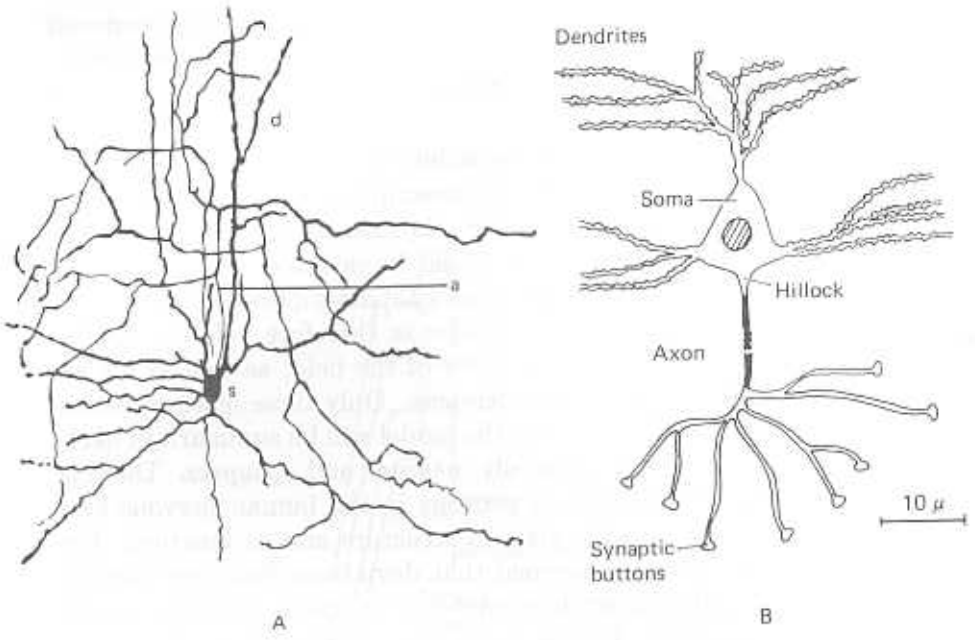


Figure 1.2: (A) A stained pyramidal cell. The rough processes (d) are the dendrites; the smooth ones (a) are the branching axons; (s) is the soma. (After ref. [16].) (B) Schematic representation of the same neuron. (After P. Peretto, unpublished.)

synapses, which are the points along the axon of the *pre-synaptic* neuron at which it can communicate the outcome of the computation that has been performed in its *soma* to the *dendrites* or even directly to the *soma* of the *post-synaptic* neuron. There may be more than one dendritic tree entering the soma. This can be seen schematically in the top left corner of Figure 1.3. The blow-up of a synapse in this drawing will be discussed in the next section, in the context of neural dynamics. The output part is the axon. Usually only one axon leaves the soma and then, downstream, it branches repeatedly, to communicate with many post-synaptic neurons.

For our concerns here, the most significant anatomical fact is that each neuron receives some 10^4 synaptic inputs from the axons of other neurons, usually not more than one input from one pre-synaptic

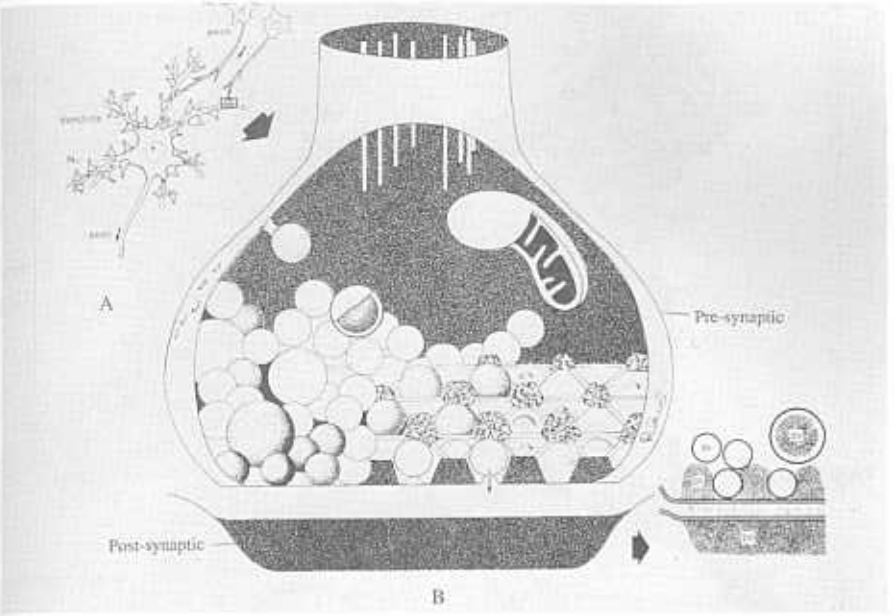


Figure 1.3: Drawing of the synaptic junction. (A) A collection of junctions connecting axons to dendrites. (B) An enlargement of one of the synapses, showing the vesicular grid and vesicles of neurotransmitter incorporated into the pre-synaptic membrane. One of those is opening into the synaptic *cleft*, under which is the post-synaptic membrane. (After ref. [17], by permission.)

neuron. This extremely high density of inputs is shown in Figure 1.4, in which each little bulb is a synapse of a neuron connecting to the surface a single post-synaptic cell. Conversely, the branching axons of each neuron form about the same number of synaptic contacts on other, post-synaptic, cells. Our cortex would, therefore, be a mosaic of assemblies of a few thousand densely connected neurons. These assemblies are the basic cortical processing modules. Their size would be of the order of a square millimeter. On a larger scale the connectivity will be assumed to be much sparser and with much less feedback, allowing for autonomous local collective, parallel processing and more serial and integrative processing of local collective outcomes.

Even though neurons of other morphologies frequently appear in recent neurobiology texts[16,19], they are usually associated with sensory functions, motor functions, or long distance transmission. Synapses

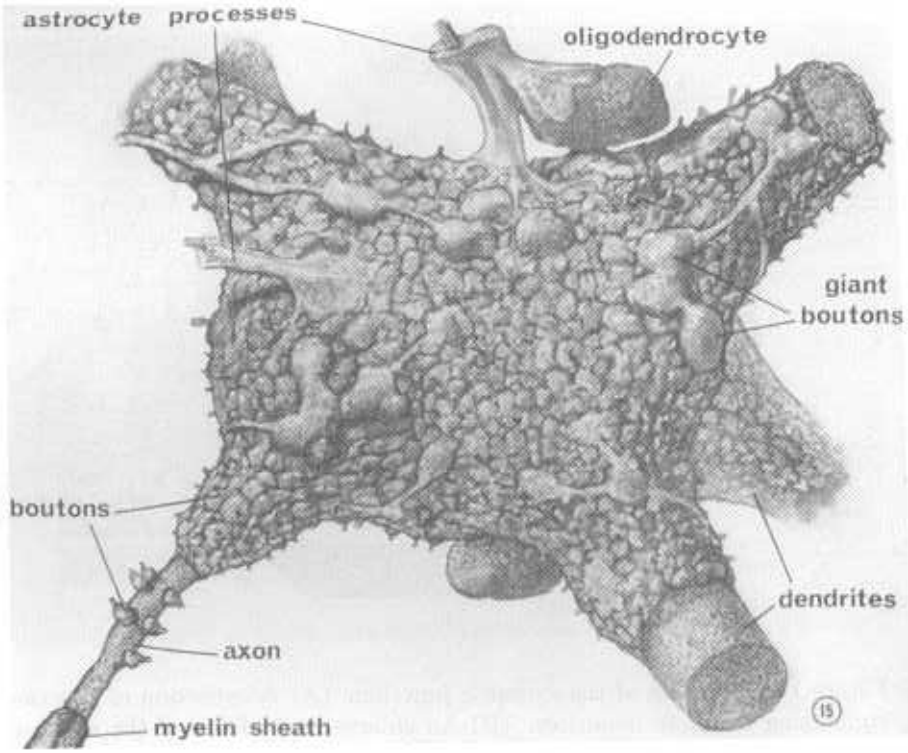


Figure 1.4: Synaptic end bulbs on the surface of a motor neuron. (From ref. [18], by permission.)

may appear between pre- and post-synaptic axons, or even between pre- and post-synaptic dendrites. Synapses of these types will also be ignored, assuming that they represent specialized remnants in peripheral regions, or are confined to lower species like invertebrates[16][p.43].

Eventually, we will allow for more than one type of neuron, but the variety will not be morphological but rather dynamical (see e.g., Chapter 7). Similarly for the synapses, they will all be basically axo-dendritic and chemical, but eventually we will allow for fast and slow ones (see e.g., Chapter 5).

1.2.2 Dynamics of neurons and synapses

The fundamental dynamical process of neural communication is depicted in Figure 1.5 and is based upon the following sequence[7] (the steps in the description follow the markings in the figure):

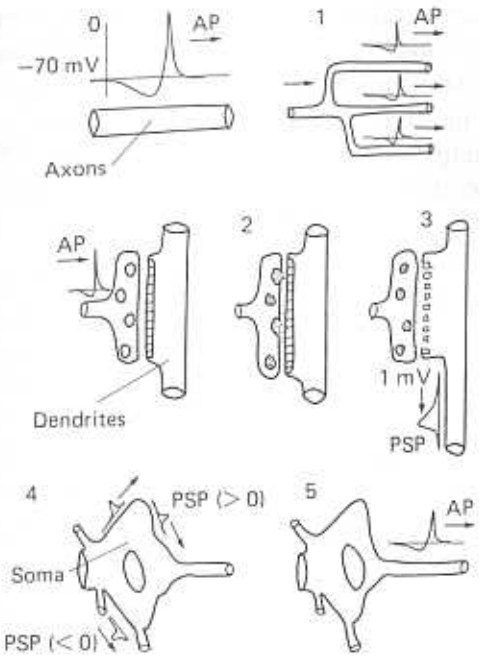


Figure 1.5: Stages in neuronal dynamics. (1) Signal in pre-synaptic axon (a); (2) signal replication at axonal branching; (3) transmission across synapse into post-synaptic dendrite (d); (4) diffusion along dendrites into soma (s); Computation in post-synaptic soma. For a full legend see text. (After P. Peretto, unpublished.)

1. The neural axon is in an all-or-none state. In the first state it propagates a signal – *spike*, or *action potential (A.P.)* – based on the result of the summation performed in the soma. The shape and amplitude of the propagating signal – the potential difference across the cell membrane – is very stable and is replicated at branching points in the axon. The amplitude is of the order of tens of millivolts. In the none state there is no signal traveling in the axon, rather there is a resting potential.

   The presence of a travelling impulse in the axon blocks the possibility of a second impulse transmission.

2. When the travelling signal arrives at the endings of the axon it causes the secretion of neuro-transmitters into the synaptic cleft. See e.g., Figure 1.3.

3. The neuro-transmitters arrive, across the synapse, at the membrane of the post-synaptic neuron. On the post-synaptic side these neuro-transmitters bind to receptors, thus causing the latter to open up and allow for the penetration of ionic current into the post-synaptic neuron. The amount of penetrating current per pre-synaptic spike is a parameter which specifies the *efficacy* of the synapse.

4. The post-synaptic potential (PSP) *diffuses* in a graded manner (unlike the spike in the axon) toward the soma where the the inputs from all the pre-synaptic neurons connected to the post-synaptic one are summed. The individual PSP's are about one millivolt in amplitude. These inputs may be *excitatory* – depolarizing the membrane of the post-synaptic neuron, increasing the likelihood of the appearance of a spike (firing), or they may be *inhibitory* – hyper-polarizing the post-synaptic membrane, reducing the likelihood of firing.

5. If the total sum of the PSP's arriving within a short period surpasses a certain threshold, which is the level at which the post-synaptic membrane becomes unstable against depolarizing ionic current flows[7], the probability for the emission of a spike, which is a manifestation of this instability, becomes significant. This threshold is again tens of millivolts and hence quite a number of inputs are required in order to produce a spike.

The cycle-time of a biological neuron of the cortex, namely the time from the emission of a spike in the pre-synaptic neuron to the emission of a spike in the post-synaptic one is about 1-2 milliseconds. This is the time the spike travels the full length of the pre-synaptic axon, the neurotransmitter crosses the synaptic gap and the post-synaptic potential difuses to the soma.

Following the dramatic event of the emission of a spike, the neuron needs time to recover. There is a period of 1-2 milliseconds in which the neuron cannot emit a second spike, no matter how large the depolarizing potential may be. This period is the *absolute refractory period* of the neuron. It sets the maximal spike frequency at 500–1000 per second. Such rates can appear at sensory input neurons, since the stimulus is externally determined and can be arbitrarily strong. In higher cortical areas rates are significantly lower. Even where bursts of

spikes are observed they may be as low as 150 per second or even, in some areas, not higher than 30 or 40 per second.[2]

1.2.3 More complicated building blocks

(For biologists only)

Inside a network, the situation is more complex. Following the short *absolute refractory period* the neuron recovers, but with a higher excitation threshold. Following a somewhat longer time interval – up to 5 or 7 milliseconds – the threshold returns to normal and the neuron can fire again with typical intra-network potentials. This is the *relative refractory period.* See e.g., Section 7.4. In a network the potentials are determined by the mutual interactions of the neurons and not by arbitrary external inputs. This prevents typical PSP's from becoming significantly greater than normal thresholds and makes it unlikely that a neuron will fire during its relative refractory period. It becomes one of the factors which limit the maximal spike rate in a network.

Similarly, a neuron that has not fired for a certain time, because its depolarizing potential has not reached threshold, loses its potential gradually, and new PSP can supplement decaying old PSP. These complications, quite relevant for the description of biological neurons, will be introduced in a second phase of the discussion, in Section 7.4, only to show that they do not modify any of the cognitive or computational features of the networks, except for a possible reduction of the maximum spike frequency.

Another neuronal variability, to be discussed in Section 7.3.3, consists of having two types of neurons, one type has only excitatory outgoing (efferent) synapses on its axons and another only inhibitory ones. This is a specificity which seems to govern most neurons. The simpler case, which will be considered for the most part, assumes that inhibitory and excitatory synapses appear randomly distributed on the terminals of the pre- synaptic neuron. It turns out that the introduction of this specificity has a very mild effect.

In addition to the single canonical nonspecific neuron, which has excitatory and inhibitory synapses scattered at random and which operates on a single fast time scale, we will find at least one additional type of neuron indispensable. This is a neuron with a longer

[2] I am indebted to Professor M. Abeles for impressing this fact upon me.

summation time in the soma and higher threshold. It will read and convey the output of the operations of the network, and will replace the homunculus. See e.g., Sections 1.4.3 and 1.4.4. A natural candidate would be the pyramidal cell, which is larger than the others and communicates over large distances. There may be a question as to whether the required neurons actually exist. This is, of course, a question for empirical neuro-physiology to answer. It should be pointed out that this type of neuron is no more than the kind of neuron that is often invoked by biologists, as a neuron sensitive to pre-synaptic firing rates.

As far as synapses are concerned, we will deal for the most part with a single type of synapse – a chemical synapse. Its excitatory or inhibitory nature will be included in allowing the efficacy to have a positive or negative sign. The unifying feature of these synapses is that they operate on the same time scale, less than a millisecond. This constraint will be relaxed to allow for slow synapses that communicate information about pre- synaptic activity a number of cycles in the past. The additional variability in synaptic delay times, we will show, has significant computational consequences, *inter alia* it introduces the possibility of operating on temporal sequences of patterns, to which we devote Chapter 5.

Most of the discoveries on the detailed properties of neurons and synapses have been made on invertebrate neurons or on motor-neurons of vertabrates[7,20]. There is still ongoing exciting research on the large, easily identifiable neurons of such creatures as squid, aplysia or tritonia. The initial studies on these simple systems have clarified the basic mechanisms of neural electrophysiology and biochemistry and most of the findings have provided a satisfactory description of more delicate neurons, such as those found in cortex. The universality of these operating principles along the evolutionary sequence is quite astounding. But there is no reason to extrapolate it uncritically to the operation required of neurons in neo-cortex.

As techniques improve, one probes the invertebrate neurons in ever increasing detail. It is then found[21] that they are not 'canonical'. For example, axonic branches sometimes do not simply replicate spike trains, different synapses can have very specific roles, etc. If, in fact, all this detail has to be carried along in analyzing the performance of multi-neuron networks, the task becomes hopeless and much of neurobiological research would lose its proclaimed motivation. In fact, it is

rather likely that neurons in simple creatures are complicated because they are few, while in vertebrates and in particular in neo-cortex, where they abound, they can afford to be simple. It is consistent with the idea that[4]:

> [evolutionary] development took place twice, independently, in vertebrates and invertebrates,....In the vertebrate the central nervous system forms a single mass of brain and spinal cord...but in the invertebrate...[it] consists of a series of separate masses (ganglia) connected by relatively thin neural strands. [p.48]

This will be our attitude here.

1.2.4 From biology to information processing

The description in Sections 1.2.2 and 1.2.3, above, indicates that the only way neurons can communicate the outcome of their computations to other neurons, or eventually, to muscles is through the emission of neurotransmitters. In saying this one has ruled out electrical synapses, through which neurons can exchange ionic currents directly. This assumption will be made throughout. In fact, we shall make an even stronger assumption, namely that

- **sub-threshold potentials do not lead to the release of neurotransmitters.**

In other words, neurotransmitters are released by spikes, or action potentials, only.

Neither of these two assumptions can be physiologically justified[16], we would like to hope that they are good approximations as far as the cognitive and computational properties of the neural networks are concerned. These assumptions simplify significantly the description of the complex, interwoven nets of neurons. The reason is that under these assumptions, information processing in the network will depend on only two sets of variables:

- **the distribution of spikes among the neurons**
- **the list of synaptic efficacies**

More specifically: Suppose there is no external input into the network. In a given time slice of the size of 1–2ms we can register which neuron is carrying a spike (action potential) and which is not. This is the instantaneous ditribution of spikes. The list of synaptic efficacies, which in the absence of learning is assumed fixed, prescribes the connectivity of the network. The distribution of spikes at any moment determines the synapses at which transmitter release will occur. Those are the synapses which emanate from neurons which happen to be carrying spikes. The list of synaptic efficacies will then determine how much every neuron will receive. This is the essential information needed by a neuron in order to compute its next state – of either spiking or not spiking – which is what keeps information processing going.

1.3 Modeling Simplified Neurophysiological Information

1.3.1 Neuron as perceptron and formal neuron

We now focus on the *logical* structure of a single neuron. The description of the previous section suggests the scheme of Figure 1.6:

- There is a processing unit, the large circle marked i, which represents the cell body, or *soma*.
- A number of input lines connect, *logically*, to the *soma*. They are depicted as lines with incoming arrows in the figure.
- Each input channel is a combination of a dendrite and a synapse, and altogether there are as many logical input channels to a neuron as there are synapses connecting to its dendrites.
- The input channels are activated by the signals they receive from the *logical* boxes to which they are connected. These little boxes are our pre-synaptic axons, and their logical nature is expressed by the fact that they can either activate the channel (carry a spike) or not activate it (sub-threshold activity in the pre-synaptic neuron).

To each input line one associates a parameter J_{ij} – the subscript i specifies the (post-synaptic) neuron we are considering, in anticipation of a multi-neuronal situation). The subscript j refers to the various input channels to this neuron. The numerical value of J_{ij} is the *synaptic*

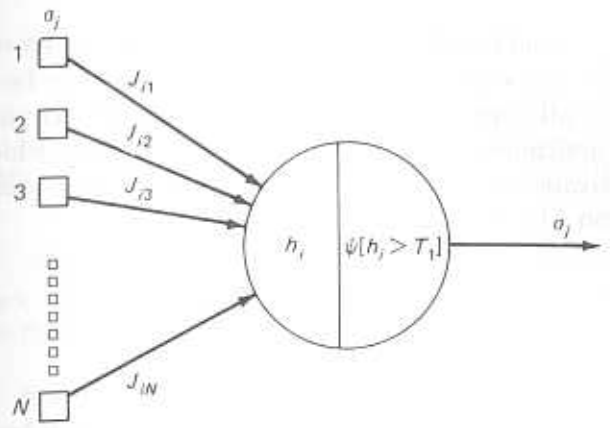


Figure 1.6: The logical structure of the neuron as a perceptron. J_{ij} are the efficacies of synapses coming into neuron i, represented by large circle. σ_j are 1-0 variables representing the arrival or non-arrival of a spike along the pre-synaptic axon connecting neuron j to i, h_i is the PSP and $\psi[h_i]$ is the decision function of the neuron. If neuron will (will not) fire, σ_i will take the value 1 (0), respectively.

efficacy which determines the amount of post-synaptic potential (PSP) that would be *added* to the soma i if channel j were activated.

There is a single *logical* output line, represented by the line with an outgoing arrow in Figure 1.6. A single output line expresses the *logical* fact that our neuron produces a single relevant output – it either emits a spike or it does not.

The operation of the unit is as follows:

- At any given moment, some of the logical inputs are activated.
- The soma receives an input (PSP) which is the *linear* sum of the efficacies J_{ij} of those channels that *were* activated.
- The sum of PSP's is compared to the threshold value of neuron i and the output channel is activated if it exceeds the threshold. Otherwise it is not.[3]

This process can be further formalized by ascribing to each of the little boxes a variable σ_j which can take on the values 1 and 0, indicating

[3] We will ignore for the moment the fact that this decision may be non-deterministic. This fact will become part of our context as we go along.

whether the channel to which the box is connected is active – ($\sigma_j = 1$), or is inactive – ($\sigma_j = 0$). The PSP at neuron i can now be expressed as the sum of **all** synaptic efficacies of the channels arriving at that neuron each multiplied by the corresponding σ-variable, which ensures that only activated channels contribute. In other words, denoting the PSP at neuron i by h_i, we can write:

$$h_i = \sum_{j=1}^{N} J_{ij}\sigma_j, \tag{1.1}$$

where N is the number of boxes, i.e., pre-synaptic neurons.

The operation of the little machine in Figure 1.6 can be expressed by the logical truth function

$$\sigma_i' = \psi[h_i > T_i] \tag{1.2}$$

where $\psi[\,]$ is a function which is 1 if the statement in square brackets is true and is 0 otherwise. The variable σ' indicates, in neuronal language, whether a spike will appear in the output axon.

Since the truth function, Eq. 1.2, is computed from variables σ_j which are themselves zeroes and ones, each of these variables can also be considered as a truth function of some statement. This then leads us directly to McCulloch and Pitts' formal neurons[22] and from there to Rosenblatt's perceptron[23,9] to which we now turn.

1.3.2 Digression on formal neurons and perceptrons

Formal neurons

As early as 1943, McCulloch and Pitts[22], in a very formal paper, showed that simplified neurons can be combined into little temporal sequences, at the end of which there would be a single output neuron, whose activity will be the truth value of any binary logical operation represented on the input neurons. With such elementary networks, it was argued, one can construct compound networks which can perform any operation of the calculus of propositions. One essential new element in this construction had been the introduction of *time*. If neurons are to compose a computer of propositional calculus, then one must take into consideration the fact that if the spikes arriving on the axons of two neurons represent truth values of some two propositions, then the

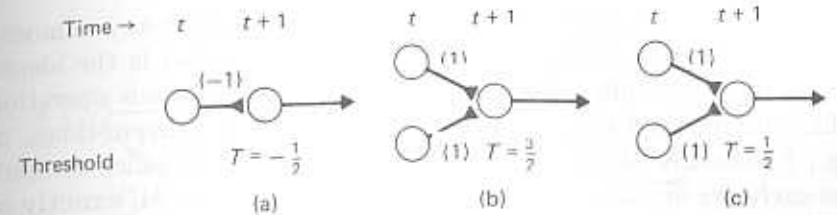


Figure 1.7: Three elementary logical operations (a) **negation**, (b) **and**, (c)**or**. In each diagram the states of the neurons on the left are at time t and those on the right at time $t + 1$.

neuron computing the logical operation on these two input propositions can give its answer only a cycle-time later than the time of production of the inputs. This cycle-time, which will be chosen as a time unit, is the spike-to-spike minimal time lapse, discussed in Section 1.2.2, above. The McCulloch-Pitts *formal neuron* is basically the entity described in Figure 1.6, usually with a small number of incoming channels. Each neuron has its proper threshold and the channels have numerical weights. The weights and the thresholds are chosen in such a way that at every *cycle-time* some logical operation is performed. The simplest such arrangements are shown in Figure 1.7. They compute (a) negation of a proposition, (b) the conjunction (**and**) of two propositions and (c) the disjunction **or**.

The diagrams should be read as follows: in (a) the left neuron emits a spike at time t ($\sigma(t) = 1$) if some proposition A is true. If the channel has the weight -1, as indicated, and the right hand neuron has the threshold $T = -0.5$, then at time $t + 1$ this neuron will not spike. If A is false, the left neuron does not spike at time t ($\sigma(t) = 0$) and consequently the right neuron will spike at $t + 1$. In other words, the spiking of the right neuron is the assertion of the *negation* of the proposition affirmed by the spiking of the left neuron. Diagram (b) in Figure 1.7 reads: the right neuron will spike at $t + 1$ if and only if **both** neurons on the left spiked at time t, indicating that the propositions tested by these two neurons have **both** been true. The parameters in (c) ensure that the right neuron will spike if at least one of its two inputs carried a true proposition.

The lore of logic has it as a fundamental proposition that one can construct any binary boolean operation from as little as the negation and one of the other two operations represented in the Figure 1.7, and

then, by the associative law, a boolean operation with any number of propositions[24]. The only additional element needed is the identity function, to provide a delay when outcomes of previous operations, which are required simultaneously, are produced at different times. See e.g., Figure 1.8. Thus, a network of formal neurons which computes the *exclusive or* (**xor**) of two propositions, which is true if **exactly** one of the two is true, could look like Figure 1.8. Note that if the truth values of the two propositions were represented on the left at time t, the result will appear at time $t+3$. Note also that we had to resort to an identity delay to keep the $(A \vee B)$ while the $(A \wedge B)$ was being negated.

It is a simple exercise, of mere historical value, to prove that no choice of weights and thresholds would allow the formal neuron to compute this same **xor** in a *single* time step. See e.g., ref. [9]. The example in Figure 1.8 shows that with a few steps, this computation can be accomplished. This is, of course, a consequence of the sufficiency of the elements in Figure 1.7 for the construction of arbitrary predicates. But, one should keep in mind that this formal universality requires real time, which increases with the complexity of the proposition. Moreover, neurophysiological evidence indicates[15] that it takes a few tens of excitatory PSP's to influence significantly the probability that a spike will be emitted by the post-synaptic neuron.

Perceptrons

Rosenblatt's *perceptron*[23] is a natural extension of the formal neurons of McCulloch and Pitts.[4] The basic idea is that a formal neuron, much like a biological one, can have inputs from many more than one or two channels. In fact, we have already observed in Section 1.2.1 that neurons in the cortex receive inputs from an extensive number of other neurons. Hence, one expects that the class of computations that could be performed at every time step, by a device of the type depicted in Figure 1.6, would be much wider than the elementary logical operations. Just for comparison, we reproduce here the drawing of the perceptron from Minsky and Papert's book. See Figure 1.9.

[4]It is quite remarkable that the perceptron idea, with most of its implications, has developed on at least one more, seemingly disconnected, branch. By 1960 Widrow at Stanford had a mechanical perceptron which classified and learnt[25]. For a mysterious reason this work has never appeared in mainstream periodicals.

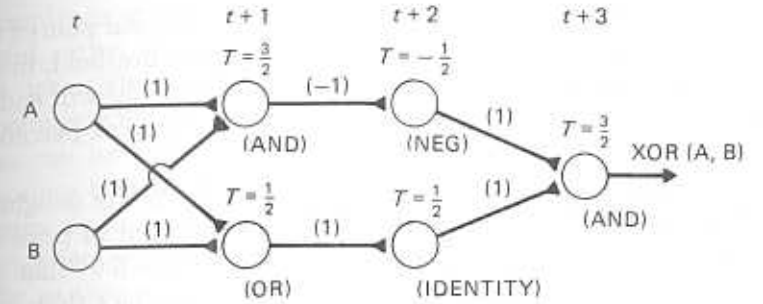


Figure 1.8: The construction for the exclusive or (**xor** by the route $[A \vee B] \wedge [\neg[A \wedge B]]$.

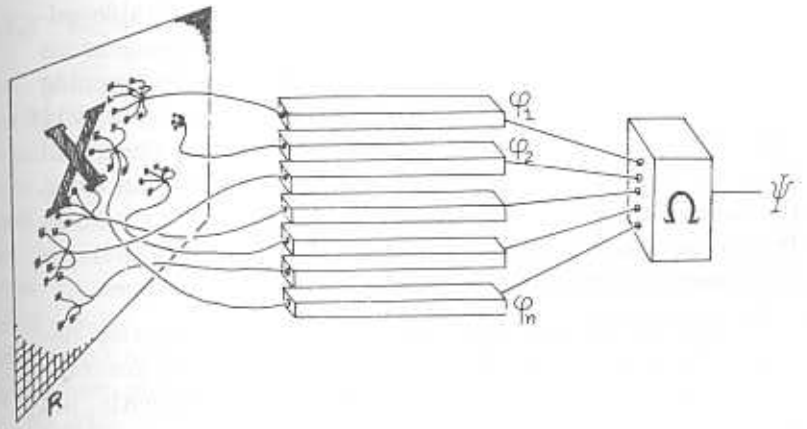


Figure 1.9: The perceptron as visualized by Minsky and Papert. (From ref. [9], by permission.)

The correspondence with Figure 1.6 should be apparent. The ϕ_i are our σ_i and the box marked Ω corresponds to the central part of the neuron in Figure 1.6, that part which sums the weighted inputs and decides, by a linear threshold computation, whether to send a '1' or a '0'. The outgoing axon is denoted in the perceptron drawing by ψ, which takes on the values 1 or 0, depending, respectively, on whether the threshold proposition, Eq. 1.2, is true or false. This corresponds to our σ_i'. What the perceptron drawing is intended to convey is that each of the little boxes, ϕ_i, is communicating some simple predicate about the perceived world. A whole variety of these predicates about

some scene are evaluated in parallel and the weighted linear sum of the outcomes of the partial predicates is compared with a threshold, in the expectation that a proper choice of weights and threshold will endow the ψ-outcome with interesting classifications of the complex perceived scene.

It is worthwhile quoting a couple of passages from the delightful book of Minsky and Papert, especially as it has been out of print for so long. They comment, somewhat facetiously, (Section 0.9) that the original attraction of the perceptron stemmed from the fact that:

> The machine is built with a fixed set of computing elements for the partial functions ϕ, usually obtained by a random process. To make it recognize a particular pattern...one merely has to set the coefficients α_ϕ [our J's] to suitable values. Thus 'programming' takes on a pleasantly homogeneous form...[one can] imagine a kind of automatic programming which people have been tempted to call *learning* by attaching feedback devices to the parameter controls...which will cause the coefficients to change in the right direction when the machine makes an inappropriate decision. The *perceptron convergence theorems* define conditions under which this procedure is guaranteed to find...a correct set of values.
>
> ...The perceptron was conceived as a parallel-operation device in the physical sense that the partial predicates are computed simultaneously...they are computed independently. [all italics in the original]

The typical computation that a perceptron can perform is a classification. There are 2^n different possible combinations on input, with each of the n ϕ's taking on the two values 0 and 1, i.e., each input is an n-bit word. There are only two possible outputs, ψ can be either 0 or 1. Hence, depending on the values of the weights, α_ϕ, the output bit will represent a classification of all the possible inputs into two classes – in one will be all those which give rise to a 0 on output and in the other will be the inputs leading to a 1.

We will not engage here in any detailed description of what has been learnt about perceptrons. It is impossible to compete with the style or the clarity of the book *Perceptrons*. For our purposes it is worth noting that while bringing about an elucidation of the properties of

perceptrons the book's message has been that perceptrons are not very useful, either as computing elements or as description of psychophysical data. This conclusion is reached by Minsky and Papert based on theorems exhibiting the restricted nature of the type of classifications that can be affected by **linear** threshold calculations, as well as on theorems that show that some of the classifications that can be *learned* require an inordinate amount of time. It must be said, in full honesty, that when computational tasks are stated as ambitiously and universally as they were in the context of the perceptron, the bounds that have been established by Minsky and Papert have not been traversed by any later device. The perceptron seems to capture something very basic and generic and its limitations seem to have a life of their own.

1.3.3 Beyond the basic perceptron

Yet, the perceptron cannot be written off. As was observed quite frankly by Minsky and Papert, all that their theoretical arguments could demonstrate amounted to showing that the perceptron in *its simplest form and most restricted interpretation* was a paltry device. As soon as one expands the scope, many possibilities open up, though they are theoretically much less transparent. See e.g., ref. [26]. One could allow for more complex predicates to be computed by the 'sensory' elements, ϕ_i, or for non-linear predicates to be computed by Ω, or as with the formal neurons of McCulloch and Pitts, to have computations of more than one layer, e.g., Figure 1.8.

Another possibility consists of a combined variation: The consecutive layers are taken to be the original predicate inputs and the computational tasks are reformulated. It is this option that will be followed here. What it implies is that the computation is not read at the output of a single perceptron after a single time-cycle. Instead, the output is fed as input to similar units. But first one engages in a more innocent modification of the perceptron. The N units at the input are connected to many linear threshold processors of the Ω type, with connection weights (the J or the α_ϕ) which can vary from processor to processor, as is depicted in Figure 1.10. See also ref. [25]. What such a setup can do is classify the elements of the space of 2^N patterns of possible combinations of elementary input predicates by projecting them into the space of 2^Q combinations of Q output spikes. In other words, the 2^N possible inputs are divided into classes, each of which contains

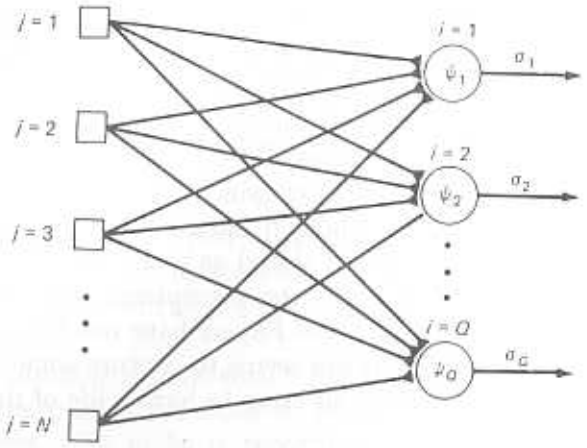


Figure 1.10: A single layer multi-perceptron.

all the inputs which produce the same combination of 0's and 1's on output. The choice of all the channel weights, the $N \times Q$ independent numbers, determines the properties of a particular projection.

The same statements can also be phrased in terms of pattern recognition. The 'seductive aspects of the multi-perceptrons' are that they allow for a reinterpretation in which the N bits, previously the simple predicates, are now viewed as pixels in an image, i.e., the black and white dots in a discretized visual pattern. The projection affected by the Q linear threshold elements is a classification, in the sense that whole groups of input images are mapped on single output combinations of Q bits. Each of the 2^Q outputs may be considered as a familiar prototype, of some character in the alphabet, for example. Thus, one may imagine character recognition of varied hand-writings. If the weights can be so chosen as to provide controlled classifications, i.e., in which the groups of input patterns that produce a single output word can be prescribed, one can read the action of the machine in either of two ways:

- as *correcting errors* in perceived patterns relative to prototype concept images, or
- as *associative* recall of concept images by varied input patterns.

This has been an active trend of research with interesting applications.

See e.g., refs. [27,28,25,29,30,31]. It has even made inroads into the interpretation of empirical neurophysiological material[32].

Such an interpretation is quite in the original spirit of the perceptron, because the value in each pixel is the truth value of the simple proposition asserting that the corresponding discrete square is, or is not white. Consequently, this approach shares the attractive features, as well as the limitations, of perceptrons. It can classify and it can 'learn', by the perceptron's mechanisms, as will be shown in Section 9.2. The main advantage of the multi-peceptron, as was pointed out above, is that while the perceptron can learn to classify patterns into two classes at most (whenever such a classification, in terms of linear threshold functions, exists), the multi-perceptron can learn to classify into a multiplicity of classes.

1.3.4 Building blocks for attractor neural networks (ANN)

A significant evolutionary leap, in the world of models, is accomplished when the multi-perceptron is closed onto itself[33,34,35,36,37]. In other words, when the output axons from the linear threshold elements – the σ_i' in Figure 1.6 – are identified with the σ_i. The neurons now form a feedback mechanism, closed on itself. Clearly, by definition, in terms of the description of the multi-perceptron, we must have $Q = N$. This is **not** an input-output system. The feeding, of external inputs, and the reading, of its output have to be dealt with separately, as will be done below. It is a part of the redefinition of the computational function alluded to at the beginning of Section 1.3.3. When the output axons become input channels there is, of course, a time shift. In other words, if at time t one has a set of N zeroes and ones, denoted by $\sigma_i(t)$ then the set of N bits composing σ_i''s becomes the set of inputs a neural cycle-time (1–2 milliseconds) later, i.e., $\sigma_i(t+1)$. A graphical transition from the multi-perceptron of Figure 1.10 to our tail-biting ANN is affected in Figure 1.11. The truth values ψ_i become input predicates ϕ_i, at the next time step.

The attractiveness of the perceptron, as well as of the multi-perceptron, is due in large part to its amenability to analysis, which allows a far reaching clarification of their potential properties. The ANN is no longer a linear threshold projector, it is a *dynamical system*, with rather complicated feed-back. For any given initial set of signals present in the channels, it goes on, cycle after cycle, to wander

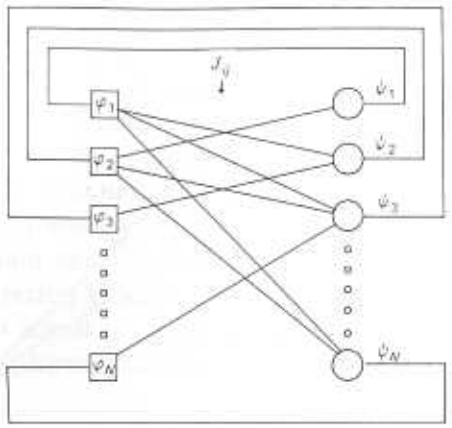


Figure 1.11: A multi-perceptron closed on itself to form an ANN.

among the 2^N possible combinations of N such signals. Supposing that the N^2 channel weights (synaptic efficacies) as well as the N thresholds are fixed, respectively, as J_{ij} and T_i, then the dynamical process can be described by a simple reinterpretation of Eq. 1.2. Formally it reads:

$$\sigma_i(t+1) = \psi[h_i(t+1) - T_i], \tag{1.3}$$

where the function $\psi[x]$ is 1 or 0, depending on whether its argument is positive or negative, respectively, and

$$h_i(t+1) = \sum_{j=1}^{N} J_{ij}\sigma_j(t). \tag{1.4}$$

A tiny example of such a dynamical process is shown in Figure 1.12. There are five neurons. Each is labeled by the state of its axon – if it conducts a spike the neuron is black, if it does not it is white. The consecutive states of the five neurons, at intervals 1, are depicted from top to bottom and the times are marked to the left of each stage. In this representation the connections are not drawn, only the logical states of the neurons. The matrix of 25 synaptic efficacies appears to the right of the temporal configurations. This matrix, together with the thresholds of the neurons, $T_i = -0.3$, determine the sequence of states given the initial list of spiking neurons – the initial *network state*.

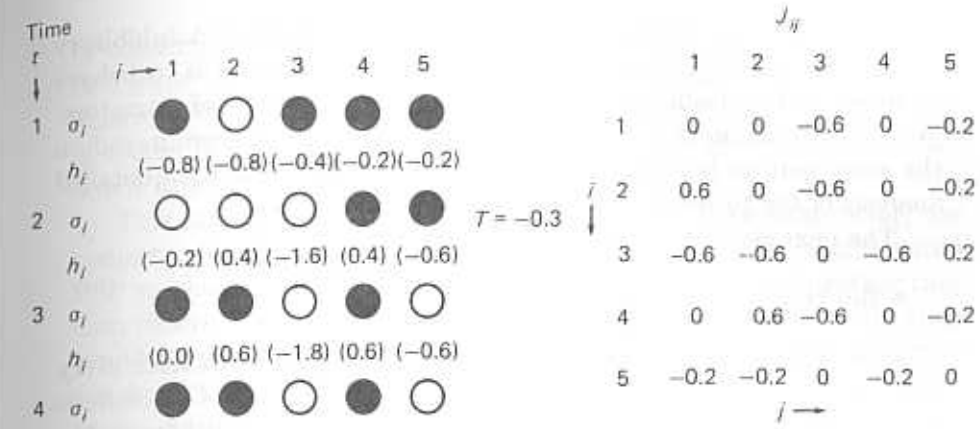


Figure 1.12: Dynamics of a five neuron network, with thresholds $T_i = -0.3$ and a connection matrix J_{ij}. White neurons are resting, black ones have spiking axons. Note that following $t = 3$ all states will be identical. This is an example of a *fixed-point attractor*.

This simple example gives a hint of our central concept – the *attractor*. Once a given configuration of firing neurons has repeated itself once, it will keep repeating indefinitely, because the very same PSP's will be generated. We can, therefore, say that the dynamics of the network has led from the initial configuration, at $t = 1$, to the *attractor* network state which appears first at $t = 3$ and at which the network will remain. Had the network been started in the state at which it is at $t = 2$ in the example, it would have been attracted to the same point in the space of its states. Chapter 2 will be devoted to an elaboration of this concept.

Already at this stage a rather strong assumption has been introduced, from the neurobiological perspective

- the individual neurons have no memory.

This is implied both by the fact that the thresholds do not vary with time, as well as by the fact that the PSP's, $h_i(t+1)$ depend *only* on the activity of the axons one cycle-time earlier. A step beyond this simplified picture is discussed in Section 7.4.

An additional simplifying assumption will be

- the restriction to a single type of neuron, which may have as many excitatory as inhibitory synapses emanating from its axon.

(See e.g., Section 1.2.3.) To have both excitatory and inhibitory synapses in the system is absolutely vital, if the network is to behave in an interesting fashion, i.e., that it can store a diversity of attractors. We will expand on this point in Section 2.3.6. To have them mixed on the same neuron is a matter of convenience which permits a detailed analysis of the properties of the network.

The next simplifying assumption will be

- full connectivity.

That is, every one of the N neurons can receive an input from every other neuron in the network and can send an output to it. Our neurons do not connect directly to themselves. While full connectivity is not very realistic biologically, it is not entirely off the mark. After all we are choosing our networks to have as many neurons as the mean connectivity in the cortex. Hence, this type of elementary module is very intensively connected. (See the discussion in Section 1.2.1). Nevertheless the extreme uniformity implicit in the assumption of full connectivity is one of the artificial features which will be lifted in due course.

Two additional simplifications allowed the initial breakthrough of Hopfield[37] . They are

- symmetric connectivity

which implies that the efficacy of a synapse communicating output from neuron j to neuron i equals the efficacy of the synapse communicating the output of neuron i as input to neuron j. This strong assumption has no biological justification. Yet, it has made the model completely tractable *and*, as will be shown in Chapter 7, the main properties of the network are not very severely affected when this assumption is lifted. In the notation of Section 1.3.2 it reads

$$J_{ij} = J_{ji}. \tag{1.5}$$

- the dynamics of the network is *asynchronous*

As the network starts to turn, one must prescribe the order in which neurons modify their activity states, each proceeding according to Eq. 1.3. It may seem natural to suppose that they are all doing it

in concert, as if guided by a clock, as was the case in the example in Figure 1.12. That would imply that every one of the N neurons makes its decision on the basis of the same activity configuration, list of spikes, of all the neurons in the network, until they have all chosen a new firing state.

This attitude is neither biologically warranted nor analytically expedient. Neurons make their decisions whenever their accumulated PSP reaches the threshold. They receive their inputs with rather random delays. Thus, even if they started in a synchronized state, they would very rapidly desynchronize. Moreover, the random delays in communication, due for example to different axonal or dendritic distances, causes neurons to respond on the basis of the firing states of their fellow neurons, some of whom have already made their new decision and some which have not. This type of disorder is difficult to model in a detailed fashion, since details may depend on the precise particular realization of the specific network. The next best choice is to avoid fragile assumptions of synchrony by assuming instead that the neurons are picked in a random order for updating, and each neuron in turn sees a different mixture of inputs, based on the sequence in which the selection has proceeded. In other words, each neuron computes its PSP, based on the particular mixture of inputs that it finds around when its turn comes up in the random sequence. This important point will be discussed in further detail in Section 2.2.

1.4 The Network and the World

1.4.1 Neural states, network states and state space

In Section 1.3.1 we have chosen to describe the dynamics of the single neuron in terms of a discrete variable which can take on two values +1 and 0, corresponding to the neuron's state of activity. These two values, or a richer description in terms of a range of continuous values (see e.g., Section 2.1.1), provide a listing of the possible *neural states*. But it is a basic tenet of the approach presented here that

- all significant cognitive events take place on the level of the network.

Here we still keep in line with

> The fundamental premise of connectionism...that individual neurons *do not transmit large amounts of symbolic information*. Instead they compute by being *appropriately connected* to large numbers of similar units[38]. [all italics are in the original]

Given that information is communicated via axonal activity, it is reasonable to classify *network states* by lists of the simultaneous axonal activities. If *neuronal activities* are described by discrete, two-valued variables, then at any moment in time the network is characterized by N bits, each bit corresponds to a specific neuron and its value indicates whether the neuron is firing or resting. There are, of course, 2^N such *network states*. At each moment the network is, by definition, in one of these states. Computer scientists and mathematicians like to view such states as forming the vertices of an N-dimensional hypercube. The dynamics of the network makes it hop and skip from one vertex to another.

Let us recall that our basic network typically has 10^3–10^4 neurons – the mean cortical connectivity, much like in Hinton and Anderson[39]. The number of possible *network states* in the elementary module is enormous. The view to be advocated here is that

- elementary cognitive phenomena (such as retrieval from memory and recognition, for example) are represented by *patterns of activity* of networks of this size.

It is perhaps these large numbers that led Feldman and Ballard[38] to try networks of ten neurons, because they 'have not seen how to carry out a program of specific modeling in terms of these diffuse models'. In my view such diffuseness is the heart of the matter. The program represented by this text is a proposal of 'specific modeling' on that scale, though the computational tasks to be performed by the networks are incomparably more modest.

In this rich space of network states we will need a measure for a *distance* between states, each of which is an N-bit word. A distance will allow us to judge quantitativly how different are two visual pictures described by pixels; how far is an attractor from the initial state of the network (as in Figure 1.12), etc. Eventually, the distance will give a

quantitative measure of how faithfully a certain memory has been recalled, or how different stimuli can be and yet be associated with the same memory. It will allow the estimation of the extent of that part of the space of possible states that the network will cover under given dynamical and initial conditions etc. Since our states are binary N-bit words, a natural distance between two states is the Hamming distance, which is simply the number of bits at which the two binary words differ. An alternative measure, which physicists would find more natural, is *similarity* or *overlap*. This is the kind of quantity which naturally maps on such familiar theoretical constructs as magnetization (see e.g., Section 3.2.4). It makes physicists feel at home and enables them to pull out the tricks of their trade. While the Hamming distance measures difference, the overlap measures similarity. They are completely equivalent and are linearly related, as we proceed to show.

1.4.2 Digression on the relation between measures

To make this relationship more formal, we shall denote the Hamming distance between two given network states, by d_H. Let us also denote by N_G the number of bits which are identical in the two words. Clearly then,

$$N_G = N - d_H,$$

and the difference between the number of agreeing bit and the number of differing bits, M, could be written as:

$$M = N - 2d_H \qquad \text{or} \qquad d_H = \frac{1}{2}N(1 - \frac{M}{N}) \tag{1.6}$$

where N is the total number of bits in each word. The quantity M/N will be called the *overlap*.

Note that if the 0's in the words were replaced by -1's, then the *overlap*, M, between two words would be

$$M = \sum_{i=1}^{N} S_i \bar{S}_i, \tag{1.7}$$

where S_i and $\bar{S}_i$ are the N ± 1 bits of the two words. The above expression for M is explained by the fact that every position at which the two words agree contributes a 1 to the sum, giving N_G. While

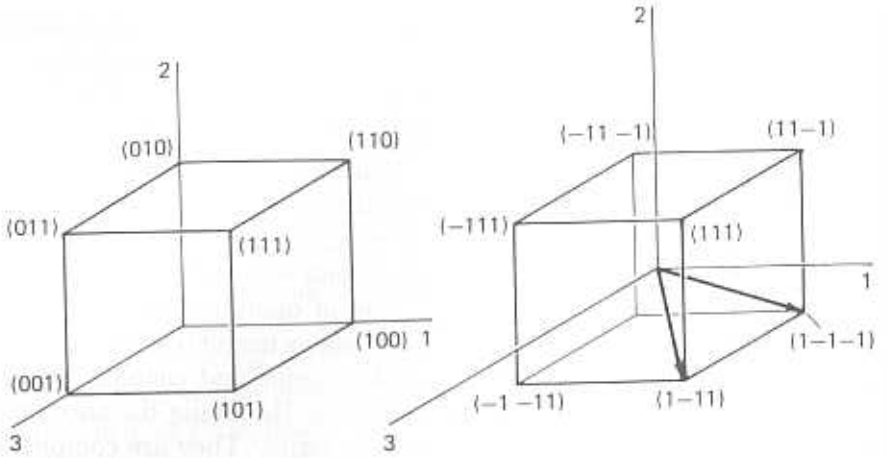


Figure 1.13: Geometrical representation of the distance between pairs of network states.

every position at which they disagree contributes a -1, subtracting d_H from N_G.

Eq. 1.7 has a very appealing interpretation: it is the scalar product of the two N-dimensional vectors whose components are the ± 1 of the two words. Consequently, M is just N times the cosine of the angle between these two vectors, which will be denoted by $m(S, \bar{S})$. This geometrical picture in terms of angles must be supplemented by a geometrical interpretation of the vectors S. To do this we observe that if the components of the initial 0-1 words had been σ_i, then

$$S_i = 2\sigma_i - 1, \tag{1.8}$$

which has a natural interpretation: if the 0-1 words were the vertices of an N dimensional hypercube, of side 1, then the new ± 1 words are also vertices of an N-dimensional cube. But this cube has its center at the origin of coordinates, has a side of length 2, and has its vertices symmetrically disposed relative to the origin of coordinates. This transformation is exemplified in Figure 1.13, for a system of three neurons, which has eight possible states and its states are the vertices of a cube in three dimensions. The quantity m, in the S-description, is nothing but the angle cosine between two vectors from the origin to the two vertices corresponding to the two states.

In terms of m, Eq. 1.6 is written as[26].

$$d_H = \frac{1}{2}N(1 - m) \tag{1.9}$$

It is easily verified that m varies from 1 for identical states, through 0 for states which differ by one half of their *neuronal states*, to -1 for maximally different states. It is rather rewarding that the transformation Eq. 1.8, which converts the Hamming distance into a geometrically natural object, is also the one that would be chosen by physicists. It will be discussed again when the model is set up in detail, in Section 2.1.2.

1.4.3 Representations on network states

Not all network states will be given equal significance. Only special dynamical sequences will be selected as candidates for the description of cognitive events, such as retrieval or recognition. A dynamical sequence of network states consists of consecutive listings of all neuronal activities. Each listing is a snapshot of the activity states of all neurons. Consecutive listings follow each other by a single cycle-time in which all neurons had a chance to update their neuronal states. Note the double temporal aspect of the concept of the dynamical sequence of network states. Each snapshot, which in network terms is static, describes a dynamical state of each neuron. We will focus attention on special sequences of states that will provide us with articulation for

- emergence
- generation of meaning
- self-recognizability, i.e., freedom from homunculus.

In the simplest realization of this program, the cognitive events are identified as *fixed points*. A dynamical sequence of network states is a trajectory on the hypercube in the N-dimensional space in which the N-bit words, representing the network states, are the vertices. A *fixed point* is a sequence which remains at the same point **in the space of network states**. It would be the sequence of states which follows $t = 3$ in the example of Figure 1.12. It does not imply that dynamics has come to a standstill. Neurons go on firing, but because the existing synaptic efficacies reproduce the same list of active and inactive neurons

at every cycle, one finds the same network state in every cycle. Here the double dynamical aspect mentioned above manifests itself in the fact that a repeating activity of the neurons is a stationary state for the network. Actually, this would be a situation in which some of the neurons are maximally active, because each neuron which is **on** in such a fixed point, is firing bursts at the top rate, the others are thunderously silent. This extreme picture will be significantly relaxed as we go along.

The special emphasis one places on fixed points is because they are used to represent our elementary cognitive events. Given a certain initial network state, which might be imposed by an external stimulus, the network follows a dynamical trajectory, determined by its synapses.

- The rapid arrival at the fixed point will be identified as a recall from memory of the pattern corresponding to the state which is fixed. This state is recalled by its similarity to the external stimulus.

To recapitulate, a pattern is recalled if under the influence of a stimulus the ANN drifts rapidly into an *attractor* such as a *fixed point*. The fixed point is the simplest kind of attractor. It attracts in that the dynamics causes the trajectories from many initial points to flow into it. But attractors could be cycles of states, or even non periodic sequences, as those discussed in Chapter 5. The network state, the N-bit word, that is repeating at the fixed point is the memory *associatively recalled* by the incoming visual stimulus. The **realization of the recall** is a **pattern of activity** which corresponds to a **fixed** network state. The relation between the possible fixed points of the dynamics of the ANN and visual patterns in the world is unspecified, and may be partly innate and partly learnt. But, in some sense, these special patterns of activity – the fixed points, or other types of attractors – are the nearest we come to having *representations* in this attempt at a description of cognitive processes.

The stimulus may be some visual pattern. The input mechanism does its thing to it, and then it is the task of the ANN to decide whether the pre-processed input (into the ANN) is *associated* with a pattern in memory. If it is, then either some biological, external reaction should ensue, or another ANN should be provoked into further processing of the distilled pattern. The latter can take place if the first ANN sends efferents (outgoing lines) for further processing.

Still on this preliminary level the network may be *retrieving, recalling* or *recognizing* some external stimulus. As such it must be connected to an input mechanism, which itself may be a computing neuronal system, directly below the surface. In other words, the ANN is **logically** located close to the interface of the central processing area with the *afferent* (incoming) input signals, which is the leftmost part of the central box in Figure 1.1. The pre-processed input (see Section 1.4.4) projects onto an ANN (out of many possible networks it may affect), by injecting PSP's into the neurons of the ANN (or into a subset of these neurons), according to the activity provoked in the input mechanism by the external stimulus.

> An important point must be clarified here. Many authors of the multi-perceptron culture have interpreted the Hopfield ANN as having all neurons in the network as input elements and as output elements. This interpretation is unwarranted. The approach that will be followed here is that the input and read-out have to be imposed externally, with much freedom. Such an approach appears natural when hardware realizations of such networks are contemplated. Actually, in terms of popular tokens **all** the neurons in an ANN are like the hidden units of the PDP approach[40]. In PDP language we have strongly interacting hidden units. Note also that the interface between the input and the ANN may be well defined by the fact that **there is no feedback from the ANN to the input neurons**. This, of course, is not a biological assertion. Rather, it is a logical possibility worth modeling. It is presented pictorially in Figure 1.14, which is a blowup of the central part in Figure 1.1 .

- The boundary is drawn where a cut goes through one way neuronal transmission, into an area with intensive feedback.

Representation is a very charged word in the context of cognitive studies. (See e.g., [41,2,13,42]). We take up this issue again in the next section. What the above discussion should clarify is that in an ANN there are states, which are N-bit words, and which are evoked by *classes* of stimuli. Yet these words, which in some sense *represent* those classes, are ephemeral, repeating activity patterns of the network. As

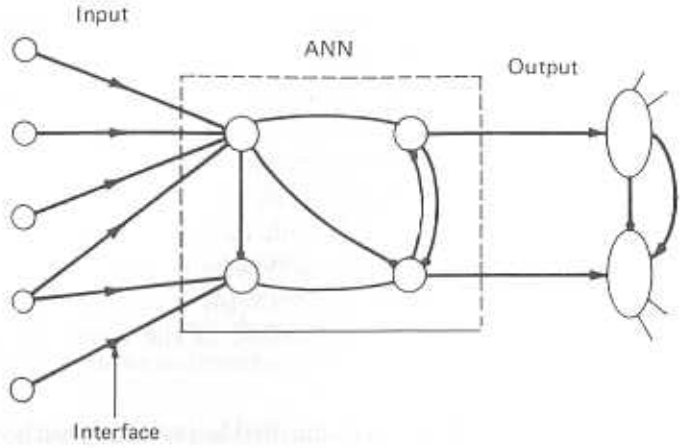


Figure 1.14: Logical interface between input and ANN. Defined by the transition from an essentially feed-forward connectivity to a network with extensive feedback.

we shall see abundantly in what follows the potential fixed points are determined by the masters of the dynamics, which are the synaptic efficacies.

1.4.4 Thinking about output mechanism

The return of the grandmother cell

In the previous section, strong emphasis was laid on repeating patterns of **neuronal** activity (fixed network states) as candidates for *self-recognizability* for *emergence* and for *generation of meaning*. To fill these pledges with content we strongly depend on the output or read-out mechanism. There must be a way for the system to realize that the ANN has reached a fixed point, or any other type of special temporal sequence of network states, which is, after all, a very special temporal pattern of *neuronal* activity and inactivity. The simplest way would be for the system to have *read-out neurons*. The ANN should connect to them efferently i.e., it should synapse on their dendrites. One can conceive of two extreme types of read-out neurons.

- Type 1 – Neurons which transmit the output of one ANN to another one, or to some motor system, by establishing a one-to-one communication between neurons in the first ANN and those in the second. See e.g., Section 8.4

- Type 2 – Neurons which identify a specific network state by connecting efferently (via their dendrites) to a *large* sample of the neurons in the ANN and communicate the outcome for further processing, or to a motor system.

In the words of a biologist:

> A person or neuron reading the output of these neurons would know that the stimulus in the receptive field was important. He or it would not know how the organism would respond in response to the stimulus: whether to make a movement, or whether just to notice and not move. The activity before visually guided saccades...is an excellent example of such a response to a generally attended stimulus[43].

Neurons of type 1 will be very useful in mediating between ANN's which process different levels of hierarchically stored data structures, as in e.g., Section 8.4. Or, for communicating identified stimuli of a special kind for higher level processing, as would be the case for conditioned counting of identical stimuli, such as chimes. See e.g., Section 5.4. It is unlikely that neurons of this type can provoke spikes in the neurons of the second network. Many more than a single excitatory input would be required. They could though contribute a coherent PSP to the second network, an effect that is amplified if the read-out neurons are inhibitory[44].

Output neurons of both types must be able to temporally average the activity of **each** of a significant sample of the neurons of the ANN over a period of a number of cycle-times of the ANN. Output neurons of type 2 must also have an excitatory synapse to a sample of neurons which are firing in the pattern, and inhibitory ones to a sample of those which are quiescent. There is, of course, a question whether such averaging neurons exist. This is an empirical question. But in the absence of an unambiguous answer, it should be pointed out, before they are ruled out, that these are the very same neurons which respond selectively to different afferent spike rates. After all attractors manifest themselves in a significant rise of spike rates.

The first requirement, of temporal averaging, is to ensure that what will be read out will be those temporal sequences of network states in

which the sampled ANN neurons are essentially repeating their activity in most cycles within the averaging period, i.e., fixed points. Those network states should generate a response in the read-out neuron. The average number of spikes per cycle, for the active neurons, will be very close to 1, while for the passive ones it will be very close to 0. Small fluctuations can be allowed if the read-out neuron is somewhat tolerant. Such tolerance will also allow some errors, in which a small fraction of the neurons will not conform with the memorized pattern. The level of tolerance is determined by the biological system, or alternatively by the way in which the artificial device is constructed.

The second requirement, of selective coupling to the output neuron, ensures that in a recalled pattern of memory the

- inactive neurons are as significant as the active ones.

Moreover, it ensures that the read-out neuron will register a response only for a small subset of the possible attractors of the ANN, or even for a single attractor only. Recall of different memories, requiring different responses, are to be recognized by different attractor selective read-out neurons. This would classify these neurons as *grandmother cells*. See e.g., ref. [45] and ref. [13][p. 113]. But before placing the present approach in this context, we should complete the discussion of our read-out neuron.

It may be expedient to have output neurons which read only a subset of the neurons in the ANN. In this way, one could account rather naturally for situations in which an N-bit network state may contain a mixture of data relevant for a certain task, such as a number, together with contextual information or specific attributes, such as shape or sound. This may permit, for example, the retrieval of a relevant memorized word associated with a number, by the visual form of the number, which may require only a fraction of the total number of bits in the memorized pattern. It may then be utilized for an arithmetical task, and may allow for a reaction based on its auditory properties, which may be coded in another fraction of its total number of bits. Every feature may draw on different parts of a generous number of bits allowed for the corresponding attractor state. Such an attitude is implied in the exercise of Section 5.4.

In a sense, these read-out cells are somewhat like perceptrons. They are considerably freer in that they should respond to small classes of inputs, the members of which are grouped together in Hamming

distance and are rather likely to be separable for perceptron classification. The extreme case in which every read-out neuron is to recognize a single attractor is clearly achievable. There remains the question as to the origin of the careful planning of the connection of the ANN to the read-out neurons. But there also always remains the answer that these connections are partly due to evolutionary programming and partly to learning in ontogeny. The latter can come about *instructively*, or *selectively*[46]. This careful connection to the output neurons is no more mysterious at the output than at the neuro-muscular junction end of the ANN. Perhaps this puzzle is not especially about vertebrates with highly evolved mental capabilities. Be that as it may, we will not elaborate on this subject in what follows.

Both types of read-out cells provide the system with the means necessary for ascertaining that the ANN has reached an attractor. The cells of the second type can tell which specific attractor it is. If the stimulus has led the ANN into an attractor rapidly, then the read-out neuron will fire and the biological system will be able to proceed in the special way, appropriate for the recognition of that particular stimulus. When the action takes place on the level of retrieval, this is the end result. But since both the attractor and the read-out mechanisms are proposed as universal mechanisms, this reasoning can be repeated on higher levels of processing, where ANN's may deal with various compositions of sensory inputs, together with inputs from other cortical components. Again, the point of view advocated here is that any cognitive operation must end up with a rapid drift toward an attractor, a fact which can be recognized by the corresponding read-out cells.

- The fact that resident cells can **recognize** special dynamical sequences and lead to biological function (such as motor response) is freedom from *homunculus*.

In other words, the neutral fact that the network enters a repeating pattern of neural activities is the signal that a cognitively significant event is taking place, without the observing eye of an all knowing little person. This is the meaning we give to self-recognizability.

But a dynamical process that leads to the appearance of such special temporal sequences of network states, **which are robust to significant amounts of noise**, i.e., to random errors that can occur in the communication between neurons, is very remarkable indeed. This is one of the main themes of our approach and of this volume. It is one of the

most important and specific contributions that physics has introduced into the subject. In fact, it is so remarkable that it is quite appropriate to refer to such attractors as *emergent* properties of the network[37]. The precise sense of this emergence will be the subject of Section 3.2, and it is just because the sense is so well defined that I insist on using it, despite its problematic status in biological vocabularies.

Finally, we also had *generation of meaning* on the label of our package (in Section 1.4.3). This concept will also be given a precise meaning, and yet it will retain an interesting edge. Input into an ANN may, or may not, lead within a biologically reasonable short time, to one of the the network's attractors. If it does, it can lead to a biological function which endows it with *meaning*[47,48]. Accordingly, we classify stimuli entering the ANN, as *cognitively meaningful* if they lead the network quickly to an attractor. Otherwise, the input is classified as meaningless and ignored. This is the coarse division between the meaningful and the meaningless. It refers to inputs into the ANN. The specific meaning is in the particular attractor, and it depends, of course, on the level of cognitive processing in which the ANN is situated, i.e., it may be a mere retrieval, a recognition or the result of an arithmetical computation. If the unfamiliar stimulus is imposed persistently enough, then it may become a candidate for learning, which is the process of formation of new attractors.

Digression on grandmother cells

It was mentioned above that the read-out neurons are in some sense *grandmother cells*. By this one usually refers to a neuron that can discriminate between complex stimuli[45]. By construction, the read-out neurons of the second type, will respond, i.e., fire an action potential, to narrow classes of recognized stimuli, or features of stimuli, or, in general, to a narrow class of cognitively meaningful input. The abstractness and complexity of the response-provoking proposition depends only on which ANN is being read.

The 'classical' neuro-physiological concept of a 'grandmother cell' is summarized by Churchland[13] as follows:

> ...the idea is that successive levels involve feature extraction of an increasingly abstract type... from cells that respond to brightness contrast, to cells that respond to light boundaries in specific orientations in specific parts of the visual field,.... Higher yet are cells responding preferentially to a hand, or to a nonspecific face,.... Finally, the logic leads us to conjecture... cells whose conditions of response are even more abstract and complex, such as a cell that would respond uniquely to the presence of Grandmother....[p. 113]

This view has become largely discredited in biological circles, and is presented as such by Churchland. Yet, it is hard to deny that such neurons exist. Even on the simplest, sensory level rods and cones have differing spectral sensitivity curves and single somato-sensory neurons on the skin can make approximate determinations of position, see e.g., ref. [49]. These examples may appear rather lowbrow as will examples that point out that at the motor end one would also tend to find neurons with specific responses, which are specific to the response that should activate the particular muscle. As one penetrates deeper into the nervous system one notices that the selectivity of ganglion cells to center-surround contrasts is already a feature of reading *collective* outputs of *interacting* cells, such as rods, cones, amacrine cells and horizontal cells in the visual pathway, see e.g., ref. [49][p. 42].

The extreme view of the hierarchically specialized neurons can be traced to interpretations of the experiments of Hubel and Wiesel[50], on the visual cortex of mammals, although it does not seem to have been explicitly proposed in their classic work. Hubel and Weisel reported finding *simple* cells and *complex* cells. The first seemed highly selective and the last seemed to participate in a variety of features. More recent experiments[51] find that incoming information finds a multiplicity of passages into the cortex, which would challenge a picture of a single information-reducing pathway. However, along every one of the parallel pathways there must be a progressive reduction and abstraction of information. Along every one of them one would expect to find simple and complex cells. In our context, we would say that if we probe cells inside an ANN, or read-out cells of type 1, we would observe *complex* behavior, since they will participate in a number of ANN attractors. On the other hand, probing a read-out neuron of type 2, one would observe a *simple* cell. It should perhaps be remarked that the above discussion is not intended as an explanation of Hubel and Wiesel's experiments. Rather, it is an indication that when the conceptual framework developed here runs into controversial issues of neurobiology it has, at this stage, enough options to remain afloat.

1.5 Spontaneous Computation vs. Cognitive Processing

1.5.1 Input systems, transducers, transformers

A wide variety of models has been proposed to account for various perceptual functions, such as pattern recognition, classification, shape and motion perception. The main categorization of these models, which will be useful in our context, is between *feed forward* networks and ANN's. In the former we would group such proposals as Widrow's multi-perceptron adaptive classifier[25], Willshaw's correlation model[27], Kohonen and Palm's error reducing projections[52,53, 31] and much of the PDP feed-forward approach[40]. What is common to all these models, from our vantage point, despite wide differences in sophistication and in biological plausibility, is that all of them can be viewed as what Fodor[54] calls *input systems*.

The most important characteristic of such systems is that they produce for each input an output with the same status. The only way to distinguish between good and bad outputs is by their content, and if the little internal observer, the homunculus, is not to be introduced, one must make strong assumptions about the statistical distribution of the possible inputs, e.g., from the visual system. If such assumptions were warranted they would require 'input systems [that] are associated with fixed neural architecture'[54]. Yet, even if there are correlations in the universe of stimuli, one would like the system to be able to distinguish between familiar stimuli and those that have to be submitted for learning. Otherwise the use of the word *memory* is rather unnatural. Hence, already on the input level, cognitive discrimination seems to be required.

An alternative approach has focused on dynamical attractor models such as Caianiello[35], Amari[34], Grossberg[55], Little[36], and Hopfield[37]. A meaningful role for such models is not to replace the *input systems* but rather to complement them. Input systems may provide a lion's share of the elaborate *transduction* of the external stimuli, to use Fodor's terminology. They should be envisaged as part of the input scheme in Figure 1.1, and they 'function to interpret transduced information and to make it available to central processing'[54]. ANN's do not produce an output for every input. There is no direct computational relation between input and output. Instead, the network produces a signal which indicates that output has been produced — it reaches an attractor — and then the output can be read.

1.5.2 ANN's as computing elements — a position

In the context of computer science or artificial intelligence, every input-output relationship is viewed as a computation. Intentions and temporal relations do not figure prominently. Thus, for example, it is sometimes expected that a network should allow for the computation of a binary operation — Boolean or arithmetic — for all pairs of inputs that can be represented on the network[56], essentially simultaneously. As an example, consider the construction in Figure 1.15, of a network with 16 input bits and 8 output bits. The 16 input bits, in the column of boxes on the left, are permanently identified as two 8-bit operands – one operand in the top 8 and the other in the bottom 8. The internal elements, hidden units, are logical gates with two inputs, which can communicate their output to any other gate or to an output unit. The connectivity and the assignment of logical functions to the internal elements determine the mapping of the 16-bit input on the 8-bit output. An appropriate choice of these two sets of variables creates a multi-layered feed-forward (no feedback) network. For every 16 bits that are imposed on the input units, the 8 output bits, 8 full boxes terminating on the right, take on the state which corresponds to the sum of any two externally prescribed 8-bit operands.

We refer to this type of operation as *spontaneous computation*, in the sense that the network does its 'thing', and it is an external observer (homunculus) that has to interpret the relation of the input and the output as representing some specific binary computation. This is, of course, a quite satisfactory situation as far as the construction of devices is concerned. Yet it is not a reasonable way of accounting for a mental calculation.

The device in Figure 1.15 is not an ANN. It is described here because often ANN's are expected to perform as devices of this type, and it is easier to identify these expectations in simple systems. It is sometimes expected that ANN *attractors* represent specific computational outcomes of certain readings of the input stimuli. This is neither possible nor desirable. It is impossible because ANN's do not produce a different result for different inputs. Quite the contrary, they group stimuli in association classes represented by an attractor. All

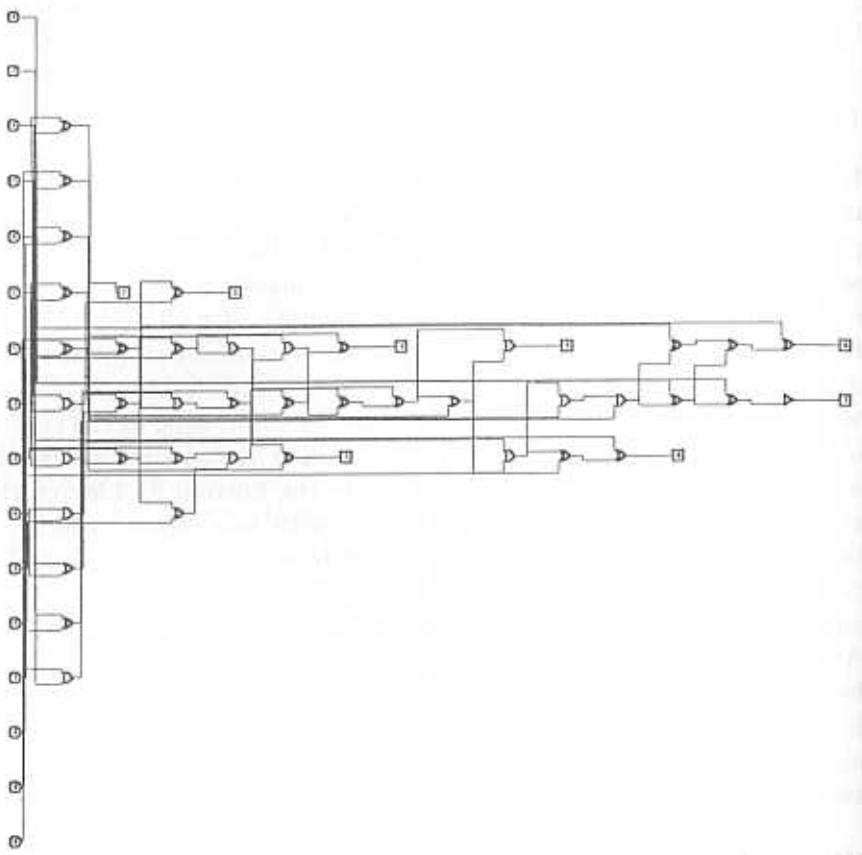


Figure 1.15: A network which *spontaneously* computes the sum of two 8-bit words, the top and bottom input units (open circles on left) and represents the sum on the 8 output units (full squares on right). The internal elements are logical gates with two inputs and several outputs. The full specification of the network includes a description of the connectivity as well as the choice of logical function at each gate. (From ref. [56], by permission.)

the stimuli in a class are *associated* with the attractor to which they flow. Moreover, it seems that mental computations, of the logical or of the arithmetical kind, are first and foremost operations on temporal sequences of data. A minimal superficial list of steps would include:

- *Operands and operation* reach the system through one sensory modality or another in some temporal order.

- Based on the operation, which must be recognized, the operands will be processed in one of a number of possible manners — it may be a retrieval from memorized results, or it may climb a level and an algorithm may be applied, using the attractors of yet another ANN.
- The imprint of the initial input sequence must be carried along, on some parallel channel, to provide the arriving outcome with specific meaning and a correspondence to the task assigned[57,58].

This list of steps is definitely not a *spontaneous computation*, and it is toward tasks of this nature that ANN's aspire to contribute. Some steps in this direction are attempted in Section 5.1.

ANN's have been criticized for their inability to perform some specific Boolean functions, or for being unable to store and retrieve specific sets of patterns, under given restrictions[59]. Specifically, Sejnowski et al have pointed out that a three neuron network cannot possibly store and retrieve the binary words $(1,1,1),(1,0,0),(0,1,0)$ and $(0,0,1)$, without storing additional patterns. As we shall see in Chapter 6 the restrictions for real storage are even more stringent than this example indicates. Yet it is not **this** limitation which will diminish the likelihood that such networks may account for some quasi-cognitive processes, within the perspective outlined here. It seems that criticisms of this type are remnants of the perceptron trauma, where it was at first expected that the scheme would have universal power for global feature extraction. Because of the high expectations, counter examples were devastating. The attitude adopted here is that such structured computations are neither elementary operations of an ANN nor of a human brain[60]. This attitude has found some empirical support in recent psychological tests[61].

This is not to say that relaxational processes, as represented by attractor dynamics, cannot be used profitably to perform *spontaneous computations*. The most notable examples are networks whose attractors are solutions to such problems as the *matching problem*[62] and the *traveling salesman problem*[63]. The first problem involves two sets with equal numbers of points. The distance between every point in one set and every point in the other set is given. The task is to choose the set of N pairs for which the total sum of distances is minimal. The *traveling salesman problem* involves a set of N points with given distances between any two of them. The task is to find an ordering

of the points such that the total distance traveled in passing through all the points in that order will be minimal. Such a trajectory will be the preferred choice of a traveling salesman, whose task it is to visit every one of the sites represented by the points. The two problems are sketched in Figure 1.16. The statement that these optimization problems can be 'solved' by neural networks implies that a network of neuronal elements can be set up having the following properties:

1. The *neurons* are given identifications in terms of the variables of the problem, e.g., representing a potentially matched pair in the matching problem or a pair of (city, position) in a trajectory in the traveling salesman problem.

2. The *states* of the neurons can be given the interpretation of truth values of propositions represented by the neuron, i.e., an active neuron implies that the conjunction represented by its definition in the context of the problem is true. An inactive neuron implies that the conjunction is untrue.

3. A set of *synapses* can be defined in terms of the parameters of the problem – distances in the two examples – and the constraints of the problem such that the optimal solutions of the particular problem will be an attractor under the normal neural dynamics.

Both networks provide a dynamical, analog method to an approximate solution of problems of high computational complexity. The pair-matching network has found a rather interesting application in modeling the visual detection of motion. Since the visual system transmits discrete stimuli, there must be a way of connecting corresponding points on two consecutive images of a moving solid body. One way of modeling this mechanism of visual analysis is by optimization of *matching* pairs of points, which in turn can be performed by an appropriately connected neural network[64].

As far as our present approach is concerned, these types of attractor networks will be considered as part of a sophisticated input system, either in the scheme of Figure 1.1, or in Fodor's theoretical sense.

1.5.3 ANN's and computation of mental representations

The issue of *representations* has been hinted at already in Section 1.4.3. The controversy as to whether the postulation of *mental representations*

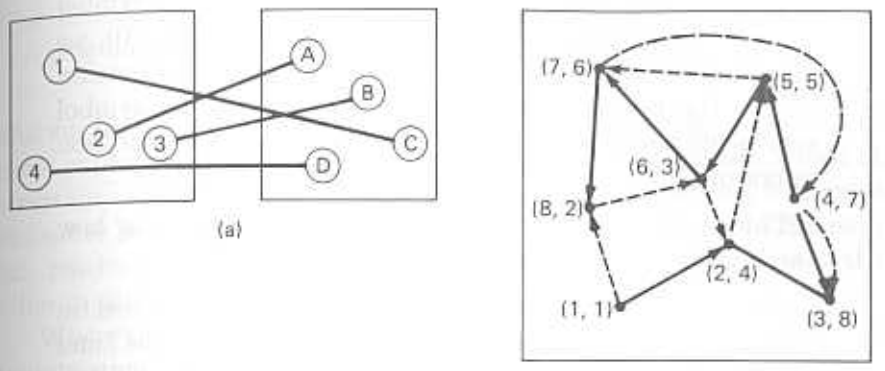


Figure 1.16: (A) One possible pairing in the matching problem. (B) Two trajectories in the traveling salesman problem. The pairs of numbers indicate the the order in which the corresponding points are visited in the two trajectories. The positions of the points in both problems are fixed.

and of *computations* in terms of these symbolic entities is a plausible description of cognitive activity, is raging on a variety of terrains — from the ontological to the epistemological to empirical psychology and psycho-linguistics. For us it may be relevant on two counts. First, it is strongly related to the anti-reductionist thesis. See e.g., ref. [13] [p. 298]. Second, it is a refreshing hope, even if only a remote possibility, that a concrete model may pave the way out of a rather abstract debate. Here we will not presume to advance this argument within the technical context in which the debate normally takes place. We will restrict ourselves to the hope that an articulation of ANN type models for cognitive activity may provide an additional dimension to it, as the PDP approach seems to have done[42,13].

Yet, perhaps more for the sake of modelers than cognitive scientists, we feel obliged not to avoid a presentation of the main lines of the controversy. The staunchest presentation of the representational-computational position is that of Pylyshyn[41]. This is at bottom a computer metaphor applied to the cognitive domain. His formulation is as follows:

> ...computation is the only worked-out view of *process* that is both compatible with a materialist view of how a process is realized and that attributes the behavior of the process to the operation of rules upon representations...

> ...in order to explain why the machine prints out the symbol '5' when it is provided with the expression '(PLUS 2 3)', we must refer to the meaning of the symbols in the expression and in the printout...it prints out '5' because that symbol represents the number five, 'PLUS' represents the addition operator...and five is indeed the sum of two and three...
>
> ...This is, of course, precisely what we do in describing how (and why) people do what they do...
>
> ...the abstract numbers and rules...are first expressed in terms of syntactic operations over symbolic expressions...and then these expressions are 'interpreted' by the built-in functional properties of the physical device...the syntactic structure of representations reflect all relevant semantic distinctions...

This is an extreme view, but in one modification or another it informs much of current research in cognitive science and in linguistic philosophy and psychology. For the opposite view we paraphrase an outspoken and persistent representative, Shanon[42]. He argues that essentially no variant of this position is plausible, or efficacious, either as a philosophical account or as a description of empirical psychological data. On the empirical side: the computational approach depends on a number of features which can be put to a test,

- Representations, being abstract, should be independent of their substrate — language, modality of presentation (visual, acoustic etc.)
- If behavior is governed by computation of representations, then rules of substitution should apply.
- In the linguistic domain, one would expect words to have single meanings, otherwise there will be an ambiguity as to which representstion is to be acted upon in the computation.
- Still in the linguistic domain, there is a difficulty as to which representations are expected to be activated by metaphors, or by irregular use of words, which seem to be essential ingredients for innovative use of language.

- If words evoke fixed, abstract representation, then performance in the linguistic domain should be independent of context.

Shanon brings a whole list of empirical data to challenge every one of these expected features[42] . There are, of course, stratagems which at a certain cost would deflect much of this evidence. The cost is a more elaborate account of cognitive behavior. This, according to Shanon, has reached proportions which are reminiscent of the defense of the ether in nineteen century physics.

Where do ANN models fall in this controversy? One can give very tentative answers at best. The models spring from the neuro-biological substrate, but their operation can be abstracted from it. It was mentioned in Section 1.4.3 that things in the world, having been learnt, have a corresponding representative. It is a network state which is an attractor or, in other words, the representative is a certain pattern of activity of the network.

But, even if the network operates only as a *memory*, when its operation is restricted to recognition and retrieval, the outcome evoked by a stimulus may be very sensitive to the context. This is because the specific N-bit word, which is the attractor-network-state, may be composed of a variety of contextual elements which may have been present at the learning stage, or that have been required by the output reaction. For example, on learning a number, one may be intrinsically associating with it its sound, its visual form, perhaps even some of the preceding or following numbers in a recitation. It may even be that the selected form of the attractor is influenced by the expectation that the number, when shown, will be voiced, or vice versa etc. It may also be that, with experience, the various aspects become separated and the canonical *representational* features of the attractor become increasingly distilled, as a 'product of cognitive activity, not the general basis for it'[42].

It appears more natural to associate network (attractor) states with concepts in an expansive way. In other words, if an ANN is to perform some manipulation on numbers which are in the range 0 to N, one needs only $\log_2 N$ bits, and hence apparently only that number of neurons in the network. However, even the 'purest' mathematical concepts can be retrieved by a wide variety of associations. This fact suggests that at least on the level of the memory, network states associated with those numbers would have many more bits (and therefore

neurons) than arithmetical frugality would require. The extra bits would allow for the retrieval of the number by a different stimulus with which it had been conditionally associated during learning. Such a process has recently begun to be implemented[65], see e.g., Section 9.3.3.

It may be imagined that by the time the necessary levels of processing have been ascended, to reach the actual arithmetical operation, data reduction has already been affected to purify the *representation* of the number. But such data reduction must be a very complex adaptive process, and arithmetic, in learning or in practice, is likely to be performed on the gross states, carrying along all the associations. Moreover, as time goes on, the same operations would, in this view, be performed on 'numbers' carrying varying associative baggage. As the accompanying baggage changes, so do the 'rules' by which these states are combined. This attitude is expressed in greater detail when we present, in Section 5.4.2, a network which counts. It is this kind of perspective that seems to suggest an interpretation closer to that of Shanon. But this is, of course, provided one can convincingly show that ANN's can perform in an interesting way.

Finally, it is not completely clear that the lines dividing this issue are very well defined. It is difficult to free oneself from a shadow of suspicion that it is merely a semantic dispute. After all, if the observation of surface behavior, which clearly *has* computational aspects, *must* imply that any deeper explanation is computational, and in terms of representations, then one is facing a tautology. On the other hand, if some explanation in terms of a model expressed on a deeper (lower) level were to be put forward, that explanation could be implemented and tested on a computer. The operation of the computer can then be reinterpretted as a manipulation of representations. All one has to do is to reread the algorithm in terms of the momentary surface behavior. So we will put this controversy aside, hoping that if a reductionist approach, such as is advocated here, produces interesting behavior, it will contribute something to the clarification of the controversy.

Bibliography

[1] C. Eccles, *The Understanding of the Brain* (McGraw-Hill, NY, 1973).

[2] J.A. Fodor, *The Language of Thought* (Thomas Y. Crowell, NY, 1975).

[3] J. Searle, Minds, brains and programs, *The Behavioral and Brain Sciences*, **3**, 552(1980).

[4] D.O. Hebb, *Essay on Mind* (Lawrence-Erlbaum Assc., Hillsdale NJ, 1980).

[5] J.J. Hopfield, Collective processing and neural states, in C. Nicollini ed., *Modeling and Analysis in Biomedicine* (World Scientific, NY, 1984).

[6] A.L. Hodgkin and A.F. Huxley, A quantitative description of current and its application to conduction and excitation in nerve, *Journal of Physiology*, 117, 500(1952).

[7] B. Katz, *Nerve, Muscle and Synapse* (McGraw-Hill, NY, 1966).

[8] Aristotle, *De Anima*, H. Lawson-Tancred (trans.) (Penguin Books, Harmondsworth, 1986).

[9] M. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, Mass., 1969).

[10] A.M. Turing, Computing Machinery and Intelligence, *Mind*, Vol. LIX, No.236 (1950).

[11] K. Popper, *The Logic of Scientific Discovery* (Hutchison, London, 1959).

[12] J.C. Eccles, The neurophysiological basis of experience, in M. Bunge (ed.): *The Critical Approach to Science and Philosophy* (The Free Press of Glencoe, London, 1964).

[13] P. Churchland, *The Neurophilosophy of Mind* (MIT Press, Cambridge Mass., 1986).

[14] G.M. Edelman and V.B. Mountcastle, *The Mindful Brain* (MIT Press, Cambridge Mass., 1978).

[15] M. Abeles, *Local Cortical Circuits* (Springer-Verlag, Berlin, 1982).

[16] T.H. Bullock, R. Orkand and A. Grinnell, *Introduction to Nervous Systems* (W.H. Freeman, San Francisco, 1977).

[17] K. Akert, K. Pfenninger, C. Sandri and H. Moor, Freeze etching and cytochemistry of vesicles and membrane complexes in synapses of the central nervous system, in G.D. Pappas and D.P. Purpura, eds., *Structure and Function of Synapses* (Raven Press, NY, 1972).

[18] R. Poritsky, Two and three dimensional ultrastructure of boutons and glial cells in the motoneuronal surface of the cat spinal cord, *Journal of Comparative Neurology*, **135**, 423(1969).

[19] S.W. Kuffler, J.G. Niccols and A.R. Martin *From Neuron to Brain* (Sinauer, Sunderland, Mass., 1984).

[20] J.C. Eccles, *The Physiology of Nerve Cells* (The John Hopkins Press, Baltimore, 1957).

[21] E.R. Kandel, *Cellular Basis of Behavior* (W.H. Freeman, San Francisco, 1976); *Behavioral Biology of Aplysia* (W.H. Freeman, San Francisco, 1979).

[22] W.S. McCulloch and W.A. Pitts, A logical calculus of the ideas immanent in neural nets, *Bull. Math. Biophys.*, **5**, 115(1943).

[23] F. Rosenblatt, *Principles of Neurodynamics* (Spartan Books, Washington DC, 1962).

[24] B. Russell and A.N. Whitehead, *Principia Mathematica* (Cambridge University Press, London, 1910-1913).

[25] B. Widrow and M.E. Hoff, Adaptive switching circuits, *IRE WESCON Convention Record*, **4**, 4-96(1960); B. Widrow, G.F. Groner, M.J.C. Hu, F.W. Smith, D.F. Specht and L.R. Talbert, Practical applications for adoptive data-processing systems, *IRE WESCON Technical Papers* p.1 (1963).

[26] P. Peretto, Collective properties of neural networks, *Biol. Cybern.*, **50**, 51(1984) and P. Peretto and J.-J. Niez, Stochastic dynamics of neural networks, *IEEE Transactions: SMC* **16**, 73(1986).

[27] D.J. Willshaw, O.P. Buneman and H.C. Longuet-Higgins, Non-holographic associative memory, *Nature, Lond.*, **222**, 960(1969).

[28] T. Kohonen and M. Rouhonen, Representation of associated data by matrix operators, *IEEE Trans. Comput.*, **22**, 701(1973).

[29] K. Fukushima, Cognitron: a self-organizing multilayered neural network, *Biol. Cybern.*, **20**, 121(1975).

[30] L.N. Cooper, A possible organization of human memory and learning, in B. Lundqvist and S. Lundqvist eds. *Collective Properties of Physical Systems* (Academic Press, NY, 1973).

[31] G. Palm, On associative memory *Biol. Cybern.* **36** 19(1980) and Neural Assemblies: An Alternative Approach to Artificial Intelligence, in E. Frehland ed. *Synergetics – From Microscopic to Macroscopic Order* (Springer-Verlag, Berlin, 1982).

[32] E.T. Rolls, Information representation, processing and storage in the brain: Analysis at the single neuron level, in J.P. Changeux and M. Konishi eds., *Learning* (Springer-Verlag, Berlin, 1986).

[33] S. Amari, Characteristics of random nets of analog neuron-like elements, *IEEE Trans. SMC*, **2**, 643(1972).

[34] S. Amari, Learning patterns and pattern sequences by self-organizing nets of threshold elements, *IEEE Trans. Comput.*, **21**, 1197(1972).

[35] E. Caianiello, TITLE, *J. Theor. Biol.*, **2**, 204(1961).

[36] W.A. Little, The existence of persistent states in the brain, *Math. Biosci.*, **19**, 101(1974); and W.A Little and G.L. Shaw, Analytic study of the memory storage capacity of a neural network, *Math. Biosci.* **39**, 281(1978).

[37] J.J. Hopfield, Neural networks and physical systems with emergent selective computational abilities, *Proc. Natl. Acad. Sci. USA*, **79**, 2554(1982).

[38] J.A. Feldmam and D.H. Ballard, Connectionist models and their properties, *Cognitive Science* **6**, 205(1982).

[39] G.E. Hinton and J.A. Anderson, *Parallel Models of Associative Memory* (Lawrence Erlbaum, Hillsdale, 1981).

[40] D.E. Rumelhart and J.L. McClelland eds. *Parallel Distributive Processing: Explorations in the Microstructure of Cognition* Vols. I and II (MIT Press, Cambridge Mass., 1986).

[41] Z.W. Pylyshyn, Computation and cognition: issues in the foundations of cognitive science, *The Behavioral and Brain Sciences*, **3**, 111(1980).

[42] B. Shanon, The role of representations in cognition, in J. Bishop, J. Lockheed amd D.N. Perkins eds., *Thinking* (Lawrence Erlbaum, Hillsdale NJ, 1987); The non-abstractness of mental representations, *New Ideas in Psychol.*, **5**, 117(1987); The semantic representation of meaning: a critique, *Psychological Bulletin*, **104**, 70(1988).

[43] M.E. Goldberg and C.J. Bruce, Cerebral cortical activity associated with the orientation of visual attention in the rhesus monkey, *Vision Res.*, **25**, 471(1985).

[44] P. Fatt and B. Katz, The effect of inhibitory nerve impulses on a crustacean muscle fiber, *J. Physiol. (Lond.)*, **121**, 374(1953).

[45] H.B. Barlow, Single units and sensations: A neuron doctrine for perceptual physiology?, *Perception* 1, 371(1972).

[46] J.-P. Changeux, *Neuronal Man* (Oxford University Press, NY,1986).
[47] A. Lwoff, *Biological Order* (MIT Press, Cambridge, Mass., 1962).
[48] H. Atlan, Self creation of meaning, in Arechi, Erikson and Haken Eds. *The Physics of Structure and Complexity, Physica Scripta*, **36**, 563(1987).
[49] H.B. Barlow and J.D. Mollon, *The Senses* (Cambridge University Press, Cambridge, 1982).
[50] D.H. Hubel and T.N. Wiesel, Responses of cells and organization of of visual cortex., *Proceedings of the Royal Society* **B 198**, 1(1977).
[51] M.M. Merzenich and J.H. Kaas, Principles of organization of the sensory-perceptual systems in mammals, *Progress in Psychobiology and Physiological Psychology*, **9**, 1(1980).
[52] T. Kohonen, Pekka Lehtio, P. Rovamo, J. Hyvarinen, K. Bry and L. Vainio, A principle of neural associative memory, *Neuroscience* **2** 1065(1977).
[53] T. Kohonen, *Self-Organization and Associative Memory* (Springer-Verlag, Berlin, 1984).
[54] J.A. Fodor, *The Modularity of Mind* (MIT Press, Cambridge, Mass., 1983).
[55] S. Grossberg, On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks, *Journal of Statistical Physics*, **1**, 319(1969). For references to many later publications in this line of research see e.g., *The Adaptive Brain: I. Learning, Reinforcement, motivation and rhythm* and *The Adaptive Brain: II. Vision, speech, language and motor control* (North-Holland, Amsterdam, 1986).
[56] S. Paternello and P. Carnevali, Learning networks of neurons with Boolean logic. *Europhysics Lettrs*, **4**, 503(1987).
[57] J.A. Fodor, Information and association, in M. Brand and R.M. Harnish eds. *The Representation of Knowledge and Belief* (The University of Arizona Press, Tucson, 1986).
[58] D. Andler, private correspondence.
[59] T.J. Sejnowski, P.K. Kienker and G.E. Hinton, Learning symmetry groups with hidden units: beyond the perceptron, *Physica*, **D22**, 260 (1986).
[60] D.J. Amit, Neural Networks counting chimes, *Proc. Natl. Acad. Sci. USA*, **85**, 2141(1988).

[61] S. Thorpe, K. O'Regan and A. Pouget, Humans fail on XOR pattern recognition problems, *Neural Networks: From Models to Applications*, L. Personnaz and G. Dreyfus, eds. (IDEST, Paris, 1989).
[62] M. Mezard and G. Parisi, Mean field equations for the matching and the traveling salesman problems, *J. Physique Lett.* **46**, 77(1985).
[63] J.J. Hopfield and D.W. Tank, 'Neural' computation of decisions in optimization problems, *Biol. Cybern.*, **52**, 141(1985).
[64] N.M. Grzywacz and A. Yuille, Motion correspondence and analog networks, *MIT Artif. Intell. Memo.* 888(1986); for a general review see E.C. Hildreth and C. Koch, The analysis of visual motion: From computational theory to neuronal mechanisms, *Ann. Rev. Neurosci.* **10** 477(1987).
[65] D. Lehmann, Memory and the formation of associations in neural nets, Hebrew University Computer Science preprint (1987).

2

The Basic Attractor Neural Network

2.1 Networks of Analog, Discrete, Noisy Neurons

2.1.1 Analog neurons, spike rates, two-state neural models

The two-state representation of neural output, which enjoys such a wide popularity among modelers of neural networks, is often considered oversimplified both by biologists and device designers. Biologists prefer to describe relevant neural activity by *firing rates*. These are continuous variables describing mean spike activity of neurons, rather than the discrete variables which describe the presence or the absence of an individual spike. Device designers prefer sometimes to think in terms of operational amplifiers, currents, capacitances, resistors, and continuous time equations. It turns out that in a wide range of parameters the performance of a network as an ANN is largely independent of the representation[1]. As we shall see in the following chapters, e.g., Chapter 5, the discrete representation provides a more transparent framework for structured manipulations of attractors.

Eventually the gap between the different descriptions, closes, because in our formulation of the output mechanism, Section 1.4.4, a significant event is said to occur when a time average over spike activity is found to be high, which is nothing but a measure of mean *firing rates*. On the other hand, in the analog description, in terms of electronic components, which is deterministic in structure, one eventually generates spikes, stochastically, at a mean instantaneous rate proportional to the continuous variable at hand, e.g., the potential excess over the

threshold. This is a concession made to the biologist who concentrates after all on measuring spikes[1]. In Hopfield's words:

> For high gain systems, the stable states of the real circuit [the analog network] will be exactly those of the stochastic [discrete] model.
>
> The continuous model supplements, rather than replaces, the original stochastic description.
>
> ...it will often be more practical to develop ideas and simulations on that model even when use on biological neurons or analog circuits is intended.

Here we will describe the network of analog neurons to provide a unified language. However, in most of what follows we will confine ourselves to discrete two-state neurons. The rest of this section can, therefore, be considered as a digression, inessential for the remainder of the discussion.

Networks of analog neurons

The dynamical variables are the membrane potentials, U_i $(i = 1, \ldots, N)$, which are the PSP's accumulated in the soma of each of the N neurons composing the network. These potentials vary due to three factors:

- Currents induced by the activity of pre-synaptic neurons.
- Leakage via the finite resistivity of the membrane.
- Input currents from outside the network.

The rate of change of the membrane potential of neuron i can be described by the equation ([2,3,1,5])

$$C_i \frac{dU_i}{dt} = \sum_{j,j\neq i} J_{ij} g(U_j) - \frac{U_i}{R_i} + I_i. \tag{2.1}$$

This equation should be compared and contrasted with Eq. 10.1. A sketch of a standard circuit is depicted in Figure 10.3. The right hand side contains the currents contributing to charging the input capacitance C_i of neuron i by the potential dU_i. The first term represents the currents induced by the activity of all the other neurons, the second

term is the leakage current due to the trans-membrane resistance R_i between the interior and the exterior of the neural cell, and the third term represents input currents from sources outside of the network. In the absence of the first and third terms, the potential U_i decays to zero with the time constant R_iC_i. While the last two terms are self-evident, the first requires an explanation.

If the potential of neuron i is equal to U_i, then the *output* of the neuron can be described by a mean *firing rate* – number of spikes per unit time, V_i. On a suitable time scale, a neuron's output can be viewed as a short-time average of its firing rate V_i. The activity of a neuron is then a continuous variable V_i, which varies between $V_i = 0$, corresponding to the quiescent state, and the maximal $V_i = 1$, which is a spike every neural cycle-time. The unit of time can be chosen differently, but then the maximal value of V will be the maximal number of spikes in the new time unit. The firing rate is related to the membrane potential by:

$$V_i = V_0 g(U_i), \tag{2.2}$$

where the function g, the *gain*, is assumed to be monotonically increasing from 0 to 1, exhibiting a *sigmoid* form, e.g., Figure 2.1[4]. The coefficient $V_0(= 1)$ is the maximum firing rate of a cell. If transmitter release is proportional to the mean firing rate in the pre-synaptic neuron, then the first term on the right hand side of Eq. 2.1 is the current into the post-synaptic neuron i. It is assumed to be a linear sum of the currents from all connecting pre-synaptic neurons, modulated by the efficacy of the corresponding synapses, which convert the transmitter into membrane potential. Clearly, many simplifications are involved, mainly about the absence of time delays and of spatial structure of post-synaptic inputs.

A convenient choice for the *gain* function g is

$$g(U) = \frac{1}{2}[1 + \tanh(GU)] \tag{2.3}$$

The parameter G describes the slope of the sigmoid function (see Figure 2.1), and for $G \to \infty$, the output assumes only the two discrete values 0 and 1, but now they have a different meaning: Either the neuron is quiescent or it is firing at its highest possible rate.

Equation 2.1 describes deterministic dynamics of the continuous variables U_i. The neural network is, of course, a stochastic system.

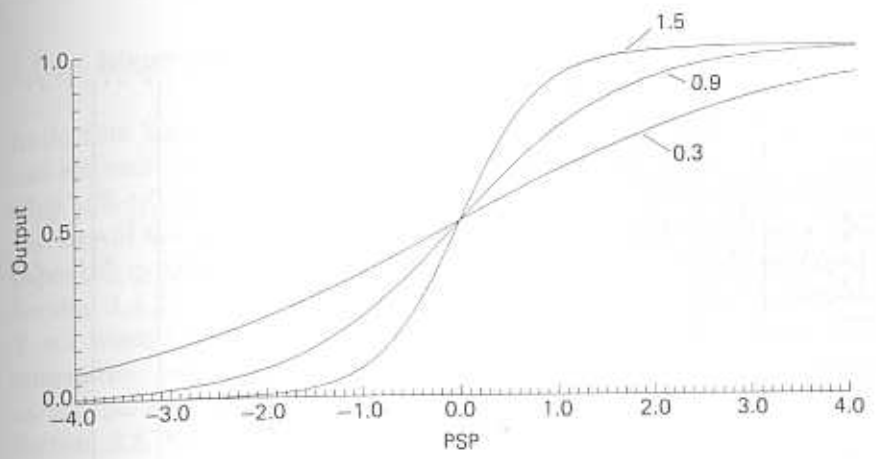


Figure 2.1: Continuous input-output relation for several values of the gain.

The discrepancy is reconciled by recalling that we have made a shift from the 'microscopic' variables – the spikes – to coarse variables – mean firing rates. Statistical physics is quite familiar with such shifts from stochastic to deterministic behavior accompanying a shift in the level of the description: Individual particles in a liquid, at finite temperature, behave stochastically, yet hydrodynamics is that discipline which describes the deterministic development of densities. Clearly, if we return to the level of spikes, the firing rates would have to be translated into stochastically realized spike trains. Moreover, if the communication between neurons is primarily affected by spikes then, as we explain in the following paragraph, the dynamics itself becomes stochastic.[1]

To reintroduce spikes one can proceed as follows: If the input potential of neuron i is U_i, the mean firing rate is $V_0 g(U_i)$. A real neuron will emit a spike during the time interval dt with probability $V_0 g(U_i)dt$. If the spike is emitted in this time interval, the potential is reduced by a fixed quantity U_0. Even in the absence of external input, neurons continue to receive their inputs via the first term in Eq. 2.1, namely based on mean firing rates, but those rates are modified stochastically by the emission of the spikes. Spikes appear in total asynchrony. Simulations of such networks are presented in Figure 2.2.

[1] I am grateful to John Hopfield for a helpful correspondence on this point, based on unpublished studies he had carried out in collaboration with David Tank.

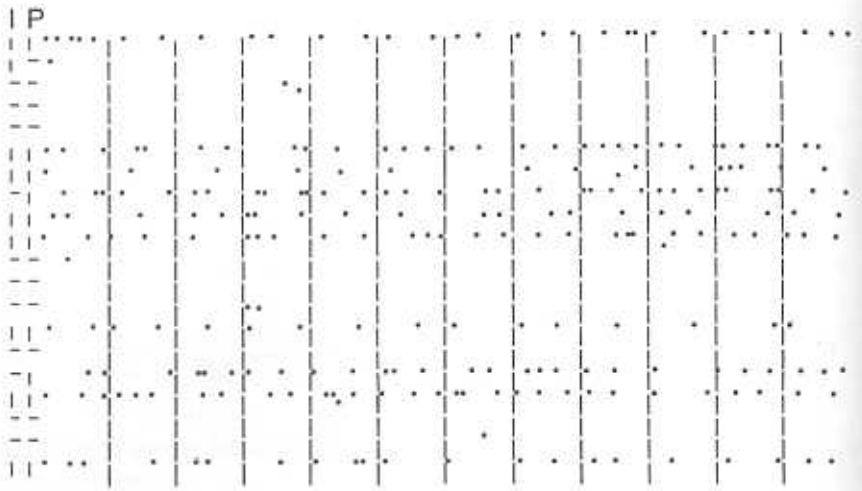


Figure 2.2: Raster of spikes of a randomly selected set of 20 out of 60 neurons generated stochastically by analog neurons receiving spike inputs. Two columns on the left are, respectively, I initial state, P retrieved pattern. Vertical (horizontal) bar on left – neuron active (passive). Network starts with 10% errors.

The same equations (2.1–2.3) describe an electrical circuit involving operational amplifiers, capacitors and resistors[1], as is described in Chapter 10. Cell bodies are mimicked by a combination of an RC circuit, which allows the accumulation of charge in the presence of a resistive leak, together with an amplifier which produces the gain function $g(u)$ from the input. The synapses are resistances with conductance J_{ij}. Since resistances are always positive, the excitatory and inhibitory nature of synapses is captured by an artefact which realizes the dynamic behavior of the cell body by doubling the output from every amplifier. One output is direct and one is inverted. This is discussed in detail in Section 10.2.

The collective behavior of such an electric network, described by Eq. 2.1 and Eq. 2.3, is qualitatively similar to that of a stochastic, discrete, two-state neural network[1]. In particular, for sufficiently large values of the parameter G in Eq. 2.3, the neural outputs in a stationary state are driven to their extreme values, of either 0 or 1. This means that some neurons are inactive while the others fire at the maximum rate. We shall see that this is exactly what happens when a stationary state is reached in networks of two-valued neurons.

2.1.2 Binary representation of single neuron activity

In Section 1.3.1 we have reduced the biological neuron to a formal logical element, which transforms a continuous input into one of two possible values of output. This output was regarded as the instantaneous activity of the neuron and was described by a binary variable σ which takes the values 1 or 0. An alternative representation was mentioned in Section 1.4.2, assigning to each neuron a binary variable S, such that $S = 1$ when the neuron is active and $S = -1$ when it is inactive. This alternative representation is particularly appealing to the physicist, as it relates naturally to Ising models of magnetic systems,[2] see e.g., Section 3.2. In the following chapters, as we proceed to construct the model and to explore its properties, we shall frequently benefit, both conceptually and technically, from analogies with Ising spin systems. Historically, this analogy seems to have been first put forward in 1954 by Cragg and Temperley[8], see also ref. [9].

The (0,1) representation and the (−1,1) are equivalent. They are related by the transformation (Eq. 1.8)

$$S = 2\sigma - 1. \tag{2.4}$$

In terms of the S-variables, the PSP accumulated on neuron i, the equivalent of Eq. 1.4, at the end of a certain *summation period*, becomes

$$U_i = \frac{1}{2} \sum_{j, j \neq i}^{N} J_{ij}(S_j + 1). \tag{2.5}$$

Where J_{ij} is the synaptic connection matrix with positive or negative elements, describing excitation or inhibition (Section 1.2.2). The response of this neuron in the absence of noise, the equivalent of Eq. 1.3, is determined by

$$S_i = \mathrm{sign}(U_i - T_i), \tag{2.6}$$

[2] At some point electrical engineers also seem to have appreciated the virtues of this representation, see e.g.,[6,7]

with threshold T_i. The sign-function is equal to the sign of its argument, namely

$$\text{sign}(x) \equiv \begin{cases} +1 & \text{if} \quad x > 0 \\ -1 & \text{if} \quad x < 0. \end{cases} \tag{2.7}$$

Let us now rearrange the argument of the sign-function as follows:

$$S_i = \text{sign}(h_i + h_i^e), \tag{2.8}$$

where

$$h_i = \sum_{j,j\neq i}^{N} J'_{ij} S_j \tag{2.9}$$

with

$$J'_{ij} = \frac{1}{2} J_{ij}$$

and

$$h_i^e = \sum_{j,j\neq i}^{N} J'_{ij} - T_i. \tag{2.10}$$

- The parameters h_i and h_i^e have a simple interpretation in the analogous system of magnetic spins S_i connected by the interaction matrix J_{ij}. See e.g., Section 3.2.1. The variable h_i is the local magnetic field experienced by the spin analog of neuron i, due to its interaction with all the other spins in the system, and h_i^e is the fixed, 'external', magnetic field at site i, which does not depend on the orientation of the other spins. Eq. 2.8 indicates that the noiseless dynamics always acts to align the spins with their instantaneous local total magnetic fields.

When the average membrane potential $\bar{U}_i = \sum_j J'_{ij}$ is approximately equal to the threshold T_i, the response of a neuron is most sensitive to changes in its inputs. As will be shown in Section 6.4.2, operation in the vicinity of the neuronal threshold increases the storage capactiy of the network. We can, therefore, simplify the model by assuming that

$$h_i^e = 0 \tag{2.11}$$

This assumption implies that the real thresholds of the neurons are balanced by the mean activity in the network. It eliminates thresholds,

in the (–1,1) description, from the present discussion to allow a more concise and transparent treatment of the most important features of the model, focusing attention on the role of the synaptic efficacies in controlling the **collective** performance of ANN's. The thresholds will be reintroduced at several places in later chapters, when their effects become relevant.

The simple model of the operation of a single neuron can now be summarized by:

$$\begin{aligned} S_i &= \text{sign}(h_i) \\ h_i &= \sum_{j,j\neq i}^{N} J_{ij} S_j \end{aligned} \tag{2.12}$$

where, to simplify the notation, we have dropped the prime on J_{ij} and incorporated the factor $\frac{1}{2}$ in its definition.

2.1.3 Noisy dynamics of discrete two-state neurons

The input-output relations described in the last two sections are *deterministic* – a given input is always followed by the same output. In reality, the synaptic transmission is a noisy process and the potential on the post-synaptic membrane is not determined precisely by the values of the J_{ij}'s. The firing rule, Eq 2.6, implies that the new activity state of neuron i will be determined with *certainty* by whether[3]

$$\frac{1}{2} \sum_{j,j\neq i}^{N} J_{ij}(S_j + 1) > T_i \tag{2.13}$$

or its opposite holds. The improbable equality of the sum of so many entries to T_i introduces some indeterminacy, but it will be ignored for the sake of the present discussion.

Yet, neurons operate in a noisy environment and we now turn to a discussion of the origins of the noise as well as to the modifications it introduces in the dynamical process. This was first done by Little[9]. See also e.g., ref. [7]. An action potential traveling down an axon of the pre-synaptic neuron reaches the *synapse*, where a chemical, the *neurotransmitter* is stored in a large number of *vesicles*. This is visualized in

[3]Note that the original notation for the J's is briefly restored.

Figure 1.3. The strong fluctuation in the membrane potential at this site causes the release of the contents of several vesicles – several quanta of chemical transmitter. These quanta of neuro-transmitter affect the membrane of the post-synaptic neuron, by activating receptor sites, so as to cause ionic *input current*, which, eventually, contributes to the potential on the membrane of the post-synaptic cell body. This process is affected by several sources of noise, all of which are well established experimentally[10]:

1. The number of vesicles discharged upon the arrival of an action potential varies at random according to the Poisson probability distribution, with a mean determined by the value of J_{ij}.

2. The size of the quanta may vary over a large range, but the mean and variance of their distribution is common to all the synapses. The contribution of each quantum to the PSP is given by a Gaussian probability density, independent of i and j.

3. Even in the absence of an action potential there is a small random rate of spontaneous discharge of chemical transmitter across the synaptic cleft.

When these effects are taken into account, the post-synaptic potential U_i becomes a Gaussian random variable with the probability distribution[11,12]

$$\Pr(U_i = U) = \frac{1}{\sqrt{2\pi\delta^2}} \exp\left[-\frac{(U-\bar{U}_i)^2}{2\delta^2}\right]. \tag{2.14}$$

The PSP defined in Eq. 2.5 is the mean value $\bar{U}_i$ of the random variable U_i. The width of the distribution, δ, is determined by the parameters associated with the different sources of noise mentioned above.

The probability that neuron i fire an action potential is equal to the probability that its membrane potential is higher than the threshold T_i. Therefore,

$$\Pr(S_i = 1) = \int_{T_i}^{\infty} dU \Pr(U_i = U) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\bar{U}_i - T_i}{\delta\sqrt{2}}\right)\right] \tag{2.15}$$

where the *error function*, erf(x), is defined by:

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x dt e^{-t^2}. \tag{2.16}$$

Similarly, the probability that neuron i does not produce an action potential is

$$\Pr(S_i = -1) = 1 - \Pr(S_i = 1) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\bar{U}_i - T_i}{\delta\sqrt{2}}\right)\right]. \tag{2.17}$$

To simplify the model, as in Section 2.1.2, one rearranges the argument of the error function as in Eq. 2.8, and adopts the assumption 2.11. The two probabilities can be combined into a single expression for the probability that the activity of neuron i assumes the value S_i, namely

$$\Pr(S_i) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{h_i S_i}{\delta\sqrt{2}}\right)\right] \tag{2.18}$$

where h_i is defined in Eq. 2.12. The last expression can be approximated to within 1% by

$$\Pr(S_i) = \frac{\exp(\beta h_i S_i)}{\exp(\beta h_i) + \exp(-\beta h_i)} = \frac{1}{2}[1 + \tanh(\beta h_i S_i)], \tag{2.19}$$

where

$$\beta^{-1} = 2\sqrt{2}\delta.$$

One should be appropriately struck by the similarity of this expression to the response function of the analog neuron, Eq. 2.3. The two do in fact play a rather similar role, but the exact relation between them has not yet been fully clarified.

An expression identical to Eq. 2.19 (with appropriate changes in the meaning of the variables) describes the single spin dynamics of an Ising system interacting with a *heat bath* at temperature $T = \beta^{-1}$, known also as the Glauber dynamics. (See e.g., Section 3.2.2). This stochastic process provides the link between the dynamics of neural networks and the thermodynamical treatment developed in the following chapters. The firing probability is plotted in Figure 2.3. The parameter β measures the width of the region of uncertainty. When $\beta \to \infty$, which is the

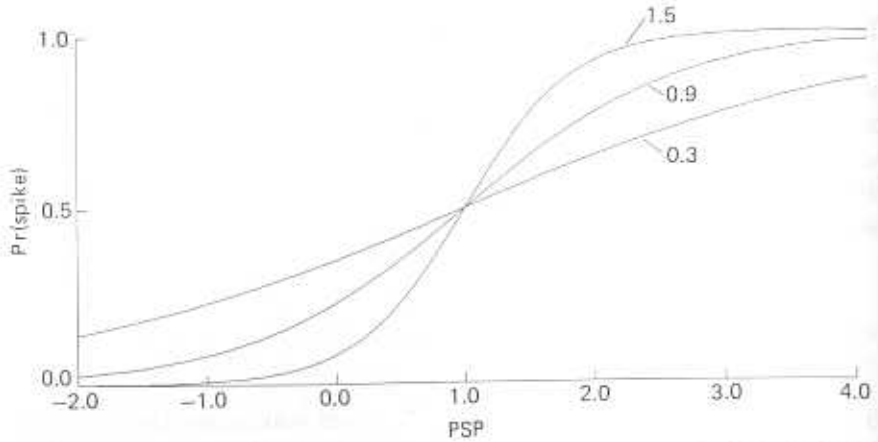


Figure 2.3: The firing probability as function of the local field (PSP) for several values of β.

noiseless limit (zero temperature), this region shrinks to zero and one recovers the deterministic firing rule of Section 2.1.2. The particular expression in Eq. 2.19 was chosen for convenience, in order to establish an analogy with the *statistical mechanics* of Ising spin systems.

Note that the curves in the figure have the same sigmoid shape as the curves in Figure 2.1. However, they have a very different meaning. In Figure 2.1, they represent *deterministic* input-output relations, while in Figure 2.3 they express a stochastic threshold response. Yet, the similarity of the two representations emerges again when one realizes that the limit $\beta \to \infty$ makes the discrete, stochastic neuron a deterministic system, while the limit $G \to \infty$ in Eq. 2.3 converts the deterministic, continuous neuron into a discrete system.

2.2 Dynamical Evolution of Network States

2.2.1 Network dynamics of discrete-neurons

In Section 1.4.1 we have elaborated on the role of network states, whose dynamical development is to represent meaningful cognitive or computational activity. A network state is defined by the collection of instantaneous activities of the individual neurons. In the network of discrete two-state neurons it can be expressed as an N-bit word, whose

elements are S_i^I. The superscript I labels the particular network state and the subscript i labels a particular neuron. A network state evolves in time, tracing a trajectory in the space of the 2^N possible firing configurations. This can also be viewed as a march on the vertices of an N-dimensional hypercube. See e.g., Section 1.4.1. Eq. 2.19 defines the elementary rule which determines how an individual neuron 'decides' whether to fire, or not to fire, an action potential. It depends only on the instantaneous PSP collected by that neuron. It does not specify, however, how this 'decision' relates in time to similar 'decisions' made by the other neurons in the network.

This is not a simple question. In reality, a neuron fires whenever its accumulated PSP reaches the threshold. If threshold has not been reached, then the accumulated PSP decays gradually, via leaks through the membrane, as is described in the dynamical equation for the analog neurons Eq. 2.1. A simple approximation to this behavior is to assume that if a neuron has not fired within the *absolute refractory period*, the basic cycle-time of the network, then the accumulated PSP decays instantly to zero and the buildup of the PSP starts all over again. The failure to produce an action potential within the interval $\Delta t = 1$ can be viewed as an attempt to fire, which resulted in a negative 'decision'. The present simplified description has it that:

- On the average every neuron attempts to fire an action potential in every unit interval – cycle-time, independently of all other neurons. This implies a mean updating rate of the inverse of the basic cycle-time.

The outcome of this attempt depends on the signals which have reached the neuron from all other neurons in the network since the last such attempt. Some of these signals represent the activity of the contributing neurons in the previous unit time slot. Other signals, which had to cover longer axonal and dendritic distances, may represent earlier activity states which may have already been modified. This means that the collection of signals, making up the input to a neuron, does not really represent a well defined *network state* in the dynamical evolution in the network, because the latter was defined as an instantaneous snapshot of the activity of all neurons in the system. This discussion highlights an ambiguity left in Eq. 1.3, where the signals contributing to the PSP were left unspecified.

To capture such complicated dynamics is not easy and one might as well resort to the full analog equations, sprayed with stochastically generated spikes, as in Section 2.1.1. Instead, two simplified versions of network dynamics, which are tractable, have become popular.

- The first assumes that all the neurons update their activity states *simultaneously* at discrete time steps n, where $n = 1, 2, \ldots$, as if governed by a clock. The inputs of every neuron in the network are determined by the same activity state of the network in the time interval $(n-1) < t < n$. This type of dynamics can clearly be described in terms of network states. It will be referred to as *synchronous* or *parallel*.

- The second is the *asynchronous* or *sequential* dynamics, in which the neurons are updated one by one, in some prescribed sequence, or in a random order. In this mode every neuron coming up for a decision has full information about all the decisions of the individual neurons that have been updated before it.

These two types of dynamics will now be discussed separately.

2.2.2 Synchronous dynamics

It is assumed that the neurons are synchronized so that they can fire only at integral multiples of a time period. The PSP on each neuron at time $t = n$ is determined by the activities of all other neurons in the time interval $(n-1) < t < n$. At the beginning of each period the neurons start from a zero PSP, carrying no trace of previously accumulated inputs. In other words, after every time unit they all return to their resting membrane potentials. This type of dynamics was introduced in earlier studies of neural networks[13,7,9] and is still a favorite among investigators with a foothold in the culture of cellular automata. Given the inherent stochastic nature of neural communication, the strong synchronicity makes this type of dynamics even more unrealistic than would be implied by all the other simplifications that have been introduced. It does in fact have some special features which are not robust, such as two-cycles. See e.g., Section 2.2.4. Yet, for some features it can still serve as a reasonable guide.

According to Eq. 2.9, the local field (PSP)[4] at time $t = n$, on neuron i is

$$h_i(t) = \sum_{j,j\neq i} J_{ij} S_j(t-1). \tag{2.20}$$

The probability of transition from a state J, which is specified by the values of all N neuronal variables at time $t-1$, i.e., $\{S_i^J(t-1)\}$, to a state I, $\{S_i^I(t)\}$, at time t, is the product of the transition probabilities of all the individual single neurons (Eq. 2.19)

$$W(I \mid J) = \prod_{i=1}^{N} \Pr(S_i^I) = \frac{\exp(\beta \sum_i h_i^J S_i^I)}{\prod_i [\exp(\beta h_i^J) + \exp(-\beta h_i^J)]}. \tag{2.21}$$

The upper index on the local fields in Eq. 2.21, h_i^J, indicates that all these fields, Eq. 2.20, are determined by the activities of the neurons in a single network state J. In this case, $W(I \mid J)$ is a transition probability matrix, which describes a Markov chain. In other words, the probability that the network at time $t+1$ be found in network state I depends **only** on the state J in which it had been at time t.

In the noiseless limit ($\beta \to \infty$), the transition probabilities of the individual neurons become *step-functions*, as indicated in Figure 2.3, namely

$$\lim_{\beta\to\infty} \frac{1}{2}[1 + \tanh(\beta x)] = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 & \text{if} \quad x > 0. \end{cases}$$

The state J makes a deterministic transition to the state I according to

$$S^I(t) = \text{sign}\left[\sum_{j,j\neq i} J_{ij} S_i^J(t-1)\right], \tag{2.22}$$

for $i = 1, \ldots, N$. The sites at which the local field h vanishes accidentally, i.e., those neurons which happen to receive a PSP exactly equal to their threshold, must be treated with a special rule. The most natural one is to have those neurons change state with probability of one half. This implies that network states in which the local fields vanish at some sites cannot be strict fixed points.

[4] The terms local field and PSP will henceforth used interchangeably. The minor risk lest context be confused is well worth taking.

2.2.3 Asynchronous dynamics

The alternative class of dynamical processes for networks of discrete neurons is still an idealization of the real picture but comes closer to capturing the asynchronous and stochastic nature of the operation of a neural network. The underlying idea is that within a unit time interval the N neurons are updated at a mean rate of N updatings per cycle-time. One way of achieving this is to choose a random sequence of N neurons in each unit time interval, i.e., the time scale for updating all the neurons in the network has to be the average time between successive attempts of an individual neuron to fire an action potential. A single neuron is updated at any given time. Since each of the neurons could have been updated anywhere within the time interval, the PSP of a neuron under consideration may partly be determined by the neural activities in the previous time interval and partly by the new activity state of neurons within the present one. This effect is captured by dividing the elementary time interval into N equal sub-intervals[14],

$$\delta t = \frac{1}{N}. \tag{2.23}$$

One neuron is then updated at each multiple of δt, with its PSP composed of inputs of the other neurons in the network state at the previous time sub-interval δt. Thus neurons which are consecutively updated have inputs which may differ by a state of a single neuron. On this shorter time scale we recover a Markov process in the space of network states, with non-vanishing transition probabilities between states which differ by a single neural state.

When the i-th neuron examines in its turn the conditions for firing, it finds a local field which is determined by the network state in the previous δt time sub-interval. The next neuron will find a field determined by the same state, except for the neural state of neuron i, which might have been modified. The *elementary* transition probability from state J to state I, $W(I \mid J)$, is composed as follows:

- If network state I differs from state J by the neural activity of more than one neuron, then $W(I \mid J) = 0$.

- If these two states differ by a single neural state, then the transition probability matrix W is a product of
  - The probability that neuron k, which is to be updated, follows in the selection sequence the neuron i which was updated in entering state J.
  - The probability for the single neuron to change its state, Eq. 2.19 with h_i determined by Eq. 2.12.

This process is often further simplified in simulations. Instead of choosing a single neuron at each time interval, in which case it may happen that the same neuron is picked more than once before the entire network is visited, and others may be skipped, one selects a random sequence of **all** the neurons in the network, and updates their activity states in that order. A further simplification is to replace the random sequence by an orderly fixed sequence. The asymptotic properties of the dynamical behavior in systems of interest should be the same for all these variants. Otherwise one can hope for very little robustness in the presence of noise. Short time transients, and convergence times prior to the asymptotic behavior may, of course, vary.

In these asynchronous dynamical processes, the matrix of transition probabilities $W(I \mid J)$, which is of order $2^N \times 2^N$, is rather sparsely populated with non-zero entries. If the updating sequence is random, the matrix will have only $N + 1$ entries per row. One entry will be in the diagonal – the probability that configuration J remains unchanged, and N corresponding to the possible transitions to one of the N states which differ from J by a change in the state of a single neuron. If the updating sequence is more orderly, the number of non-vanishing probabilities at each step decreases further. In each row there will be exactly two non-vanishing entries per row in $W(I \mid J)$ – one diagonal term, which represents the probability that the particular neuron selected for updating does not change its state, and one for the probability that state J make a transition to state I in which the selected neuron has changed its state.

In terms of the march of the network on the vertices of an N-dimensional hypercube, the asynchronous dynamics implies a trajectory connecting adjacent vertices only. The random sequence allows steps from a vertex to any of its N neighbors. In the ordered sequence, the network is allowed to step only to one of the N neighbors of the vertex it happens to be on.

The matrix of transition probabilities from state J to state I in an entire sequence of N updatings, which would describe a transition taking place in a unit time interval, is the product of the N transition matrices for the elementary steps, described above. In general, this matrix is different from the corresponding matrix in synchronous dynamics, and both define different trajectories in configuration space. Moreover, for asynchronous dynamics this matrix depends, in general, on the order of updating and not only on the initial and final states. On the original unit time scale this dynamical process is, therefore, not a Markov chain. The exception is the random sequence which remains Markovian even on the larger time scale.

2.2.4 Sample trajectories and lessons about dynamics

In the previous section we have discussed a variety of possible dynamical processes which can accompany the modeling of a neural network described in terms of discrete, spike variables. Eventually each of these processes must be considered as an attempt to approximate some generic features of a stochastic, asynchronous dynamical system. It may, of course, be logically possible to have the operation of the cortex fully synchronized by a clock. It does seem rather unlikely considering the vast number of elements, the variability in dendritic and axonal lengths, etc. Moreover, as it will turn out even an asynchronous dynamical process can give rise to temporary collective behavior which appears rather synchronized on various time scales. We will therefore argue for an attitude which has it that even if synchronized network activity is observed empirically it is a result of robust cooperative behavior, rather than a *deus ex machina* on a scale of 10^{11} neurons and 10^{15} synapses.

But then what are the generic features which are well simulated by the various processes listed in Sections 2.2.2, 2.2.3? In order to have a pictorial reference for this discussion we present in Figures 2.4–2.9 (almost) complete trajectory maps of two networks of 4 neurons. There are two networks in the sense that there are two matrices J_{ij}. Each network is 'run' with three types of dynamics –

- synchronous,
- asynchronous with ordered sequence,
- asynchronous with random sequence.

For each type of dynamics we present the trajectory of each network starting from every possible initial condition. Each figure represents therefore 16 trajectories in a condensed notation which will be explained below. Before explaining the code we point out that the qualification 'almost', which appeared above, refers to the impossibility of generating a complete list of trajectories for the cases in which the sequence of updatings is random. The trajectories depend on the specific sequence of random numbers which determine the consecutive selections of neurons to be updated. Yet, this will not deter us from drawing general conclusions based on partial lists in the two corresponding cases, Figures 2.6 and 2.9.

The states of the network are 16 4-bit words. They are represented by the corresponding hexadecimal numbers, if every -1 is replaced by 0. Thus the circles with the symbols $0, 1, 2, \ldots, F$ stand for instantaneous states of the network $(-1,-1,-1,-1), \ldots, (1,1,1,1)$, correspondingly. Arrows indicate noiseless transitions between states in consecutive time intervals, namely after all spins have been updated. Arrows emanating from black dots point to the initial state of a particular trajectory. Trajectories which merge, by flowing into fixed points or cycles, are drawn together. The first network is specified by the synaptic matrix:

$$J_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ -1 & 0 & -1 & -1 \\ 1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 \end{pmatrix} \tag{2.24}$$

It corresponds to Figures 2.4–2.6. The second synaptic matrix is

$$J_{ij} = \begin{pmatrix} 0 & -1 & 1 & 1 \\ -1 & 0 & 1 & -1 \\ 1 & 1 & 0 & 1 \\ 1 & -1 & 1 & 0 \end{pmatrix} \tag{2.25}$$

which defines the dynamics in Figures 2.7–2.9.

The first two figures, 2.4 and 2.5, point to the fact that special types of updating procedures induce rather exotic dynamical asymptotic trajectories. In the synchronous procedure we find two limit cycles, one of two states, $7 \Leftrightarrow 8$, which attracts states 1 and E, and one cycle of four states, $2 \Rightarrow 5 \Rightarrow D \Rightarrow A \Rightarrow 2$, which attracts everyone else. On the other hand, the asynchronous sequential procedure of Figure 2.5

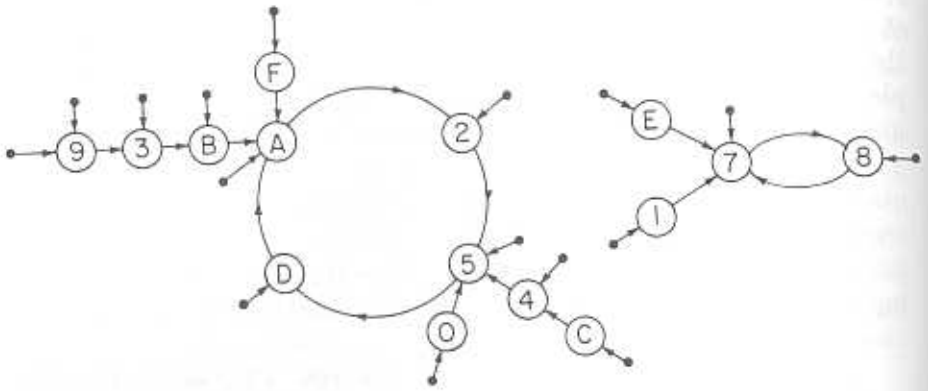


Figure 2.4: Trajectory map of 16 initial states in network 1 (J_{ij} of Eq. 2.24) with synchronous dynamics as described in Section 2.2.2.

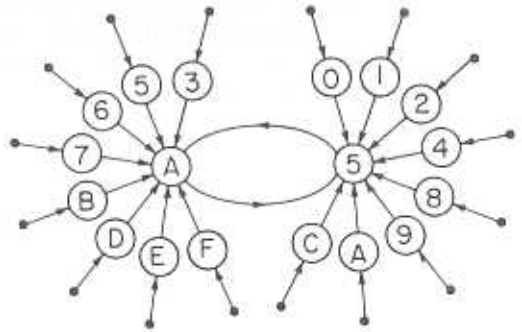


Figure 2.5: Trajectory map for network 1 with sequential, asynchronous dynamics.

brings forth a single limit cycle of two states, 5 ⇔ A, which attracts all possible 16 states of the network. The two states participating in this 2-cycle have nothing in common with the 2-cycle which appeared in the synchronous procedure. Instead, these two states are two of the four that make up the synchronous 4-cycle. One can safely assume that intermediate deterministic variants of these two procedures, in which some given subgroup of neurons is updated synchronously while the rest are updated in an ordered sequence, would produce further curios.

This point is further highlighted by the sample trajectories in Figure 2.6. The updating in a random sequence erases all semblance

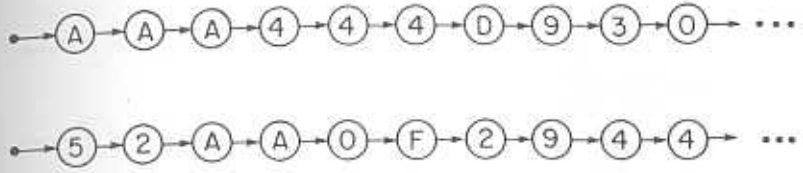


Figure 2.6: Sample trajectories in network 1 with asynchronous dynamics and random updating sequence.

of organized asymptotic behavior in a network with the same synaptic connections. Such extreme sensitivity to the precise updating process cannot but lead to the conclusion that no significance can be attributed to any of the sequences in Figures 2.4–2.6. This is a lesson which should be seriously contemplated by practitioners of cellular automata, where miraculously rich asymptotic behaviors are exhibited by systems which are highly synchronized in their dynamics.

Should one give up then? Figures 2.7–2.9 suggest that not all is capricious. Now the network is specified by the matrix Eq. 2.25. Synchronous dynamics produces a richer structure than before. Figure 2.7 shows five different attractors. Two are fixed points – 1 and B. They attract 0,6 and 9,F, respectively. Two are isolated 2-cycles – (1 ⇔ 8) and (7 ⇔ E), and a fifth is a somewhat more popular 2-cycle – (2 ⇔ D), which attracts 3, A, 5 and C. The significant point is the appearance of fixed point attractors. Once the network has reached any of those, it will persist there no matter what the updating procedure. The reason is that at a fixed point every neuron has a PSP which agrees with its activity state. Every neuron, in whatever order it is selected, will find it advantageous (at $T = 0$) to remain in its state, since it is a common feature of all updating procedures that individual neurons choose their new state according to Eq. 2.22.

Turning now to Figure 2.8 we find an even simpler situation. All trajectories terminate in one out of two fixed point attractors, 4 and B. These two are reversed states of each other, namely they are obtained from each other by reversing the states of all neurons. The origin of this symmetry is the simple fact that since the PSP's are linear in the S_i, reversing the signs of all S_i's reverses the sign of the h_i at every neuron. See also Section 4.5. There are initial states like 0, 6, 8, C or E which flow in a single step to the attractor 4. And initial states like

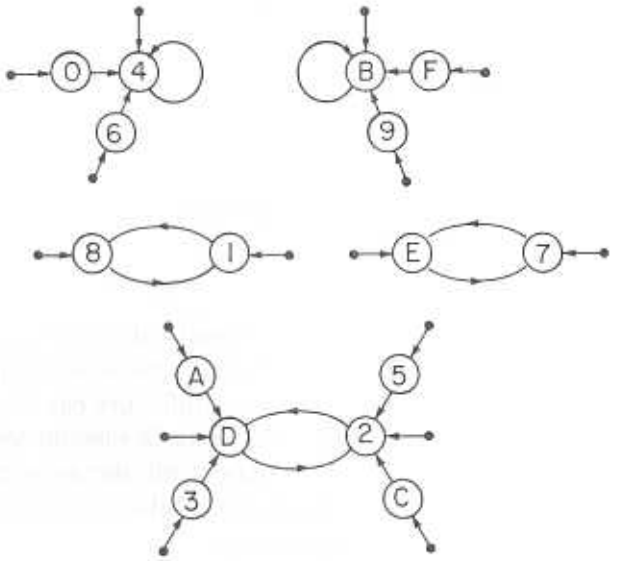


Figure 2.7: Trajectory map of 16 initial states in network 2 (J_{ij} of Eq. 2.25) with synchronous dynamics as described in Section 2.2.2.

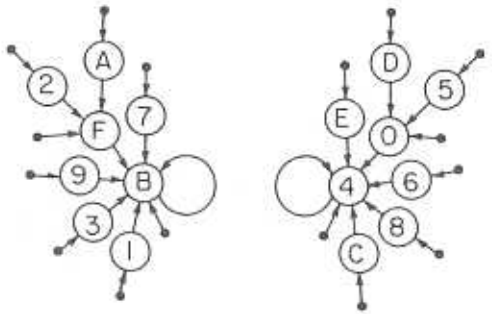


Figure 2.8: Trajectory map for network 2 with sequential, asynchronous dynamics. See e.g., Section 2.2.3.

5 or D which go there via state 0. The important point is that now the two different dynamical processes, the synchronous and the sequential asynchronous, display a common feature of the asymptotic dynamics of our little network: the existence of the two fixed point attractors and those are the same. This is a feature that will surface in any updating

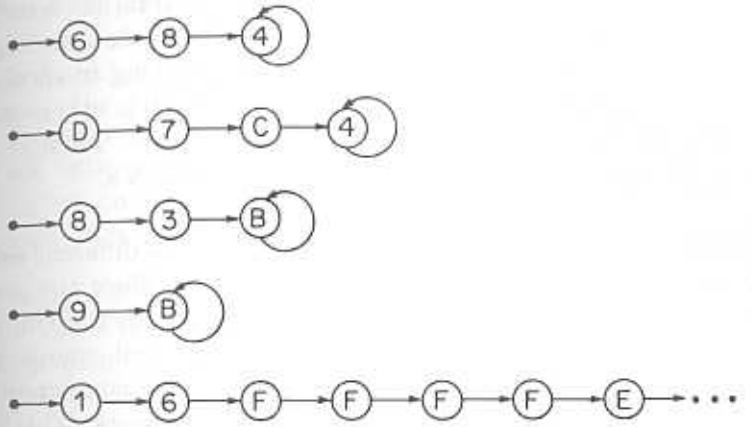


Figure 2.9: Sample trajectories in network 2 with asynchronous dynamics and random updating sequence.

procedure, given the decision rule Eq. 2.22. It is, therefore, an intrinsic feature of the network's organization, i.e., of the synaptic connection matrix.

Figure 2.9 presents five sample trajectories of network 2, in an asynchronous, random updating sequence. The common fixed point attractors of the previous two examples are present here as well. In fact, every trajectory ends up in one of the two attractors, which supports the suggestion that these are the only asymptotic features of this network, while the various two cycles in Figure 2.7 are artifacts of the particular updating procedure. But Figures 2.8 and 2.9 should be contrasted in order to bring out another important generalization. The trajectory starting from state 1 exhibits six steps without reaching any of the attractors. It has stayed in state F four cycle times before continuing on its way. This indicates that even if the common asymptotic features are unravelled, retrieval times and *basins of attraction* still depend on the updating procedure.

2.2.5 Types of trajectories and possible interpretation – a summary

Let us summarize the tentative lessons which can be *extrapolated* from the discussion in the preceding sections and from the examples

discussed above. These will not be theorems, but rather assertions which can be tested on such toy networks as the two studied in the previous section. They should appear on better footing to those who will venture to suffer through the next chapter, which is otherwise not essential for the comprehension of the rest of this book. These extrapolations help us in formulating expectations from ANN's.

- Different forms of the synaptic matrix give rise to different asymptotic dynamics under any updating sequence. Since the asymptotic dynamical behavior is to be associated with the computational performance of the network – memory, recall, counting etc. – we are led to the paraphrase that the acquired competence for performance resides in the connectivity of the network. This is further restated by saying that the competence of the network is in its organization, and hence learning is a process of self-organization.

- Three basic types of asymptotic dynamics can be discerned:

  - Chaotic — trajectories which wander in an uncorrelated way in the space of network states. Such trajectories would usually be very sensitive to the precise initial state.
  - Limit cycles — trajectories which lead, rapidly, to small cycles of states. The asymptotic behavior is correlated, yet here these types of dynamics will be discarded because
    * they seem too sensitive to the updating procedure,
    * we perceive a much simpler option – fixed points.
  - Fixed points — trajectories which lead the network to dwell for an appreciable period on a single state. These are the least sensitive to the updating prescription and, as will be shown later, they are rather insensitive to initial conditions, which widens the scope for modeling associative recall.

To reiterate, the asymptotic behavior of the network, on which we focus our interest, may depend also on the dynamical procedure, but such dependence is unwanted because no particular procedure is a faithful representation of the activity in the biological network. We therefore look for asymptotic properties which are insensitive to the updating procedure.

The *fixed point* attractors, to which we will refer as *attractors* in what follows, deserve a reiteration of some of the comments that have already been made about them:

- An attractor represents a dynamic situation of the neurons. When the network is in an attractor state all neurons with an $S = 1$ fire action potentials at the highest possible rate. Neurophysiologically they should be identifiable by the appearance of *bursts*.

- When the behavior of the network is translated back to the language of the analog network of Section 2.1.1, the attractor is manifested by a period in which some of the neurons have firing rates near maximum, while the others have very low rates.

- A network can reach an attractor even in completely asynchronous dynamics with a random updating sequence. On reaching an attractor state, a large number of neurons appear to be acting in concert, as if synchronized. The discrete description would have them fire spikes simultaneously, but this should not be taken too literally. A neuron can and will fire its spike stochastically anywhere in the basic time interval.

In the next section we will elaborate on various aspects of possible attractor scenarios and on their interpretation as a way of motivating the emphases that mark the present approach.

2.3 On Attractors

2.3.1 The landscape metaphor

The identification of significant network activity with fixed point attractors suggests the image of landscapes with valleys and hills on which the dynamical process executes a march which terminates at the bottom of one of the valleys[15]. This is a picture of special appeal to physicists, who have developed powerful tools and intuitions around energy functions, processes which make systems relax toward energy minima, etc. In the next chapter we expand on the formal aspects associated with the existence and the definition of such landscape functions for systems with many degrees of freedom. Here we will consider the pictorial features of a landscape as an associative memory, to express what a landscape is not and what it might be.

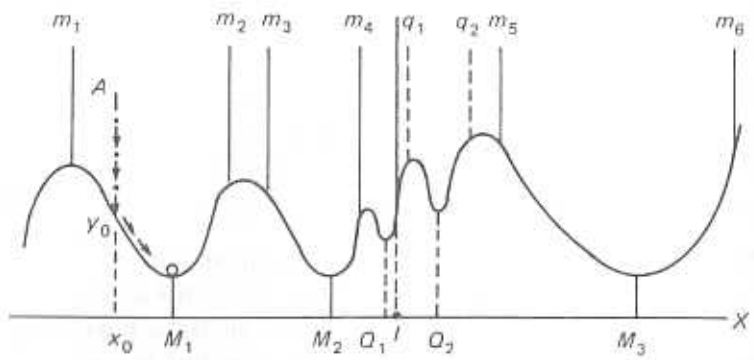


Figure 2.10: The one-dimensional landscape metaphor for associative, content addressable memory. M_1–M_3 are memories, Q_1, Q_2 are spurious states, $m_1, \ldots, m_6$ are maxima delimiting basins of attraction.

The most elementary landscape metaphor for *retrieval* from memory, for *classification*, for *error correction*, etc. can be represented as a one-dimensional surface with hills and valleys, Figure 2.10. The stored memories are the coordinates of the bottoms of the valleys, M_1, M_2, M_3 in Figure 2.10. Stimuli are drops which fall vertically on this surface, each stimulus carries with it x_0 – the coordinate along the horizontal axis of the point it starts from, which would be also the coordinate of the point y_0 it first touches on the surface.

The dynamical route of the drop, once on the surface, can be imagined as frictional gliding. A drop would, therefore, always move toward lower points, if there are any. After a long enough time such a drop will be found at the bottom of a valley which is found strictly downhill from the point of initial contact with the surface. Once the drop arrives at such a minimum, it will find no points of lower height and will stay there. The minima are, therefore, fixed point attractors. The coordinate of the particular minimum arrived at, such as M_1, M_2, M_3, Q_1 and Q_2, will be referred to as the memory retrieved by the stimulus. All stimuli between m_1 and m_2, for example, will retrieve the same memory M_1. The fact that they all retrieve the same memory is referred to as *associative recall*, or *content addressabilty*. This range is the analog of the *basin of attraction* of the particular memory.

It is not generally the case that a dynamical process can be described by a landscape picture, even if it has attractors. A landscape

picture applies only if some function of the variables of the system can be found that is decreased by every step of the dynamical process. In our little metaphor, such a function is, by definition, the potential energy, or the height of the drop. It is decreased by every step *and* is bounded from below. One may be tempted to follow the trajectories in a system and to try to construct a landscape function (Lyapunov function) by ascribing values to states in such a way that if there is a transition from state A to state B then B will be assigned a lower value than A. Yet for a general dynamical process this will not work because the construction will run into contradictions. A general dynamical process will allow transitions between the same two states by two different routes one ascribing a higher value to the first and the other to the second. A special case would be a cycle, as in Figure 2.4. If there are no limit cycles then an energy function can be constructed.[5]

Two additional features find their analog in this simple metaphor. In Figure 2.10 essentially every stimulus will arrive at an attractor. Yet, drops may take short or long times to arrive. The length of time may depend on the initial distance of the stimulus from the memory as well as on the type of dynamical process, which in our example is represented by the steepness of the slopes. This may have important consequences in the domain of perception which will be discussed in the next section. The second point is connoted by marking three of the attractors in the figure by M's and two by Q's. The difference is in their height. The M's are absolute minima while the Q's are local minima. Typically, when the dynamics of a non-linear process, such as the network's dynamics, is organized to have a set of attractors where memories are supposed to reside – by choosing the J_{ij}, for example – additional, uncalled for, attractors appear as well. See e.g, Sections 4.1.3 and 4.5.4. Those will be referred to as *spurious states*. Here we will try to dispose of them, or to consider as erroneous retrieval of an attractor that has not been ordered. How one can dispose of such attractors will be indicated in Section 2.3.5 and systematically discussed in Chapter 4. A possible speculative empirical situation that may be associated with their appearance and removal will be mentioned in Section 2.3.4.

[5] I am indebted to Professor Daniel Lehmann for this comment.

2.3.2 Perception, recognition and recall

The arrival of a trajectory, initialized by a given stimulus, at an attractor is the realization of retrieval and at the same time it is the assignment of meaning, as has been suggested in Section 1.4.4. Two basic times are involved

1. time of arrival at the attractor, to prevent loss of perception by refreshment.

2. time of presence in the attractor, to make readout possible.

The time in 1 has to be short enough so that arrival to the attractor will take place before the system turns to a new task, either by the arrival of a new stimulus or by some internal refreshment mechanism, which resets the network to another initial state. By 2, we imply again a time of persistence in the attractor, which must be long enough to allow a biological readout mechanism to ascertain that the network has reached an attractor. As was already mentioned in Section 1.4.4, if attractors are to be given a prominent cognitive role, the output mechanism must be able to detect the fact that the system has actually reached an attractor in order to distinguish the activity of the network from transient activity, in which consecutive network states are largely uncorrelated. This implies that

- The output mechanism must average the activity of the network over a long enough period of time, so that presence in an attractor is clearly distinguished from transient behavior by the especially large value of this average.

Note that the averaging must take place even if the network performs in noiseless conditions and rests perfectly aligned once it reaches an attractor.

In empirical situations of perception, one can make a gross distinction between recognition and recall. The first, which is the simpler task, involves the response that indicates a recognition that an object being perceived had been perceived previously – 'it is in memory'. Recall, on the other hand, implies a reconstruction of a complete item in memory based on a fragment of the familiar item. In terms of our metaphor in Figure 2.10 we may picture the distinction in the following terms: Recognition is an output device which is sensitive to the arrival of the

drop to an attractor, any attractor. Such an indiscriminate reading indicates that the stimulus was familiar. On the other hand, recall is a process by which a detailed item of information, specific to the particular attractor which had been reached, is propagated in the wider system to generate a response based on the specific detailed memory. See e.g., ref. [16].

Even the simpler task of recognition must take a number of neural cycle-times to distinguish an attractor from a transient. The total time for arrival to the neighborhood of an attractor, together with the time required for the output device to realize that an attractor has been reached, cannot be much longer than 40 milliseconds. An estimate for an upper bound of such a time may be inferred from experiments on fast recognition of pictures[17,18]. The tentative lesson we would like to draw from such experiments is that the network dealing with recognition is ready to accept inputs in a very rapid succession. Consequently, if a stimulus had not led to an attractor within a time shorter than the minimum time between two stimuli that the network can process it will get lost, as far as the perceptive, or cognitive, systems are concerned.

The existence of such a bound on the recognition time may limit further the basins of attraction of the attractors. Not all stimuli within the 'physical' basin will be cognitively perceived, but only those that can make it to the bottom of the valley within a biologically prescribed time interval.

2.3.3 Perception errors due to spurious states – possible role of noise

As we shall see in Chapter 4 the imprinting of a certain number of memories as attractors in a network is always accompanied by the appearance of spurious attractors, such as Q_1 and Q_2 in Figure 2.10.[6] A stimulus originating between m_4 and q_1, may retrieve Q_1. Since the Q-attractors are considered spurious, this will be an error of recognition or recall.

Often it is the case that the spurious states lie higher in the landscape than the stored memories. See e.g., Section 4.5.2. The reason why they remain attractors, despite the fact that there are lower states, is

[6] In one dimension the appearance of the attractors at Q_1 and Q_2 is rather arbitrary. In higher dimensional situations the spurious minima are topologically induced by the stored memories.

that they are surrounded by barriers, as in Figure 2.10. If the dynamics is strictly downhill, then the system cannot escape such local minima. Usually, however, neuronal dynamics is noisy and hence stochastic. As we have seen in Section 2.1.3, Eq. 2.19, neurons can make transitions into states which are opposed to the direction of their PSP. The probability for making such a transition depends on

- The size of the difference between the PSP and the threshold – the larger it is, the lower the probability of making a step in the wrong direction.
- The level of the noise – the lower it is, the lower the probability of steping in the wrong direction.

In the landscape metaphor, the noise may be pictured as heating the surface from below, causing the surface to vibrate.

The psychiatrist Hoffman[19] has likened the effect of noise in a model of a neural network to that of heating a bumpy frying pan in which a kernel of popcorn has been placed.[7] Usually the movement will be downhill, but with increasing temperature more and bolder jumps up will take place. In fact, the probability for a jump in the wrong direction decreases with the ratio of the size of the step to the level of noise. See e.g., Eq. 2.19. In the presence of noise, the system never settles quietly in an attractor. At low noise levels it will wander near the bottom of one of any of the five attractors in Figure 2.10. The activities of all these states near the bottom of an attractor have similar values of the variable x, and this will be recognized by the output mechanism, which has been assumed to average the activity over a certain time period. In the metaphor, the averaging must distinguish between cases in which the system is localized near an attractor and those cases in which it wanders over large uncorrelated regions.

At higher levels of noise, the drop may hop across barriers between adjacent minima. In this way, noise may be an agent for eliminating the effect of spurious states while preserving the retrieval in the stored memories. The surface in Figure 2.10 should make it clear that the barrier for crossing from a local minimum to a global one is lower than the barrier for the reverse process. This promises that there be

[7] I am indebted to Dr. M. Horenstein of the Institut National Marcel Rivière for bringing this article to my attention.

a *window* in noise values for which the spurious states be destabilized, while the stored memories remain good attractors.

The detailed study of the ANN at low storage levels, Chapter 4, confirms this possibility in a very remarkable way. It turns out that at low noise levels the dynamics teems with spurious states. At high levels of noise there are no attractors – the dynamics becomes *ergodic*, allowing the network to arrive at any of its states. In between there is an interval in the values of the noise parameter, in which noise is high enough to destabilize all spurious states and yet is sufficiently low to provoke very few errors in the retrieval of the stored memories. See e.g., Section 4.1.4.

In this context we find the role of noise rather counter-intuitive. It improves the retrieval of information. For some time now there has been a feeling in biological circles that this beneficial role of noise is rather universal in living systems[20]. Any system which qualifies to be called live must be rather complex. As such it necessarily has many metastable states, most of which would probably be mortal if the system were to persist in them. Noise, it is argued, prevents the freezing of biological systems, on the small scale, into metastable states.

2.3.4 Psychiatric speculations and images

The metaphors of landscapes, attractors, spurious attractors and noise has begun to give rise to psychiatric speculation. The first bold step in this direction has been taken by Hoffman[19]. He writes that in spite of the fact that

> Artificial models may never duplicate the richness and novelty of human cognition under either normal or pathological conditions. This, however, does not preclude the possibility that aspects of cognition and behavior can converge with artificial models. This convergence, if robust, could suggest ways of understanding underlying brain events.

The suggestion is that the ANN be considered as a metaphor for rather high brain functions which can lead to manifestations of *schizophrenia* and *mania* in speech disorders. Proper, normal speech, which in psychiatric terms is governed by a fixed *gestalt*, is likened to the presence of the ANN in a proper attractor, one that has been learned by the network. Speech is disorganized both in people suffering from *mania*

and from *schizophrenia*. The difference between the two disturbances is characterized as[19]:

- 'at any given time, the manic speaker is capable of accessing a single coherent plan or gestalt; disorganization derives from unregulated "shifts" from one gestalt to another as speech is being generated.'

- 'a schizophrenic utterance reflects a discourse plan whereby multiple disparate gestalt fragments are "patched" together as a single, stable incoherent structure.'

This description of the phenomena invites an immediate image from the repertoire of ANN's. The distinction between the two disturbances is likened, respectively, to that between

- noisy network behavior, whereby noise provokes 'unregulated' shifts from one attractor to another, and

- spurious states, which are 'stable' patches of attractor fragments.

Based on this imagery, Hoffman carried out a fair number of simulations on neural networks finding further analogues between network phenomena and psychiatric events. For details, the reader should consult the original article. We mention this bold speculation here not in order to support it. What we would like to emphasize rather is the power and richness of the ANN metaphor.

Spurious states and noise can be combined in a simpler paradigm.[8] It is observed that when recognition tests are given to epileptic children they react quite quickly but make many recognition errors. Given certain drugs, ritalin for example, recognition takes significantly longer, but the number of errors decreases. In the absence of a systematic study of this cognitive effect and of the biochemical effects of ritalin on cortical dynamics, we feel free to put forward the following metaphor: Consider in Figure 2.10 an input stimulus such as I. At low noise levels the network will retrieve rapidly the spurious attractor Q_1. At a higher noise level a jump across the barrier between Q_1 and M_2 becomes possible. As a result, in response to a stimulus like I the network will retrieve M_1, a true memorized pattern. Such retrieval will take significantly longer because I is further away from M_2 than from Q_1 and because it will take some extra time to climb over the barrier.

[8] This was suggested to me by neurologist Dr. R. Amit, of Haddassa Hospital, Jerusalem.

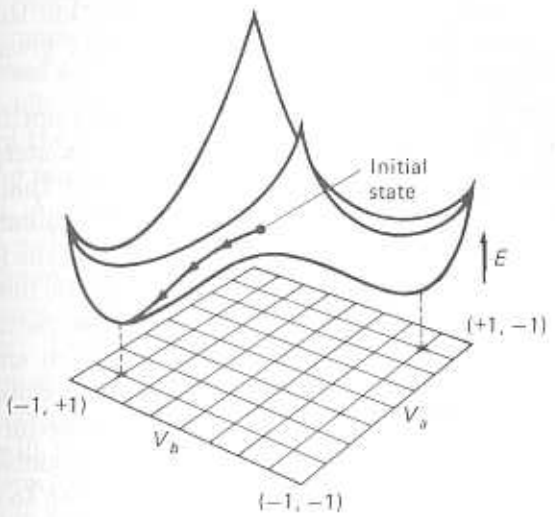


Figure 2.11: Landscape over a two-dimensional parameter space. The positions of the two minima are marked in the parameter plane by crosses. A ball rolling on the left will end up in one of the two minima, as indicated by trajectory. The two coordinates of the minimum will be the content of the retrieved memory. Which of the two minima will be retrieved depends on the basin in which the initial condition was. (From ref. [21], by permission.)

To conclude this section we present in Figure 2.11 a landscape over a two dimensional parameter space. While two is a small number compared to the dimensions of interesting parameter spaces, the figure does help to extend our intuitions.

2.3.5 The role of noise and simulated annealing

Noise has been employed by physicists in improving the outcome of various optimization procedures[22]. One may be interested in the lowest energy states of some physical system, or in the shortest trajectory of a traveling salesman. In the first example, the energy, expressed in terms of the degrees of freedom of the system, provides a landscape of the type discussed above. It should be supplemented by some procedure which steps downhill. In this way one obtains what is known as *steepest*

descent procedure for the minimization of the energy. In a problem like the traveling salesman, one can also define a function that should be minimized. It is simply the length of the trajectory given a choice of the ordering of the towns to be visited.

There is, however, a familiar difficulty with such optimization procedures. If the function has many local minima, a steepest descent procedure will usually lead into one of these, rather than into one of the desired global optimal solutions. In physics, a familiar situation of this kind is the spin-glass, which will come up again in the following section. This is a system of mutually interacting magnetic moments which exert conflicting influences on each other; where interacting magnetic moments are as likely to try to align each other as to anti-align each other[23]. Such a system's energy is replete both with global minima as well as with local minima[24]. Using a standard procedure of steepest descent will be extremely ineffective for arriving at global minima of the energy. A similar situation is encountered in trying to optimize the traveling salesman's path length[22].

A method of dealing with such optimization difficulties, where landscapes exist, namely when at every step the process reduces some function which is bounded from below, has become known as *simulated annealing*[22]. Such a method is not only important for probing low energy states of physical systems with a multitude of local minima, but is of considerable import in many practical fields such as engineering, planning and economics. It consists of widening the choice of allowed steps according to a probability just like in Eq. 2.19. If x represents the set of relevant variables and Δx is an attempted step in the space of x which is associated with a change ΔE in the landscape function E, then simulated annealing allows Δx to be made with probability

$$\Pr(\Delta x) = \frac{\exp[-\beta \Delta E(x)]}{\exp[-\beta \Delta E(x)] + \exp[\beta \Delta E(x)]}. \tag{2.26}$$

The parameter β is a fictitious inverse temperature, which measures the level of noise introduced by the stochastic enlargement of the space of trajectories.

In the vocabulary of the preceding section we can expect that steps which increase the value of E by as much as β^{-1} will be likely, while much greater increases will be very improbable. In particular, as β becomes very large, i.e., with the suppression of the noise, the only steps that will be allowed are those which decrease E, and one is back to

deterministic steepest descent. For moderate values of β the stochastic optimization process will not settle into local minima, which are the analogues of meta-stable states, with barriers lower than β^{-1}. Every time the process approaches one of them it will linger in the neighborhood for a while. After a time that is inversely proportional to the exponential of the height of the barrier surrounding the local minimum divided by the temperature β^{-1}, the system will be likely to move into a deeper well. As time proceeds, it will tend to be found in the deeper minima. Yet, as long as the temperature remains high, it will not settle, just like Hoffman's popcorn kernel. In order to actually reach optimal states, the noise must be gradually decreased, and eventually it must be eliminated. The proper temporal schedule of the variation of the fictitious temperature is a matter of considerable sophistication[22,25]. There is, of course, no guarantee that at the end of such a schedule the system will be found at a strict global minimum, since the process is not deterministic. It turns out, however, in problems as varied as the traveling salesman[26], the optimization of micro-electronic chip component placing[22,27], to the investigation of low lying states in spin-glasses[25], that the improvement over traditional methods is very significant.

2.3.6 Frustration and diversity of attractors

In the context of models of neural networks, spin-glasses are often invoked. There are two basic reasons for the association.

- Spin-glasses have very many energy minima – this is a combination of *stability* (attractors) and *diversity* (plenitude). Such a combination is quite desirable in an associative, classifying memory.

- In spin-glasses, mutual influences between constituent magnetic moments are conflicting in tendency. This is an intrinsic feature of neural systems ever since Sherrington, whereby neurons interact synaptically via intense mixtures of excitatory and inhibitory synapses.

It is hard to resist quoting Sherrington's peerless prose on the matter

> The nerve nets are patterned networks of threads. The human brain is a vast example, offering immense numbers

> of determinate paths, and immense numbers of junctional points... These junctional points are often convergent points for lines from several directions. Arrived there signals convergent from several lines may coalesce and thus reinforce each other's exciting power.
>
> At such points too appears a process which instead of exciting, quells and precludes excitation. This inhibition... does not travel. It is evoked, however, by traveling signals not distinguishable from those which call forth excitement.... The two are relatively antagonistic[28].

We now proceed to describe an example which should give some initial feeling about the connection between conflicting mutual interactions and the appearance of a diversity of minima. For the non-physicist the example may appear initially *ad hoc*, but we recommend that it not be side-stepped.

Consider four variables S_i (to be referred to as spins, as they would be in the magnetic analog), which can take on the values ± 1, and which are placed at the four corners of a square, as in Figure 2.12. Three such instances are displayed. They differ in the signs on the lines connecting the spins. In (a) all four have + signs; in (b) there are three + signs and one − sign. In (c) there are two pluses and two minuses. These signs indicate the preference of relative orientational alignment. In other words, a positive (negative) sign indicates that two neighboring spins prefer to have the same (opposite) sign.[9] Suppose that the energy of a given configuration of spins (S_1, S_2, S_3, S_4) is given by the expression

$$E\{S\} = -\frac{1}{2}\sum_{\{i,j\}} J_{ij} S_i S_j. \tag{2.27}$$

where the sum is over all pairs which are connected by lines in the figure and $J_{ij} = +1$ for lines marked +, and -1 for lines marked −. Because of the minus sign in front of the sum, every pair of spins on an edge would lower the energy by having the product of its values equal the sign of the corresponding J_{ij}. For example, the pair 1–4 in Figure 2.12(b) has a negative coupling. The spins at its two ends will

[9]The reader may find it a very useful exercise to try and restate the above in neuronal language.

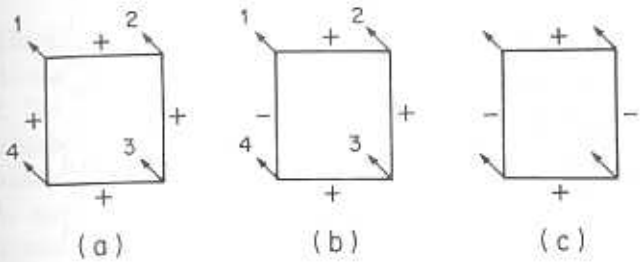


Figure 2.12: Three sets of four spins arranged on squares. The numbering is only for labelling purposes. (a) a ferromagnetic square; (b) a frustrated square – three aligning and one anti-aligning interactions; (c) a non-frustrated square – two aligning and two anti-aligning interactions.

lower the energy if they have opposite signs, since then $J_{14}S_1S_4 = 1 > 0$ and the contribution to the energy is -1.[10] Similarly, the two spins on the edge 2–3 of the same figure would prefer to have equal signs.

The reader can easily verify that in Figures 2.12(a) and (c) it is possible to arrange all four spins so that every single pair is happy, namely its $J_{ij}S_iS_j = 1 > 0$. For such an arrangement of spins, the total energy E will attain its lowest possible value, since each pair of spins contributes to its lowering. That value will be $E = -4$. The same cannot be said of the arrangement in Figure 2.12(b). There is no configuration of the four spins which can satisfy all four bonds. The reason is rather simple. The bond 1–4 would like its two spins to be of opposite signs, but all other bonds would like to align 1 with 2, 2 with 3 and 3 with 4. These three bonds would tend finally to align 1 and 4, leading to a contradiction. There is, therefore, no configuration which can give an energy as low as -4. Such a system is called *frustrated*[29]. It should be pointed out that it is not the mere presence of positive and negative connections which makes the system frustrated. This is made clear by contemplating Figure 2.12(c). If $S_1 = S_4$ and $S_2 = S_3 = -S_1$, all spins are happy and $E = -4$.

A system of four spins has sixteen possible states in all. In Table 2.1 we list eight of the sixteen states and give their energies in the three networks in Figure 2.12. Each of the other eight states is the same as one of the listed eight with all spins reversed. As such it has the

[10]Note that each term in the sum enters twice.

	(a)	(b)	(c)
+1 +1 +1 +1	−4	−2	0
+1 +1 +1 −1	0	−2	0
+1 +1 −1 +1	0	0	0
+1 −1 +1 +1	0	0	0
−1 +1 +1 +1	0	−2	0
+1 +1 −1 −1	0	−2	+4
+1 −1 +1 −1	4	+2	0
+1 −1 −1 +1	0	+2	−4

Table 2.1: Eight of the sixteen states for the three networks in Figure 2.12. The first column lists the states with the order of the spins corresponding to the numbering on the squares in the figure form left to right. The columns labelled (a)(b)(c) correspond to the different parts of Figure 2.12.

same energy. The striking fact in this table is that the two unfrustrated systems, (a) and (c), have each one state of lowest energy, $E = -4$. There are really two such states if we restore the reversed states. These are the ground states of these systems. In contrast, the frustrated system (b) never reaches an energy as low as -4, but has four states of lowest energy $E = -2$ (eight low energy states altogether).

- Frustration prevents the energy from becoming as low as in an unfrustrated system, but it creates *diversity* – a variety of ground states.

For our purposes here, the implication of frustrated systems is the following. Attractors in network dynamics will be found in certain circumstances to be equivalent to minima in a landscape such as that of a system of interacting spins. Pre-spin-glass physics had been mostly familiar with spin systems that have a very small number of energy minima, and those they have are all equivalent to a single one, via *symmetry transformations*. Such was the doubling of the single ground state in the two unfrustrated systems above. If networks are to serve for classifying memories, they must be able to store a diverse collection of memories which are significantly different. Since memories are associated with attractors and attractors with minima, frustration provides a way out of the poverty of usual spin systems and will allow a rather rich (sometimes too rich) structure of attractors. This will serve as a basis for much of the subsequent discussion.

Bibliography

[1] J.J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, *Proc. Natl. Acad. Sci. USA*,**81**, 3088(1984).

[2] S. Grossberg, Some networks that can learn, remember, and reproduce any number of complicated space-time patterns, I, *J. of Math. and Mechanics*, **19**, 544(1969) and II, *Studies in Applied Mathematics*, **XLIX**, 135(1970).

[3] S. Grossberg and J. Pepe, Spiking thresholds and overarousal effects in serial learning, *J. Stat. Phys.*, **3**, 95(1971).

[4] J.J. Hopfield and D.W. Tank, Computing with neural circuits: a model, *Science*, **233**, 625(1986).

[5] T.J. Sejnowski, Skeleton filters in the brain, in *Parallel Models of Associative Memory*, G.E. Hinton and J.A. Anderson, Eds. (Lawrence Erlbaum, Hillsdale, N.J., 1981)

[6] B. Widrow and M.E. Hoff, Adaptive switching circuits, *IRE WESCON Convention Record*, **4**, 4-96(1960); B. Widrow, G.F. Groner, M.J.C. Hu, F.W. Smith, D.F. Specht and L.R. Talbert, Practical applications for adoptive data-processing systems, *IRE WESCON Technical Papers* p.1, (1963).

[7] S. Amari, Learning patterns and pattern sequences by self-organizing nets of threshold elements, *IEEE Trans. Comput.*, **21**, 1197(1972).

[8] B.G. Cragg and H.N.V. Temperley, The organization of neurons: a cooperative analogy, *Electroenceph. Clin. Neurophysiol.*, **6**, 85(1954).

[9] W.A. Little, The existence of persistent states in the brain, *Math. Biosci.*, **19**, 101(1974).

[10] B. Katz, *Nerve, Muscle and Synapse* (McGraw-Hill, NY, 1966).

[11] G. Shaw and R. Vasudevan, Persistent states of neural networks and the nature of synaptic transmission, *Math. Biosci.*, **21**, 207(1974)

[12] P. Peretto and J.-J. Niez, Stochastic dynamics of neural networks, *IEEE Transactions: SMC* **16**, 73(1986)

[13] E. Caianiello, Outline of a theory of thought processes and thinking machines, *J. Theor. Biol.*, **2**, 204(1961).

[14] P. Peretto and J.J. Niez, Collective properties of neural networks, in *Disordered Systems and Biological Organization*,

E. Bienenstock, F. Fogelman Soulié and G. Weisbuch eds. (Springer-Verlag, Berlin, 1986).

[15] J.J. Hopfield, Neural networks and physical systems with emergent computational abilities, *Proc. Natl. Acad. Sci. USA*, **79**, 2554(1982).

[16] L. Standing, Learning 10,000 pictures, *Quarterly Journal of Experimental Psychology*, **25**, 207(1973).

[17] S. Sternberg, High speed scanning in human memory, *Science*, **153**, 652(1966).

[18] S. Sternberg, Memory scanning: Mental processes revealed by reaction-time experiments, *American Scientist*, **57**, 421(1969).

[19] R.E. Hoffman, Computer simulations of neural information processing and the schizophrenia-mania dichotomy, *Archives of General Psychiatry*, **44**, 178(1987).

[20] H. Atlan, *L'Organization Biologique et la Theorie de l'Information* (Hermann, Paris, 1972).

[21] D.W. Tank and J.J. Hopfield, Collective computation in neuronlike circuits, *Sci. Am.*, December, 1987.

[22] S. Kirkpatrick, C.D. Gelatt Jr. and M.P. Vecchi, Optimization by simulated annealing, *Science*, **220**, 671(1983).

[23] S.F. Edwards and P.W. Anderson, Theory of spin glasses, *J. Phys.*, **F5**, 965(1975).

[24] S.F. Edwards and F. Tanaka, The ground state of a spin glass, *J. Phys.*, **F10**, 2471(1980).

[25] See e.g., G.S. Grest, C.M. Soukoulis, K. Levin and R.E. Randelman, Monte Carlo and mean field slow cooling simulations for spin-glasses: Relation to NP-completeness, in *Heidelberg Colloquium on Glassy Dynamics*, J.J. van Hemmen and I. Morgenstern eds. (Springer-Verlag, Berlin, 1987).

[26] S. Kirkpatrick and G. Toulouse, Configuration space analysis of travelling salesman problems, *J. Physique*, **46**, 1277(1985).

[27] See e.g., I. Guyon, L. Personnaz, P. Siarry and G. Dreyfus, Engineering applications of spin glass concepts, in *Heidelberg Colloquium on Glassy Dynamics*, J.J. van Hemmen and I. Morgenstern eds. (Springer-Verlag, Berlin, 1987).

[28] C.S. Sherrington, *The Brain and its Mechanism* (Cambridge University Press, London,1933).

[29] G. Toulouse, Theory of the frustration effect in spin glasses I, *Comm. on Physics* **2**, 115(1977).

3

General Ideas Concerning Dynamics

3.1 The Stochastic Process, Ergodicity and Beyond

3.1.1 Stochastic equation and apparent ergodicity

In Section 1.4.1 we have seen that the dynamics of an ANN is a march on the vertices of a hypercube in a space of N dimensions, where N is the number of neurons in the network. Every one of the vertices of the cube, Figure 1.13, is a *bona fide network state*. Let us first consider the case of asynchronous dynamics, Section 2.2.3, with a single neuron changing its neural state at every time step. In this case the network steps from one vertex to any of its N nearest neighbors. The question of *ergodicity*, and correspondingly of the ability of the system to perform as an associative memory, is related to the dependence of trajectories on their initial states. In other words, the initial states of the network are those states which are strongly influenced by an external stimulus. If the network were to enter a similar dynamical trajectory for every stimulus, no classification would be achieved. This will be the sense in which we will employ the term *ergodic* behavior. In terms of our landscape pictures, Figure 2.10 and 2.11, this would be the case of a single valley to which all flows. Alternatively, if trajectories in the space of network states depend strongly on the initial states, and correspondingly on the incoming stimuli, then the network can recall selectively, and have a variety of items in memory.

In Section 2.2.3 we have seen that in the typical noisy case there

is always a finite probability that the network move from a vertex to *any* one of its neighbors. As a result, one would conclude, intuitively, that no matter what initial vertex the network has started in, it will eventually visit all other vertices. Such *ergodicity* will be harmful, of course, only if the spread takes place in a short time. Equivalently, if the dynamics restricts the network for long times to a narrow region of the space of network states, a region which depends on the initial condition, then the network may still be a useful selective memory. Before dealing with these fine issues we shall first study the idealized cases.

The mathematical argument which is presented below can be summarized as follows:

- In a generic, noisy case any system will be ergodic — after a long enough time a system will sample its possible states in a way which is independent of its initial state.

If this were indeed generally true, then some of the most exciting phenomena around us would have been ruled out — in particular there would be no *phase transitions* and *symmetry breaking*. It should perhaps be added for the sake of the non-initiated, that symmetry breaking is tantamount to the appearance of form and structure. How nature escapes this bind, how it breaks ergodicity, is the subject of the rest of this chapter. The particular way in which this same escape route may provide *emergent* properties of neural networks is the subject of the rest of this book. It seems to be the only concrete idea that has been put forth with sufficient depth to legitimately aspire to capture the rich intuitions associated with *emergence*, whether or not emergence is necessary for a description of mental functions.

There are two possible ways to escape the predicament of ergodicity

1. If the network is noiseless, which is rather exceptional, only special moves are allowed from each vertex and the network is non-ergodic.

2. If noise is present there must be a real *cooperative phenomenon*.

A first example of what it takes to have a *bona fide* cooperative behavior will be presented in Section 3.2.1, below.

Stochastic equation in the asynchronous case

Suppose that time is discretized as in Section 2.2.3. The network can start at time $t = 0$ and proceed at intervals of δt, Eq. 2.23, where at the end of each sub-interval a single neuron, at most, can change its *neural state*. We denote the probability that the network be in *network state* J at time $n\delta t$ by $\rho_J(n)$. The probability that network which is in state J at time $t = 0$ makes a transition to state I within the interval δt, will be denoted by $W(I \mid J)$. Then, clearly, one has

$$\rho_I(n+1) = \sum_J W(I \mid J)\rho_J(n). \tag{3.1}$$

Sums over J go over all 2^N network states. In matrix notation it reads:

$$\rho(n+1) = \mathcal{W}\rho(n) \tag{3.2}$$

where $\rho(n)$ is a vector of 2^N components and $\mathcal{W}$ is a $2^N \times 2^N$ matrix.

Because the elements of both ρ and $\mathcal{W}$ are probabilities, they are normalized when summed on all possible outcomes, namely

$$\sum_I \rho_I(n) = 1, \tag{3.3}$$

$$\sum_I W(I \mid J) = 1. \tag{3.4}$$

When Eq. 3.4 is inserted in Eq. 3.1 the latter can be rewritten in the form:

$$\begin{aligned}
\rho_I(n+1) &= \sum_J W(I \mid J)\rho_J(n) \\
&= \sum_{J \neq I} W(I \mid J)\rho_J(n) + W(I \mid I)\rho_I(n) \\
&= \sum_{J \neq I} W(I \mid J)\rho_J(n) + \left(1 - \sum_{J \neq I} W(J \mid I)\right)\rho_I(n) \\
&= \rho_I(n) + \sum_{J \neq I} [(W(I \mid J)\rho_J(n) - W(J \mid I)\rho_I(n)].
\end{aligned} \tag{3.5}$$

The last equality can be easily interpreted. It has the form of a *master equation*[1]. What it says is that the change in the probability

that the system be in state I after a short time interval is composed of two parts: an increase due to transitions from other states into that state and a decrease due to transitions out of this state into other states.

Since the matrix equation 3.2 is a linear transformation, it can be iterated to obtain the distribution function ρ at time $n\delta t$ as a function of the initial distribution. It reads, in matrix form,

$$\rho(n) = \mathcal{W}^n \rho(0). \tag{3.6}$$

What one is after is the probability distribution of states after very long time, i.e., as $n \to \infty$. If ρ is decomposed as a linear combination of eigen-vectors of $\mathcal{W}$, each component will be multiplied by the corresponding eigen-value raised to the n-th power. See e.g., Eq. 3.9, below. Consequently, in the limit, it will be the terms with the largest eigen-value which will dominate the behavior.

Since the elements of the matrix $\mathcal{W}$ are probabilities the highest eigen-value is equal to one and there is a corresponding left eigen-vector whose elements all equal one. This is shown in Appendix 3.6.1, where some of the definitions are included as well. As probabilities, the element of $\mathcal{W}$ are non-negative. Moreover, generically (finite noise), in every column of the matrix there are at least $N+1$ non-vanishing elements corresponding to the N neighboring vertices to which a single-neuron update can transfer the network. For synchronous dynamics, all matrix elements of $\mathcal{W}$ are non-zero. But even for sequential dynamics, with finite noise, the matrix is technically speaking *irreducible*. As a consequence, the largest eigen-value of $\mathcal{W}$, which dominates the long-time behavior of the solutions of Eq. 3.1, is unique, i.e., non-degenerate[2,3].

The matrix $\mathcal{W}$ has a *spectral expansion* in terms of its left and right eigen-vectors, $\mathbf{V}^L$ and $\mathbf{V}^R$, of the following form:

$$W(I \mid J) = \sum_{i=1}^{2^N} \lambda_i V^L_{\lambda_i}(J) V^R_{\lambda_i}(I), \tag{3.7}$$

which is generally not symmetric. Left and right eigen-vectors are orthogonal if they belong to different eigen-values, namely

$$\sum_I V^L_{\lambda_i}(I) V^R_{\lambda_j}(I) = \delta_{\lambda_i \lambda_j}. \tag{3.8}$$

Therefore they provide an expansion of the identity.

When Eq. (3.7) is substituted in Eq. 3.6 and the latter is written in terms of components, one has:

$$\rho_I(n) = \sum_J \sum_{i=1}^{2^N} \lambda_i^n V^L_{\lambda_i}(J) V^R_{\lambda_i}(I) \rho_J(0). \tag{3.9}$$

If one keeps in mind the normalization of ρ at all times, Eq. 3.3, then for large n the sum on i, on the right hand side, is dominated by the high powers of the highest eigen-value and hence

$$\begin{aligned} \rho_I(n) &\overset{n\to\infty}{\longrightarrow} \sum_J V^R_{\lambda=1}(I) \rho_J(0) \\ &= V^R_{\lambda=1}(I). \end{aligned} \tag{3.10}$$

This implies that after a long enough time the system has effaced its history and for **every** initial distribution it reaches the same asymptotic distribution in network state space, the one that corresponds to the right eigen-vector of the *unique*, largest eigen-value.

Ergodicity in the synchronous case

In the synchronous case, all N neurons are candidates for updating in every time step[4,5]. (See e.g., Section 2.2.2). In the presence of noise there is non-zero probability for the system to make a transition from any vertex to any of the 2^N other vertices. Eq. 3.1 again describes the dynamical development of the probability distribution. The only difference is that the matrix $\mathcal{W}$ has 2^N non-vanishing matrix elements in each column, rather than merely $N+1$ as in the asynchronous case. The entire argument proceeds as before, with a unique (*non-degenerate*), maximal eigen-value, equal to unity, dominating the dynamics. The noisy network would seem to be ergodic again.

3.1.2 Two ways of evading ergodicity

Remaining on the level of the general arguments of the previous section one immediately concludes that short of having a degenerate largest eigen-value for the transition matrix $\mathcal{W}$, there is no escaping ergodicity. This should appear as a paradox, since many physical systems are non-ergodic even in the presence of noise. There must be some ruse by

which this eigen-value could be made degenerate. In that case the situation would be quite different[5].

Suppose, for the sake of the argument, that the maximal eigen-value, which remains equal to one, is doubly degenerate, namely, there are two independent left (and right) eigen-vectors with that eigen-value. Let us denote them, respectively, by $V_1^L, V_2^L, V_1^R, V_2^R$. At large times, the equation for ρ, Eq. 3.9, will read

$$\begin{aligned} \rho_I(n) &\stackrel{n\to\infty}{\longrightarrow} \sum_{i=1}^{2}\sum_J V_i^L(J)V_i^R(I)\rho_J(0) \\ &= \sum_{i=1}^{2}[\mathbf{V}_i^L\cdot\rho(0)]\mathbf{V}_i^R, \end{aligned} \tag{3.11}$$

instead of Eq. 3.10. The first factor in the sum on the right hand side of this equation is the projection, in the space of network states, of the initial distribution on each of the left eigen-vectors of eigen-value unity. It has been expressed as a scalar product. The extension to a higher degeneracy of the largest eigen-value should be self-evident.

Since $\rho(0)$ still appears in this equation, such a dynamical process will lead to different asymptotic (long times) trajectories depending on the initial situation. On the one hand it will be a classifier, in the sense that initial conditions (initial network states) belonging to different groups of states will have different asymptotic dynamical development. On the other hand, each such group will contain many initial states and therefore one may be able to capture associativity. The question then boils down then to finding the conditions which may violate the theorem about the uniqueness of the largest eigen-value of the transition matrix $\mathcal{W}$.

The noiseless way

The 'trivial' way of evading the conditions of the uniqueness theorem is to have a situation in which the transition matrix is *reducible*[2]. A sufficient condition for reducibility is to have a column with a one in the diagonal and zeroes everywhere else. Each such column represents a vertex of our state hypercube **out of** which no transitions are allowed into any of the other network states. Clearly, this can happen only in

the noiseless, zero temperature, limit. For example, if the transition matrix has the form

$$\mathcal{W} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} \tag{3.12}$$

then it has two left eigen-vectors with unit eigen-value. They are:

$$\begin{aligned} \mathbf{V}_1^L &= (1,1,1,1) \\ \mathbf{V}_2^L &= (1,\tfrac{1}{2},\tfrac{1}{2},0) \end{aligned}$$

as can be verified by inspection. But what system does such a matrix correspond to?

It is a network of *two* formal neurons. The neural states will be $+1$ and -1 and the threshold will be taken to be zero. The coupling coefficients (synaptic efficacies) are given by the matrix

$$J_{ij} = K\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

which implies that each of the two neurons synapses on the other one with excitatory efficacy $+K$. Let us choose asynchronous dynamics first.

The PSP on the neurons is

$$\begin{aligned} h_1(t+\delta t) &= KS_2(t) \\ h_2(t+\delta t) &= KS_1(t) \end{aligned}$$

as in Eq. 2.20, keeping in mind Eq. 2.23. In that case the noiseless dynamics would give for the new states of the two neurons, Eq. 2.12,

$$\begin{aligned} S_1(t+1) &= \text{sign}[KS_2(t)] = S_2(t) \\ S_2(t+1) &= \text{sign}[KS_1(t)] = S_1(t). \end{aligned}$$

Clearly, if the network is in either state $(+1,+1)$ or $(-1,-1)$, which we will label 1 and 4, respectively, then both neurons will remain in these neural states. Hence one can write:

$$W(I\mid 1) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \qquad W(I\mid 4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{3.13}$$

On the other hand, if the network is in the state $(+1,-1)$, which we call 2, then in asynchronous dynamics the network will move with equal probability either into state 1 or 4, depending only of which neuron is updated first. The probability is equal. Similarly, for the last possible state $(-1,+1)$, which is number 3. Now, if the rows and columns of the matrix $\mathcal{W}$, Eq. 3.12, are labeled by the state numbers assigned above, then one obtains the transition matrix of Eq. 3.12.

If this network is updated synchronously, then columns 1 and 4 are the same, but state 2 always skips into 3 and vice versa. We have a two-cycle. In this case the matrix $\mathcal{W}$ will have the form:

$$\mathcal{W} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

In the asynchronous case we have two eigen-values equal to 1 and two equal to 0, while in the synchronous case there are three eigen-values which are 1 and one equal to -1.

The noisy way

In both cases discussed above ergodicity is broken and state trajectories will depend strongly on the initial conditions. But, it should be observed that as soon as noise is introduced there will be no non-zero matrix elements. For low levels of noise, these additional matrix elements will be small, but that is not sufficient relief. To find necessary conditions for breaking ergodicity at low noise levels is a rather involved matter. Instead we will restrict ourselves in the present context to pointing out that in situations in which one may be tempted to invoke *cooperativity* as an explanation of broken ergodicity, ergodicity prevails. On the other hand, we will show that the type of network connections which have been introduced in the physics models are sufficient to ensure that, for low enough noise levels, the special property of *broken ergodicity*, which we would like to associate with *cooperativity* and *emergence*, holds.

Basically, the alternative way to breaking ergodicity is the *thermodynamic limit*. What it implies is that it may be the case that as the size of the matrix $\mathcal{W}$, and correspondingly of the system itself, increases, the gap between the largest eigen-value and the next largest

one shrinks. Then, in the limit of an infinitely large matrix – a thermodynamic system – there would be asymptotic degeneracy. As the gap shrinks, the time, or number of steps in Eq. 3.9, it takes for the one largest eigen-value to dominate the dynamics and efface memory of the initial conditions becomes larger and larger.

- Ergodicity will strictly break then only in a limiting sense.

This ideal sense we associate with *emergence*. A necessary condition for it is the asymptotic degeneracy of the eigen-values.

3.2 Cooperativity as an Emergent Property in Magnetic Analog

3.2.1 Ising model for a magnet – spin, field and interaction

One of the most fruitful models in modern physics has been an extremely simplified description of magnetism — the Ising model[6]. To obtain a vivid perspective of the extent of the simplification within the context of the theory of magnetic systems one should refer to such comprehensive books as Mattis' monograph[7]. Despite the simplifications and perhaps due to them the model has provided deep insights into the properties of magnetic systems, as complicated as Manganese Fluoride (MnF_2), or Chromium Bromide ($CrBr_3$), without ever mentioning that it is only the Manganese and the Chromium that are magnetic, or that all the participating elements are composed of different numbers of nucleons and electrons.

But it has done much more. It has opened the way for understanding the properties of a whole universe of systems of large numbers of strongly interacting elements. From liquids and solids through the most fundamental theories of elementary particles and fields. It has been the birthplace and the testing ground for a treasure of new concepts in essentially all fields of physics. Such fundamental ideas as *symmetry breaking*, *cooperative phenomena*, *order parameters*, *disorder parameters*, *critical exponents*, *symmetry restoration* etc., have had their first explicit, precise articulation in the framework of this apparently simple, naive model.

We will elaborate on the Ising model in this chapter for two main reasons:

- To make our use of the catch phrase *cooperative phenomena* as well-defined and explicit as possible. In particular, it will provide us with a semi-transparent context in which to observe mechanisms which bring about cooperative behavior, while exhibiting situations which might have been conceived as such and are not.
- To give a flavor of the reality of a physical system, whose extensions are the basis for physicists' activity in modeling neural networks.

The model is defined as a collection of N elementary magnetic moments of a very simple type: they can point only either up or down. These moments will be called *spins* and will be denoted by S_i, where i enumerates the particular member of the collection. They are allowed to take on only two values, corresponding to the two discrete orientations – $S_i = +1$ or -1. This magnetic system will have 2^N possible system states.

Another important concept is the *magnetic field*, which may spread throughout space. It applies a torque to every magnetic moment, trying to orient it in its own direction. We will consider two types of magnetic fields:

- external, h_i^e, independent of the dynamics of the system;
- internal, h_i^i, induced by the system's spins themselves.

In the neural analogy, the first are like the thresholds while the last are like the PSP's. If, at the site of spin i the magnetic field is h_i^e, then the total energy of the entire system will be:

$$E\{S\} = -\sum_{i=1}^{N} h_i^e S_i. \tag{3.14}$$

The curly brackets accompanying E indicate that the right hand side is the energy of the system in that particular system state. The fields h_i^e are considered to be fixed and external to the system, which implies that in order to lower the energy, as every physical system would like to do, every spin must take on the *sign* of the magnetic field which

acts on it. Only then will every contribution to the expression for the energy, Eq. 3.14 be negative.

The spins can be seen to influence each other when one takes into account the fact that magnetic moments are themselves a source of *magnetic field*, and exert a torque on each other. This influence has a feed-back effect because if the field produced changes the orientation of some of the spins then this change modifies the field and acts back on the spins that had given rise to it. To be more specific, the *internal* field h_i^i, generated at the site of spin i by the other spins is given by:

$$h_i^i = \sum_{j,j\neq i}^{N} J_{ij} S_j. \tag{3.15}$$

The coefficients J_{ij} are referred to technically as the *exchange interaction*. They may depend on the type of underlying mechanism which brings about the mutual forces, as well as on the distance between the spins i and j. Note that the term of self-interaction, $j = i$, is omitted. Moreover, since each term in the sum corresponds to a mutual force (or torque) between two spins, Newton's third law imposes the symmetry

$$J_{ij} = J_{ji}.$$

To go from the internal fields to the total energy of the spin system one has to keep in mind that in summing the effect of these fields on all spins each pair energy is counted twice, once when the first neuron is contributing to the internal field and the second is acted upon and once when the second spin is present in the field and the first is acted upon. This induces a factor of $\frac{1}{2}$ relative to Eq. 3.14. Hence, one has:

$$E\{S\} = -\frac{1}{2}\sum_{ij,j\neq i}^{N} J_{ij} S_i S_j. \tag{3.16}$$

The total of the two energies Eqs. 3.14 and 3.16 gives the energy of any configuration of spins of the Ising model in an external magnetic field. The total magnetic field acting on any of the spins can be read off from the total energy as *minus* the coefficient of the particular S_i (See e.g., Eq. 3.14). It is:

$$h_i^t = \sum_{j,j\neq i}^{N} J_{ij} S_j + h_i^e. \tag{3.17}$$

3.2.2 Dynamics and equilibrium properties

The Ising model has been so simplified that it has lost its dynamical rules. This has never bothered physicists, because as far as the equilibrium properties of a system are concerned they are completely determined by the energy. This deep result, due to Gibbs, states that

- if we know the energy, E, for every state of a system, then the properties of the system, *in equilibrium* at temperature T, can be computed **as if** we had an ensemble of identical systems and the probability for finding one of them in any of the possible states is proportional to $\exp(-E/kT)$, where k is Boltzmann's constant.

This is the celebrated *canonical ensemble*. Given the probability distribution for the states, the mean of every observable quantity is determined. How this is related to our dynamical interests will be clarified below.

If $O\{S\}$ is any property of the system, which has a value for every one of the 2^N system states, then the equilibrium value of this property is:

$$\langle O \rangle = \frac{\sum_S O\{S\} \exp(-E\{S\}/kT)}{\sum_S \exp(-E\{S\})} \tag{3.18}$$

where both sums are over all possible states of the system. Examples of such properties O may be the *magnetization* of the system $M\{S\}$ which is the sum over the values of all the spins in a given state, and its value gives the balance between the number of spins pointing 'up' and 'down'. The expression for it is:

$$M\{S\} = \sum_{i=1}^{N} S_i. \tag{3.19}$$

This quantity is a close relative of the overlap defined in Section 1.4.1, since it is in fact the overlap of state S with a state of N ones. It is, therefore, related to the quality of retrieval from memory. Another property could be the correlation between the values of two spins i and j, such as:

$$g_{ij}\{S\} = S_i S_j, \tag{3.20}$$

whose value gives the propensity of the spins at sites i and j to be aligned.

The relevance of Eq. 3.18 to our subject is due to the fact that if one allows the system to develop dynamically for a long enough time, at fixed temperature T, then the time average of any observable quantity is given by an *ensemble average* over the *canonical distribution*. In other words,

- **the time average equals an average over many systems which appear in different states with a relative frequency proportional to** $\exp(-E/kT)$.

For physical systems, this is a rather general result. But if one inspects what in the dynamical, stochastic process leads eventually to this equivalence, one finds a rather simple sufficient condition — *detailed balance*[8]. This is a condition on the transition probabilities from state to state in the resulting Markov chain – our master equation 3.1. It is formulated in terms of the matrix elements of $\mathcal{W}$ of Section 3.1.1 and it reads:

- If the transition probabilities $W(I \mid J)$ satisfy

$$W(I \mid J)F(J) = W(J \mid I)F(I) \tag{3.21}$$

where F depends on a single network state, then as

$$n \to \infty \qquad \rho_I(n) \to F(I). \tag{3.22}$$

A few comments are in place at this point:

- If Eq. 3.21 is satisfied it follows that

$$\rho_I(n) = F(I)$$

is a time independent solution of Eq. 3.1. It is an *equilibrium* situation.

- If the transition probabilities satisfy Eq. 3.21 with

$$F(I) = \exp(-E/kT),$$

then the process will lead the distribution function ρ to the equilibrium distribution of Gibbs' canonical ensemble.

- A dynamical system, operating under *bona fide* equations of motion and thermal noise, necessarily satisfies *detailed balance*.

- To reach the same asymptotic equilibrium distribution of states (and hence the same averages of observables) one can have any dynamical process, as long as it satisfies *detailed balance* with the same function F.

This has led in 1953 and 1963 to two seminal papers. The first[8] has established the Monte-Carlo technique for the approximate evaluation of means of physical observables for systems with very large numbers of interacting particles. In this technique one generates a large number of states, of the particular system under study, by a stochastic process which is chosen to obey *detailed balance* with $F = \exp(-E/kT)$, where E is the actual energy of the system. The process has nothing to do with the system's dynamics. It was originally applied to the computation of the *equation of state* of a liquid. This technique has probably been the predominant single technique in essentially all domains of physics in the last twenty years. See e.g., refs. [9,10,11].

The second[12] has done somewhat the reverse. Observing that the Ising model, which is 'the first, and most successful...[model] in an attempt to explain the ferromagnetic phase transition,' has not been so 'in dealing with systems which undergo large-scale changes with time'. What Glauber proposed is to endow the Ising model with a stochastic dynamics such that on the one hand it reflect the system's contact with a *heat bath* at a fixed temperature T, while on the other it remain faithful to the Ising energy via the detailed balance relation. As we shall see, this dynamical process is identical to our noisy neural dynamics, Eq. 2.19.

Historically, this suggestion had two independent effects. The first has been the introduction of an additional Monte-Carlo algorithm – the *heat-bath*, out of the large variety of possible ones. This does not seem to have been an intended message of Glauber's paper. It differs from the Metropolis method[8], its main competitor in popularity, in that the single spin transition probability is independent of the preceding state of that spin. The second effect has been a simple plausible stochastic dynamical process which ensures a temporal relaxation to an equilibrium governed by a prescribed energy and a given temperature. This is the trend that will be followed in this section.

The dynamics of our Ising model will be defined as follows: The system is assumed to be at temperature T and it moves from state to state by changing the spin-state of one spin at a time. The probability,

$\Pr(S_i)$, of spin i taking on the value S_i, when its turn comes up, is determined by the **total** field h_i, Eq. 3.17, acting on it. It reads:

$$\Pr(S_i) = \frac{\exp(h_i S_i/kT)}{\exp(h_i/kT) + \exp(-h_i/kT)}. \tag{3.23}$$

This expression should be compared with the noisy neural dynamics of Section 2.1.3. The similarity is, of course, not accidental. The probability is independent of the state of the neuron concerned and depends only on the states of all the neurons which interact with it and on the external field. The functional form of Pr is given by Figure 2.3. Note that we have not prescribed the order of the updating of the spins, not even the level of synchronicity.

For the Ising model, we will prescribe asynchronous dynamical updating. In this case, two consecutive states can differ by at most a single spin flip. We can now verify that detailed balance is indeed obeyed by the dynamics and the equilibrium distribution is the canonical ensemble corresponding to the Ising energy of Section 3.2.1. Note first that the probability for spin i to go into the state S_i is independent of the previous state of that spin. It depends only on the final state of that spin and on the states of other spins, who create the local field. Therefore, the probability for the system to go from a *configuration* (system, or network state) J, in which spin i is in state S, to configuration I, in which that spin is in state $-S$, is $\Pr(-S)/N$ and from a state in which the same spin goes from $-S$ to S is $\Pr(S)/N$. The denominator expresses the a priori probability for picking the particular spin. Hence:

$$\frac{W(I \mid J)}{W(J \mid I)} = \frac{\Pr(-S_i)}{\Pr(S_i)} = \frac{\exp(-h_i S_i/kT)}{\exp(h_i S_i/kT)} = \exp(-2h_i S_i/kT). \tag{3.24}$$

Note that it was essential to have h_i independent of S_i in order to ensure detailed balance. As was pointed out in the discussion at the end of Section 3.2.1 the total field on spin i is just the coefficient of S_i in the total energy. Therefore, the expression in the exponent on the right hand side of the last equation is just the difference in energy between the two configurations I and J, having the same spins at all sites except i. At i, the spin is in state $-S_i$ in configuration I and

in S_i in configuration J. One can write, using the relations Eqs. 3.17 and 3.16,

$$W(I \mid J)\exp[-E(J)/kT] = W(J \mid I)\exp[-E(I)/kT] \tag{3.25}$$

which is the statement of detailed balance with a Gibbsian F. Hence, the distribution of configurations will relax to the equilibrium distribution

$$\rho_I = C \cdot \exp[-E(I)/kT]. \tag{3.26}$$

3.2.3 Noiseless, short range ferromagnet

Attractors in noiseless dynamics

Let us first consider the Ising spin system in the absence of noise, i.e., T=0. The transition probabilities, as prescribed by Glauber's dynamics Eq. 3.23, are

1. If the local field $h_i \neq 0$ then $S_i(t+\delta t) = \text{sign}(h_i)$ with probability one.
2. If $h_i = 0$ then $S_i(t+\delta t) = \pm 1$ each with probability $\frac{1}{2}$.

In both cases, when a spin changes its state $h_i S_i \geq 0$ in the new state. Consequently, according to Eq. 3.16, if the spin at i *has* changed into S_i, the change in energy is

$$\Delta E = -2h_i S_i \leq 0.$$

The energy is reduced if the field was non-zero, or remains unchanged if the field at the changing site happened to vanish.

A very important consequence is:

- In the absence of noise (T=0), fixed points of this dynamical process are at configurations in which each spin-flip increases the energy. Such configurations will remain unchanged with probability one.
- Conversely, all local minima of the energy are fixed points of the dynamics. In particular, so are the configurations of lowest energy – the *ground states*.

This state of affairs is easily visualized on a landscape. If motions are restricted to go only downhill, then the bottom of every valley is a fixed point of the dynamics.

Ferromagnetic interactions

So far, we have kept the interactions, the J_{ij}'s, and the external fields quite general. Let us now simplify further by choosing the external field to vanish, so that any special behavior that the system may exhibit is ascribable to its internal workings, rather than to external influences. We shall also choose all $J_{ij} \geq 0$. With these choices, all pairs of spins which are connected by a non-zero J_{ij} contribute to lower the energy if they align, i.e., if they take on the same sign. This is due to the negative sign in the definition of the energy, Eq. 3.16. This tendency toward alignment is essential in making the phenomenon of *ferromagnetism* possible. This phenomenon manifests itself in the appearance of equilibrium states which have a non-zero magnetization, a non-vanishing balance of positive and negative spins.

It should perhaps be pointed out that if the system behaves ergodically, then the time average of a quantity like the magnetization vanishes. This is because for every configuration with a non-zero magnetization there is a configuration, with all spins reversed, that has the same energy, the same probability and a magnetization of opposite sign. This feature is an intrinsic symmetry of the system.

Suppose that the collection of spins has no islands. In other words, if a non-zero J_{ij} is pictured as a line connecting the spins i and j, then no partial group of spins can be separated from the system without severing some of these lines. (In the neural analogue it implies that no portion of the network can be removed without breaking some synapses.) See e.g., Figure 3.1. It is then easy to see that the ground states of the system – the states of lowest energy – have all their spins aligned. Any non-aligned pair of spins will give a positive contribution to the energy, even if they are not connected directly. This can be seen by contemplating a chain like A to Z in Figure 3.1. If these two spin are of opposite signs, then somewhere along the line connecting them there must be a pair, directly connected, with two misaligned spins.

There are two such ground states, one with all spins +1 and one with all spins −1. These states have *ferromagnetic ordering*. The fact that these are the lowest energy states, and as such are attractors of

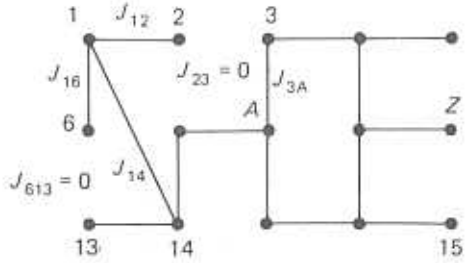


Figure 3.1: A spin system with coupled and uncoupled spins. Spins which are not connected by line segments imply that the corresponding $J_{ij} = 0$.

the noiseless dynamics, is a result of the mutual influences of the spins. In this sense the ferromagnetic phenomenon may be perceived as a *cooperative phenomenon*. It undergoes a gradient flow, with the energy as a *Lyapunov function*, which is reduced at every step and is bounded from below. The system must reach an attractor. In fact, it will reach an attractor even if the number of spins is finite.

Yet this is not the sense in which we will use the term *cooperativity*. For a physical system, and *a fortiori* for a neural network, operation without noise is not very realistic. The question then naturally arises as to whether a small amount of noise will have a small effect. This is often **not** the case, and unless the mutual interactions fall into special classes, any amount of noise, however small, will make the system *ergodic*. In particular, the attractors of a finite system will always be undone by noise, and all finite spin systems become ergodic in the presence of noise.[1] To make the situation even more transparent we will simplify our system further.

Ferromagnetic linear chain with nearest neighbor interactions

The magnetic Ising spins are placed on a linear chain. Each spin interacts only with its two neighbors, the one on its right and one on its left, as is indicated in Figure 3.2. All interactions are equal in magnitude and positive. The linear chain is placed on a ring in order to avoid complications of boundary conditions. It is a ferromagnetic model, with

[1] This fact should serve as a warning to various cellular automata networks, which perform magic in the absence of noise, but unless further arguments are brought forward, have a doubtful fate in the presence of noise.

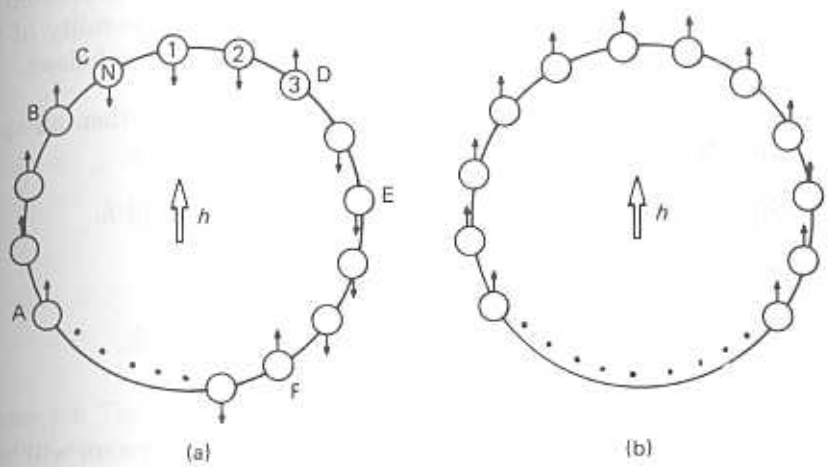


Figure 3.2: A nearest neighbor ferromagnetic Ising chain with small external field h. The arrows represent sample spin configurations. (a) is discussed in the text; (b) is an ordered ferromagnetic arrangement.

no islands. Yet, it is not a ferromagnet, in the sense that the slightest amount of noise will destroy any alignment of the spins.[2] The energy, the Lyapunov function, for this system is

$$E = -J \sum_{i=1}^{N} S_i S_{i+1}. \tag{3.27}$$

Note that the $\frac{1}{2}$ of Eq. 3.16 is absent. This is because here each pair is summed over once only. There is a term $S_N S_{N+1}$ in Eq. 3.27. It represents the closing of the ring, namely $S_N S_1$.

It is clear that the lowest energy states are the two ferromagnetic states, with all spins aligned either up or down. Both have energy $-NJ$. Any pair of misaligned spins adds energy $2J$. The magnetic field at site i is

$$h_i = J(S_{i-1} + S_{i+1}). \tag{3.28}$$

In the absence of noise, if the two spins are aligned they will align the spin between them. There is a minor delicate point when the two spins $i-1$ and $i+1$ are of opposite signs, then the local field vanishes and

[2] This system is, of course, also a network of cellular automata.

the spin i will flip half of the times. This is artificially prevented by adding a tiny external field, h, without limiting the generality of the ensuing argument. Now, the dynamical process will be as follows:

- If we start updating the spins at A in Figure 3.2, then all spins until B will remain 'up', because of their neighbors.
- Spin of type B will remain 'up', because of the field h.
- Spin of type C will flip 'up', because of the field.
- Spins between C and D follow C in a domino effect.
- All spins will follow suit.

From the overwhelming majority of initial states the system will very rapidly relax into the 'up' ferromagnetic state. The time unit will be chosen as the time necessary to update all N spins, since the system operates in some parallel fashion.

An exception is a situation like that of the spins between E and F in Figure 3.2. If most spins point down and there are individual spins pointing up, then if those are picked for updating they will flip down, and the system may drift to the 'down' ferromagnetic state. But this will happen rarely. If the updating starts at E then this spin and the next one will stay down, but the next one will flip up and the avalanche will start. The reason 'up' is preferred is due to the external field, of course. Some time courses are plotted in Figure 3.3. The curves describe the development of the average *magnetization* (overlap) per spin, $m(t)$, of the instantaneous state with the 'up' aligned state, i.e.,

$$m(t) = \frac{1}{N}\sum_{i=1}^{N} S_i(t). \tag{3.29}$$

When $m(t)$=1, every spin is in the 'up' state, and when $m(t)$=−1 they are all in the 'down' state. As one can see from Figure 3.3, the arrival at these attractors is rather rapid. The initial conditions are represented by the value of $m(t$=0). This is a clear case of non-ergodicity. For most initial configurations the system drifts into the 'up' attractor (because of the presence of h), and for the rest it drifts into the 'down' one. Both fully aligned states are fixed points of the dynamics, and there are no other fixed points.

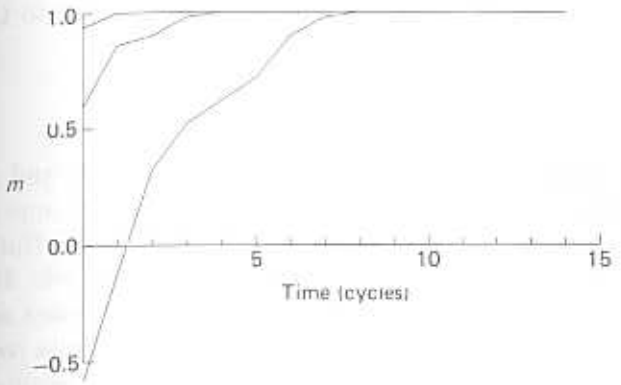


Figure 3.3: The time development of the magnetization per spin (overlap with the 'up' state) in the noiseless Ising chain. In each time unit on the abcissa all spins in the system are updated once, sequentially.

The return of ergodicity

The crucial point concerns the behavior of this same system upon the introduction of noise. It is old wisdom in physics that

- one-dimensional order with short-range interactions cannot persist[13] at any finite temperature.

The two attractors in our one-dimensional chain are clearly ordered states of the system. Noise must therefore destroy them. We will now proceed to identify the main ingredients which destroy the apparent cooperativity. In the process we will find some necessary conditions for the persistence of attractors in the face of noise.

Suppose that the noise is low, yet larger than the small external field we have put on, i.e., $T \gg h$. Suppose that the system finds itself in one of its ground states, the full 'up' state. What will be its future course? The probability for flipping one of the spins away from the fully aligned state is, according to the Glauber dynamics,[3] Eq. 3.23,

$$p_{flip} = \frac{\exp(-2J/T)}{\exp(+2J/T) + \exp(-2J/T)}.$$

Normally the spins will prefer to remain in the ordered direction. The

[3] We shall omit hereafter the Boltzmann constant since the temperture units have no special significance.

ratio of the probability for a flip into the 'wrong' state to that of remaining in the original state is

$$\frac{p_{down}}{p_{up}} = e^{-4J/T}$$

because the flipping of a single spin breaks two bonds and raises the energy by $4J$. Typically, one would have to make approximately $\exp(4J/T)$ updates in order to provoke a single flip. But we have to keep in mind that each time unit involves N updates. Hence, the number of time units required to provoke a flip decreases as the size of the system increases. The actual number of time units required for one flip is

$$t_{flip} \approx \frac{\exp(4J/T)}{N} = \exp\left(\frac{1}{T}(4J - T\ln N)\right).$$

The argument now proceeds as follows:

- Since the energy cost $(4J)$ is independent of N, as N increases t_{flip} decreases, and can be made arbitrarily small.
- Once a single spin has flipped, its neighbors have only the external field to keep them in the right direction, but if $T \gg h$ then the probability of flipping such a spin is $\frac{1}{2}$, provoking a domino effect.
- The same fate awaits every section of the ring which is long enough.
- After a while, the ordering of every long segment of the ring is broken and the distinction between up and down spins is erased, except for the tiny effect of the imposed external field.

This is how ergodicity is restored.[4] The reason is that while the system probes N spins per cycle, the balancing energy barrier, and hence the probability ratio between flip and no flip, is independent of N. Once a spin has been flipped it becomes very significantly easier to flip further spins, those sitting at the boundary between up and down sections. Any ordered section of the system that becomes large is broken. A further reflection reveals that the instability of the ordered regions is caused by two agents:

[4]The physicist should recognize here a dynamic version of Peierls' thermodynamic argument[13], with $4J - T\ln N$ playing the role of a free-energy increase due to a spin flip and $\ln N$ is the corresponding entropy.

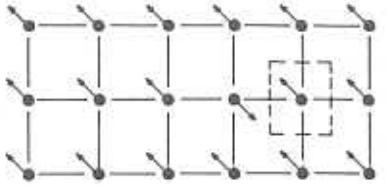


Figure 3.4: An Ising model on a square lattice with a single spin out of line. Note that a neighbor to the flipped spin still has a balance of two aligning bonds.

1. One-dimensionality.
2. Short range interactions.

If one contemplates, for example, a square lattice of spins, as in Figure 3.4, then clearly the energy of flipping a spin, in a sea of aligned spins, breaks four good bonds and hence costs $8J$ energy units, about as much as in the one-dimensional case. However, the neighbors of the flipped spin are not much easier to flip than the first one. The cost of flipping one of them is reduced to $4J$, from the previous $8J$ energy units, which implies that the probability of the disease spreading remains very low. This would also be the case for the one-dimensional system if each spin were to interact with two neighbors. In the one-dimensional case, however, flipping two adjacent spins would suffice to produce easy-to-flip spins, which would not be the case for the two-dimensional system. In fact, it can be proven rigorously that the one dimensional model is ergodic[12], while the nearest-neighbor two-dimensional model is not[14].

The decomposition of order in the one-dimensional case is shown by the dashed curve in Figure 3.6, where a system of 10,000 spins, with the small external field, is started in one of the fully aligned states and its overlap with this state is plotted as a function of time. The system is at temperature T=0.6J, and hence $4J$–$T\ln N$≈–1.4J .

3.2.4 Fully connected Ising model: real non-ergodicity

As the number of neighbors with which each spin interacts increases, so does the number of low-probability flips required for easy-to-flip spins to appear. This is true even in the one-dimensional chain. If this

number becomes as large as N itself, and N is large,[5] then the ordered state is restabilized. This will be demonstrated in the present section.

First a minor technical point must be made, one that will recur in our discussion of ANN's. If the number of interacting neighbors becomes very large, the energy has to be rescaled by that number to keep the local fields independent of the number of spins. If N spins interact with J_{ij}'s of order one the local fields can themselves become proportional to N. See e.g., Eq. 3.15. This implies that we would need a temperature (noise level) proportional to N in order for the noise to have any noticeable effect on the dynamics of the system, which is governed by the values of h/T. So, if h were of order N and T were independent of N, the system would always appear noiseless.[6]

The remedy is rather simple. If we would like to keep the temperature fixed as the number of interacting neighbors becomes infinitely large, the size of the interaction between each two spins should be rescaled. In other words, J should be replaced by J/N. Now we can discuss the Ising chain in which **all** spins interact ferromagnetically, with interactions $+J/N$, and allow the system to become arbitrarily large. The local field, Eq. 3.15, and the energy of the system, Eq. 3.16, take on the following forms, respectively:

$$h_i = \frac{J}{N}\sum_{j,j\neq i}^{N} S_j \rightarrow \frac{J}{N}M\{S\} \equiv Jm\{S\} \tag{3.30}$$

$$E\{S\} = -\frac{J}{2N}\sum_{ij,j\neq i} S_i S_j \rightarrow -\frac{J}{2N}M^2\{S\} \equiv -\frac{1}{2}NJm^2\{S\}. \tag{3.31}$$

The curly brackets indicate a dependence on a collection of all N spins. In both equations M stands for the total magnetization, as in Eq. 3.19. The m in both equations is the magnetization per spin, i.e., M/N, which is also the overlap (or cosine) between the state $\{S\}$ and the spin-up attractor state, as in Eq. 1.7 with $S_i = 1$. The right arrows

[5] As the range of the interaction becomes of the size of the system and the system becomes very large, the meaning of dimensionality fades away. There is no difference between systems of one, two or any other number of dimensions, if all spins interact. Such a system is sometimes described as being effectively of an infinite number of dimensions.

[6] The field (analog of the pressure), like the temperature, should, technically speaking, be an *intensive variable*, i.e., independent of the size of the system.

express the fact that the term with $j = i$ – the self-interaction of a spin – is absent on the left hand side and is included on the right. But this is one out of N terms of the same magnitude and as N becomes very large its relative effect becomes negligible.

Let us now return to the dynamical process. The transition probability for a spin to flip **into** state S_i, Eq. 3.23 now reads

$$\Pr(S_i) = \frac{\exp(Jm\{S\}S_i/T)}{\exp(Jm\{S\}/T) + \exp(-Jm\{S\})/T}, \tag{3.32}$$

with $m\{S\}$ depending on the particular state the system is in. The self-interaction has been treated as in Eq. 3.31, but this is justified when N becomes very large.

The probability distribution of the states, $\rho(\{S\}, t)$, is determined by the master equation, Eq. 3.5, which can be converted into a differential equation by subtracting $\rho_I(n)$ from both sides of the equation, and dividing by the cycle-time constant, which we shall denote by Γ. The equation can be written as[12]:

$$\frac{d\rho(\{S\},t)}{dt} = -\sum_{i=1}^{N}\sum_{\bar{S}=-1}^{+1} S_i\bar{S}_i w(-\bar{S}_i)\rho(S_1,...,\bar{S}_i,...S_N,t). \tag{3.33}$$

This is just Eq. 3.5, except for the fact that in the Glauber dynamics different states that can be dynamically connected differ by exactly one spin-flip. That is why on the right hand side we have a sum over spins i rather than over states I. The w's are *transition rates*, namely they are probabilities for transition per unit time, and are simply $w = p/\Gamma$. The special factor $S_i\bar{S}_i$, summed over $\bar{S}$, is a reflection of the positive and negative terms in the master equation. When $\bar{S}_i = S_i$ there is a probability flow out of the state I, and we have a negative term, and when $\bar{S}_i = -S_i$ the flow is into the state I, and the term is positive. This also explains the fact that the argument of w is $-\bar{S}_i$.

Let us now focus on the time development of the **mean** magnetization per spin. Given the probability distribution at any given time, the mean of any observable can be simply expressed as

$$\langle m\rangle(t) = \sum_{\{S\}} m\{S\}\rho(\{S\},t), \tag{3.34}$$

where the sum is over all 2^N possible states of the system, as in Eq. 3.18. The time dependence enters only via the probability distribution and

so with the aid of the master equation, Eq. 3.33, one can obtain a differential equation for $\langle m\rangle(t)$. The details of the derivation are relegated to Appendix 3.6.2. We should, however, mention the proviso, for the reader who will prefer not to consult the appendix, that the derivation becomes exact only when N becomes very large. The equation reads

$$\Gamma\frac{d\langle m\rangle(t)}{dt} = -\langle m\rangle(t) + \tanh\left(\frac{J\langle m\rangle(t)+h}{T}\right), \tag{3.35}$$

where for the sake of forthcoming discussion the external magnetic field was included in the dynamics.

Phase transition and physicists' non-ergodicity

Let us start with the equilibrium solutions of Eq. 3.35. Equilibrium implies that $\langle m\rangle$ does not vary with time, hence

$$\langle m\rangle = \tanh\left(\frac{J\langle m\rangle+h}{T}\right). \tag{3.36}$$

This is a celebrated equation in physics, it is the Weiss molecular mean-field equation for ferromagnetism, see e.g., ref. [14] ch. 4. The properties of its possible solutions for $h = 0$ depends, clearly, on the ratio J/T, i.e., the ratio of the aligning energy to the thermal (disordering) energy per spin. The possible equilibrium solutions can be very easily reviewed graphically. In Figure 3.5 the two sides of the equation are plotted on the same graph. In (a), $J/T < 1$ – the *high temperature phase* – the only intersection is at $\langle m\rangle = 0$. There is **no magnetization without external field**. In (b), $J/T > 1$ – *the low temperature phase* – there are three possible solutions, one at $\langle m\rangle = 0$ and two at the two non-zero intersections, symmetrically disposed about zero. The latter are **spontaneously polarized**. These are the two ferromagnetic equilibrium phases.

Stability analysis will show that the two polarized solutions are stable, while the one with $\langle m\rangle = 0$ is unstable. Still, a choice has to be made between the different stable equilibrium solutions. In thermodynamics, which is a theory of equilibrium, this is done by studying relative *free-energies*. The choice among solutions with equal free-energies is provoked by the introduction of infinitesimal external magnetic fields. See e.g., Section 3.4. Here the choice will be made, naturally, by the

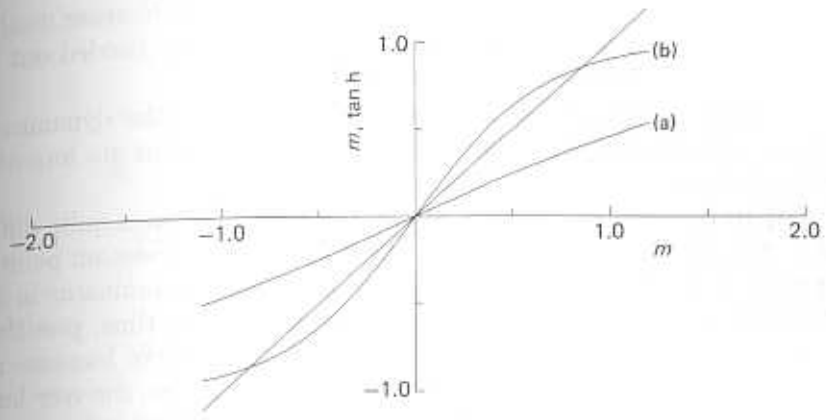


Figure 3.5: The graphical solution for the equilibrium magnetization. (**a**) The high temperature phase: solution only at m=0. (**b**) The low temperature, ferromagnetic phase: three solutions.

dynamical process by way of the initial conditions – this will be the explication of *non-ergodicity* and *cooperativity*.

Eq. 3.35 can be solved by quadrature, leading to

$$t - t_0 = \int_{m_0}^{m(t)} \frac{dm}{-m + \tanh(Jm/T)}, \tag{3.37}$$

where $\langle m\rangle$ was replaced by m, we have taken $h = 0$ and t_0 and m_0 are the initial time and mean magnetization, respectively.

Non-ergodicity is a statement about the properties of the solution, $m(t)$, of Eq. 3.37 for large t. Since t is to become very large, $m(t)$, the upper limit of the integral, will have to vary in such a way as to make the integral grow without limit. This cannot be the result of a very large upper limit, since $m(t) \leq 1$. Hence it must be a consequence of the upper limit arriving at a special point of the denominator of the integrand. The conclusions differ depending on whether the system is in the high or low temperature phase, corresponding to parts (a) and (b) in Figure 3.5.

(a) In the high temperature phase the denominator can vanish only at $m = 0$. Hence, as $t \to \infty$ $m(t) \to 0$, for any initial condition. If $m_0 > 0$, then the integrand is negative, and we must have $m(t) < m_0$ for all $t > t_0$, or the right and left sides of Eq. 3.37 will differ in sign.

For the integral to increase, as t increases, $m(t)$ must decrease until at very large t it tends to 0. The same argument can be carried out for the case $m_0 < 0$, *mutatis mutandis*.

In other words, in the high temperature phase, the dynamics is ergodic. After a long enough time, the initial conditions are forgotten by the system.

(b) In the low-temperature phase, the system behaves quite differently. Suppose that $0 < m_0 < \bar{m}$, where $\bar{m}$ is the intersection point in Figure 3.5(b). This is just the point at which the denominator in the integrand vanishes. At $m = m_0$ the integrand is, this time, **positive**. As a consequence, as t increases $m(t)$ must increase, to increase the value of the integral. This goes on until $m(t)$ reaches $\bar{m}$, for very large values of t. There the integrand diverges after an infinitely long time.

If the initial conditions are $\bar{m} < m_0 = m_1 < 1$, as in Figure 3.5, then the denominator is negative and to correct for the sign of the integrand one must have $m(t) < m_1$. This is very much like the situation in **(a)** above. The only difference is that after a very long time $m(t)$ will reach $\bar{m}$, rather than 0.

The conclusion is that for any initial condition for which there is positive mean magnetization, i.e., a surplus of positive over negative spins, the system will develop to a distribution of states which gives a *unique* value for the mean magnetization. This unique value depends, of course, on the value of J/T. When the argument is repeated for negative values of m_0, either above or below the intersection point, it proceeds along very similar lines. Now, though, the system drifts to a distribution which has $m(t)=-\bar{m}$.

This is a simple realization of non-ergodicity. Note that the dynamical process has lifted the strange symmetry between the three solutions $(-\bar{m}, 0, \bar{m})$ which was left behind by the equilibrium analysis, described above. Even the noisy system will flow to an *ordered* attractor. The choice of the attractor will depend on the value of the magnetization in the initial state.

To appreciate the difference between the short range chain and the present system, whose phase transition is an aspect of its robustness to noise, consider the three full curves in Figure 3.6. They exhibit the dynamical development of $m(t)$, for three initial conditions, in a system with the same temperature as the Ising chain, represented by the dashed line.

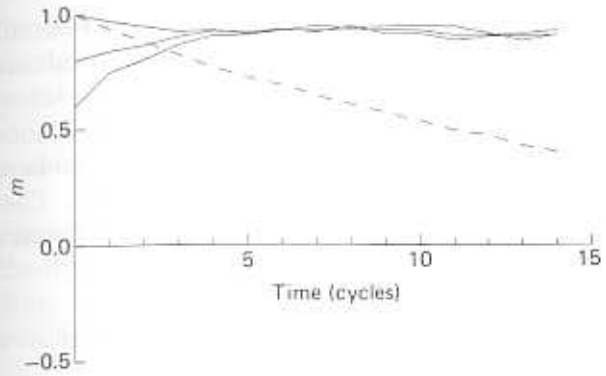


Figure 3.6: Simulation of time development of magnetization per spin in an Ising model. Dashed line – chain; full lines – fully connected, three initial conditions. $T/J = 0.6$. Compare Figure 3.3.

3.3 From Dynamics to Landscapes – The Free Energy

3.3.1 Energy as Lyapunov function for noiseless dynamics

In principle, the dynamical equations we have derived contain the answers to all possible questions. But as the situation becomes more complex, the analysis of the dynamical equations becomes extremely difficult. On the other hand, many of the pertinent asymptotic properties of the solutions can be read off from *free-energies*, about which one has collected many useful intuitions and methods in the context of equilibrium statistical mechanics. Most of the study to follow here is carried out within this culture. So, having clarified the dynamical side of non-ergodicity and cooperativity in this simple example, we now proceed to construct in this context the bridge to the free-energy.

We have seen in Section 3.2.3 that in the absence of noise the dynamics will drive the Ising model toward the minima of the *energy*. In a more general context of dynamical systems, the energy goes under the name of *Lyapunov function*, which has been widely used in the discussion of cellular automata[15], and even in the description of the dynamics of more complicated systems with some continuous variables[16,17]. This drift toward energy minima is intuitively clear, since in the absence of noise a physical system can either decrease its

energy by transferring some to a very large and cold reservoir, or it can change at constant energy. Consequently, any local minimum will be an attractor and those will either be fixed points, or the system may hop between local minima with equal energy. The existence of the energy function allows also for the introduction of the landscape picture. This was already raised as a metaphor in Section 2.3.1. The dynamical laws governing physical systems are the reality behind such metaphors.

3.3.2 Parametrized attractor distributions with noise

At finite temperature the system can gain energy, at random, in amounts which are proportional to the absolute temperature T. As the energy gain becomes larger, its probability decreases exponentially. Hence some of the local **energy** minima are destabilized. Eventually, as the temperature becomes high enough the energy minima become irrelevant and the system wanders randomly among all its possible states. Even the lowest amounts of noise make the existence of individual states as attractors impossible, since there is always a finite probability that a spin will flip, against all odds. The best one can hope for is that within some of the energy valleys there would be **restricted** regions in the space of states from which the system will have *zero escape probability*. Such behavior can be expected only in infinite systems which can break ergodicity. Within such a region the system can still wander under the influence of noise, but only within a part of the total space of states.

- In the noisy system, one cannot speak of attractor states, but only of attractor probability distributions of states.

Each probability distribution is associated with one of the regions of broken ergodicity and describes the relative probabilities of various states within the subspace. Alternatively, one can parametrize the distributions in terms of a collection of mean values of *observables*. For example, in Section 3.2.4 we have chosen $\langle m(t) \rangle$ as a single parameter to characterize the distribution. As we shall see throughout this volume, in different situations the collection of expectation values to characterize the distribution may vary, depending on the quantities of direct relevance.

From the discussion in Section 3.2.2, we can conclude that if detailed balance holds, the asymptotic distribution will be the Gibbs distribution (or ensemble)

$$\rho(\{S\},t) \to \rho_G(\{S\}) = C \cdot e^{-E\{S\}/kT}$$

and every observable takes on expectation values according to Eq. 3.18. But from the dynamical point of view, the situation is somewhat more delicate. Since the dynamical process may, at low temperatures, become stranded in a restricted part of the space of states, then

- the expectation values should include only those states that are actually accessible to the system.

3.3.3 Free-energy landscapes – a noisy Lyapunov function

It is a hundred year old wisdom, due to Boltzmann, that if a system changes in conditions of constant energy it will *monotonically* increase its entropy. Equilibrium will be reached when entropy has been maximized. In our case, we also have a system which drifts toward equilibrium. After all, we know that if detailed balance holds in the dynamical process, then the distribution of states tends, after a long enough time, toward the Gibbs distribution, and the latter represents equilibrium in that it does not vary with time. See e.g., Section 3.2.2.

Yet, the systems under discussion here are not changing at constant energy. They are not isolated systems. Instead, they change at constant temperature, or noise level. As such, the system neither minimizes its energy, nor maximizes its entropy, but instead it minimizes its *free-energy*. The free-energy, F, is a combination of the energy and the entropy, namely

$$F = E - TS \tag{3.38}$$

where S is the entropy, which is simply related to the logarithm of the number of available system states at a given energy. The greater this number of states the higher the entropy and with it the disorder.

The dynamical expression of these thermodynamical insights is summarized in the following assertion:

- A dynamical process which preserves detailed balance monotonically reduces its free-energy.

Our Ising model is a special case of such a system and hence in every step of the Glauber dynamics the free-energy is reduced. The same applies for all the dynamical processes introduced in Section 2.2.1 to describe the behavior of neural networks. This assertion provides an extension of the landscape picture to noisy, or finite temperature, systems. Combining these remarks with the description of the probability distribution in terms of its mean values, Section 3.3.2, one can summarize:

- The landscape is the outline of a free-energy surface as a function of a set of mean values which parametrize the probability distributions of the states.

The rest of this chapter will be an effort to make these assertions as clear, as plausible and as useful as possible.

A natural way of defining a dynamical version of the free-energy, Eq. 3.38, in a situation which is both stochastic and time dependent, is to use the probability distribution of the states of the system, $\rho_I(t)$ from Section 3.1.1, which encompasses both aspects. Given ρ one can write

$$F\{\rho\} = \sum_I \rho_I(t)E(I) + T\sum_I \rho_I(t)\ln\rho_I(t) \equiv \langle E\rangle - T\langle S\rangle \qquad (3.39)$$

where the mean values $\langle O\rangle$ of the energy and the entropy are defined as usual, Eq. 3.34, in Section 3.2.4, namely

$$\langle O\rangle = \sum_I O\{I\}\rho_I(t) \qquad (3.40)$$

The mean energy needs no definition. The entropy, defined as

$$-\ln\rho,$$

has the usual properties one ascribes to an entropy: it is largest if all states are equi-probable – maximal disorder; it is smallest, zero, if one state has probability unity and the rest have zero probability – complete information.

The energy E is bounded from below and the entropy, as defined above, has an upper limit (for the equi-probable case) and this limit, multiplied by T, is the largest amount that can be **subtracted** from the energy, hence

- The quantity F is bounded from below, and if it decreases in the dynamical process, it can replace the Lyapunov function.

Moreover, from Eq. 3.39 it follows that the free-energy tends continuously to the energy as $T \to 0$, as the noise disappears.

It remains to be shown that F decreases in the course of the dynamical development of the system. This should be a consequence of Eq. 3.5, which can be written as a differential equation, as in Eq. 3.33, namely

$$\Gamma\dot{\rho}_I(t) = \sum_{J\neq I}[w(I \mid J)\rho_J(t) - w(J \mid I)\rho_I(t)]. \qquad (3.41)$$

In Appendix 3.6.3 it is shown that on its way to equilibrium the system decreases its free-energy at every step, provided it satisfies *detailed balance*. This result implies that

- **The free-energy is the natural extension of the energy landscape description to situations with noise.**

3.3.4 Free-energy minima, non-ergodicity, order-parameters

In the previous section we have seen that a stochastic process of the heat-bath type, as a special case of a process which respects detailed balance, reduces the free-energy at every step. This has rescued the landscape metaphor, and ensures that the process will tend toward minima of the free-energy, on its way to a stationary probability distribution. We have simultaneously established that the system will reach a stationary state when it arrives at the Gibbs distribution, $\exp(-E/T)$. This may seem like a contradiction, since the latter statement can be interpreted as implying ergodicity, while the former, if the free-energy has more than one minimum, implies breaking of ergodicity. The latter statement may be interpreted as implying that all states are available, with relative probabilities which depend exponentially on their energy difference. The former clearly implies that the system's trajectories are confined in the space of states.

There is, of course, no contradiction. Let us recall that we are dealing with infinitely large systems. Noisy finite systems are always ergodic. If the free-energy has a variety of minima the trajectories will be confined to the basin of a particular one, selected by the initial state of the system. In the stationary (equilibrium) situation the system will continue wandering, since the minimum in the free-energy selects

a probability distribution, not a state. The relative probabilities of different states, all consistent with the particular minimum condition of the free-energy, will be dictated by the Gibbs distribution. In other words,

- the system, after having developed for a long enough time, will select one out of the available Gibbs distributions, each corresponding to a particular minimum of the free-energy.

Each of these Gibbs distributions is *restricted*, in that it does not cover the entire space of system states. Different *sectors* are characterized by different values of the set of parameters – expectation values of some selected observables – which were used to classify the probability distributions, as has been suggested in Section 3.3.2.

In the Ising system with long range interactions we have parametrized the distributions by the mean magnetization, or the overlap with the 'up' ferromagnetic state. In that case we have found a breakdown of ergodicity with two possible outcomes for $\langle m \rangle$, Section 3.2.4. The present discussion implies that the free-energy of the fully connected Ising ferromagnet has two minima, symmetrically disposed, **as a function of the mean magnetization**. Moreover, it also implies that there are two equilibrium Gibbs distributions, both with relative probabilities $\exp(-E/T)$. Each is restricted so that $\langle m \rangle$ has the value of one of the minima. The fact that $\langle m \rangle = 0$ is not reached dynamically indicates that at this value the free-energy is not minimal. At high temperature this is the only minimum, and the system is ergodic.

The essential wisdom involved is in the choice of the expectation values for the characterization of the probability distribution. A casual choice would have been rather unlikely to expose a good parameter – one that brings out the non-ergodicity. Parameters (expectation values of observables) that do are called *order-parameters*. The magnetization is a rather obvious candidate in a system like the Ising model. In general the proper choice of the *order parameters* is a major part of the creative process of analyzing a phase transition. Not only does one have to find parameters which detect the breaking of ergodicity, but one must find them all. Otherwise one may encounter unanticipated non-ergodicity within sectors. As we go along we shall encounter cases in which additional order-parameters will be required, in order to fully describe the long-time behavior of the neural network.

3.4 Free-Energy of Fully Connected Ising Model

3.4.1 From minimization equation to the free-energy

When we discussed the breaking of ergodicity in the fully connected Ising model, in Section 3.2.4, we have reached the crucial equation, Eq. 3.36,

$$m = \tanh\left(\frac{Jm + h}{T}\right),$$

after omitting the brackets denoting the average. This equation, according to Section 3.3.4, determines the minima of the free-energy of the system. Therefore it must be related to an equation of the type

$$\frac{\partial \phi(m,h)}{\partial m} = 0,$$

with ϕ being the free-energy per spin. One possible guess for ϕ would be

$$\frac{\Phi(m,h)}{N} \equiv \phi(m,h) = \frac{J}{2}m^2 - T \ln\cosh\left(\frac{Jm+h}{T}\right). \tag{3.42}$$

This is, of course, no method for the derivation of the free-energy. But, one can check by differentiation that the minima of ϕ are the solutions of the non-linear equation for m. One can also check that as $T \to 0$ this expression tends to the energy, Eq. 3.31,

$$E = -\frac{1}{2}NJm^2.$$

Note the Φ is the total free-energy of the system.

What is of real significance is that in terms of a free-energy like ϕ one can write the dynamical equation 3.35 as:

$$\Gamma\frac{dm}{dt} = -\frac{\partial \phi(m,h)}{\partial m} \tag{3.43}$$

even when $h \neq 0$. This demonstrates, in the case of the fully connected Ising model, that the free-energy may reconstitute a *gradient flow* even in the noisy situation. We return to this observation below, in the discussion of the appropriate free-energy for the analysis of the dynamical process.

In fact, strictly speaking, ϕ is not a standard representation of a free-energy (see e.g., ref. [18]), being a function of T and both m and h, which at equilibrium are dependent variables. In Appendix 3.6.4 it is shown that this function is closely related to the real free-energy and, more importantly, that its minima are **exactly** at the equilibrium values of m, for any values of the externally imposed variables h, T. See e.g., Section 3.4.2. At $h = 0$ the minima of this function are those of the standard free-energy, a function of m and T, as is verified explicitly in the next section. But the statement is very general, as is indicated in Section 3.4.2. The particular importance of ϕ in the present context is two-fold

- It is the free-energy – a function of h and T (see e.g., Section 3.4.2) – which is minimized in the dynamical process.
- The physicists' techniques, to be described below, are especially well suited for arriving at this type of function.

Appendix 3.6.4 provides a general prescription for transforming a free-energy like ϕ, a function of T and h, into a free-energy expressed in terms of m and the corresponding T. But we will encounter no good reason for affecting this transformation.

Before proceeding with more systematic ways of arriving at free-energies, we draw ϕ vs. m, Eq. 3.7, for a few values of T in Figure 3.7. There are four types of curves: (1) $T > J$ has a single minimum at $m = 0$, which is slightly displaced when $h \neq 0$. This is the *paramagnetic*, ergodic phase. (3) has multiple minima, which are equally low when $h = 0$. The $m = 0$ solution becomes a **maximum**. The dotted curve exhibits the effect of an external field on the two minima (see the discussion at the end of Section 3.4.2). (2) $T = J$ where all three solutions merge. This discussion should be compared with the discussion following Figure 3.5.

It transpires that this is a third equivalent language for discussing the classification of non-ergodicity in the dynamics of the system. Alongside non-ergodicity in the dynamical equations and the minimization of a dynamical free-energy one finds a pure equilibrium property of the system which brings out all the salient features. An important point to note, one that will be amplified in Section 3.4.2 below, is that the symmetry of curve 3 in Figure 3.7 is broken by the presence of the external field. This lifts the *degeneracy* between the two minima and in the dotted curve in Figure 3.7(1) there is a single absolute minimum.

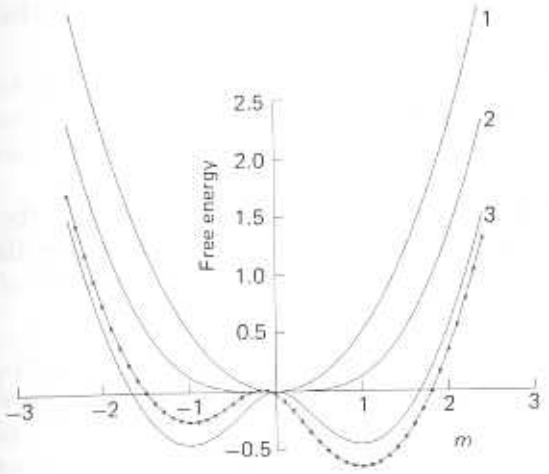


Figure 3.7: The 'free-energy' of the fully connected ferromagnetic Ising model. Undotted curves have h=0: (1) $T > J$, there is a single minimum at m=0; (2) T=J, the transition point; (3)$T < J$, there are two non-zero minima. Dotted curve has $T < J$ and h=0.2.

3.4.2 The analytic way to the free-energy

Statistical mechanics since Gibbs has developed powerful techniques for arriving at free-energies in rather general situations. Such techniques will be often used below, so we devote this section to a description of the techniques in the relatively simple context of the Ising model. But first we must recapitulate a few standard elements of statistical mechanics.

Digression: a teaspoon of statistical mechanics

Since the dynamics leads the system to equilibrium, which is **both** a *Gibbs ensemble* and a minimum of the free-energy, the task is to compute the free-energy in the Gibbs ensemble. In this ensemble, a state's probability is proportional to $\exp(-E/T)$. Following a rather common practice in statistics, one forms from the probability distribution a *characteristic function*, Z, by adding to the energy of a state an additional term

$$E' = -hO\{S\}$$

where O is any observable defined for every state of the system. The function Z is defined as:

$$Z(h,T) \equiv \sum_{\{S\}} \exp\left(-\frac{(E+E')}{T}\right). \tag{3.44}$$

This function is both a normalization constant for the probabilities and, more importantly, it is a *generating function* for the expectation values of powers of the observable O. Every derivative of $Z(h,T)$ with respect to h brings down a power of O and so

$$\langle O^n \rangle \equiv \frac{1}{Z(h,T)} \sum_{\{S\}} O^n \exp\left(-\frac{E}{T}\right) = \frac{1}{Z(h,T)} \frac{\partial^n Z(h,T)}{\partial h^n}\bigg|_{h=0}. \tag{3.45}$$

Since we will be almost exclusively concerned with means of various quantities, $Z(h,T)$ is a function of central importance. In statistical mechanics it has been named *partition function.*

We will not be interested in the most general partition function, but will restrict ourselves to

$$O = m\{S\} = \frac{1}{N}\sum_i S_i,$$

i.e., the magnetization, which is later to become the retrieval variable. In that case, the parameter h is just the external magnetic field introduced in Section 3.2.1. Common lore[1] has it that the free-energy, $f(h,T)$ is given in terms of Z as

$$f(h,T) = -\frac{T}{N}\ln Z(h,T). \tag{3.46}$$

Without going into detail, we just point out that this identification holds because both sides satisfy the thermodynamic relations:

$$\frac{\langle E \rangle}{N} = -T^2 \frac{\partial}{\partial T}\left(\frac{f}{T}\right) \tag{3.47}$$

$$\langle m \rangle = -\frac{1}{T}\frac{\partial f}{\partial h}, \tag{3.48}$$

which is an exercise in the application of Eqs. 3.45 and 3.46. See e.g., refs. [1,18].

Computation of free-energy of fully connected Ising model

The task is to obtain $f(h,T)$, or equivalently $Z(h,T)$. To compute Z we must evaluate a sum over all 2^N states, in which each spin takes its two possible values. Z can be written as

$$Z(h,T) = \sum_{\{S\}} \exp\left[\frac{J}{2NT}\left(\sum_i S_i\right)^2 + \frac{h}{T}\sum_i S_i\right]. \tag{3.49}$$

The difficulty in carrying out this N-fold sum is caused by the presence of products of S_i's with different i's in the quadratic term in the exponent. But due to the full connectivity, the interactions between different spins enter only through one single square of the total magnetization. It can be simplified by a Gaussian transform, which is nothing more than the identity

$$\exp\left(\frac{A}{2}O^2\right) = C\int_{-\infty}^{\infty} dx \exp\left(-\frac{x^2}{2A} + xO\right). \tag{3.50}$$

Applying this identity to the exponential of the quadratic term in M in the expression for Z, and setting $J = 1$, one has:

$$Z(h,T) = C\int_{-\infty}^{\infty} dx Tr_S \exp\left(-N\frac{x^2}{2T} + \frac{h+x}{T}\sum_i S_i\right). \tag{3.51}$$

In the last two equations C is some immaterial constant, not necessarily the same. The notation Tr_S is short-hand for the N-fold sum over all system states.

In this last expression for Z, the sum over states can be easily carried out, since the summand is a product of N terms, each depending on a single S_i with different i. The sum reduces, therefore, to a product of N single spin sums, each containing two terms – corresponding to the two possible values of the spin. Each factor in this product is:

$$\sum_{S=-1}^{1} \exp\left(\frac{h+x}{T}S\right) = 2\cosh\left(\frac{x+h}{T}\right). \tag{3.52}$$

The N terms are all equal, and when they are introduced into Z one has:

$$\begin{aligned} Z(h,T) &= C\int_{-\infty}^{\infty} dx \exp\left(-N\frac{x^2}{2T}\right)\cdot\left[2\cosh\left(\frac{x+h}{T}\right)\right]^N \\ &= C\int_{-\infty}^{\infty} dx \exp\left[-N\left(\frac{x^2}{2T}+\ln\cosh\left(\frac{x+h}{T}\right)\right)\right]. \end{aligned} \tag{3.53}$$

Let us make a short summary at this point:

- The evaluation of Z has been reduced to a single integral.
- With Z in hand, we have direct access to the mean values of all powers of m, which are expressed as derivatives of Z.
- Given Z, we have the free-energy via Eq. 3.46 – a tool for the determination of the structure of non-ergodicity.
- The above statements are exact for a fully connected system and are approximate for systems with sparser connections. This goes under the name of *mean-field theory*.

To complete the calculation the integral has to be carried out. Even in this simple case the explicit integration is not possible, so in order to prepare for the more general cases to be dealt with in the following chapters we shall analyze it by a general method – the method of *steepest descent*. This is an exceptionally powerful method for computing the asymptotic behavior of integrals of an exponential with a very large coefficient in the exponent[19]. The mathematical statement is that

$$I(a) = \int \exp\left[N\psi(x,a)\right]dx \approx \sum_j C\exp\left(N\psi[\bar{x}_j(a),a]\right) \tag{3.54}$$

as $N\to\infty$, where $\bar{x}_j$ are the points in the integration interval at which the function $\psi(x,a)$ has its **absolute** maxima. They would have been the minima were there a negative sign in the exponent. These points depend, of course, on the parameter a. The above equation gives the leading term for large values of N. Corrections to this leading term can be computed systematically[19].

Note that Eq. 3.54 is particularly well suited for the right hand side of Eq. 3.53. After all, we are interested in the free energy in the limit of a very large system. Otherwise, ergodicity will not, strictly speaking, be broken. After the asymptotic expansion Eq. 3.54 is applied to the partition function Eq. 3.53, the free-energy is derived from Eq. 3.46. It reads

$$f(h,T) = \frac{\bar{x}^2(h,T)}{2} - T\ln\cosh\left(\frac{\bar{x}(h,T)+h}{T}\right) \tag{3.55}$$

where $\bar{x}(h,T)$ is a minimum of $f(x,h,T)$, due to the minus sign in Eq. 3.53. In other words, $\bar{x}$ is a solution of the equation

$$\frac{\partial f(\bar{x},h,T)}{\partial\bar{x}} = \bar{x} - \tanh\frac{\bar{x}+h}{T} = 0 \tag{3.56}$$

which is equivalent to Eq. 3.36.

Still, three points require explanation:

- First is the relation between the mysterious $\bar{x}$ and the mean magnetization m.
- Second is the relation between this free-energy and the conventional one, expressed in terms m and T.
- Third is the whereabouts of the sum over minima, as implied in Eq. 3.54.

The saddle-point and the mean magnetization

We proceed to show that at the *saddle-point*,[7] the point in the integration interval which dominates the integral, $\bar{x}$ is equal to the mean magnetization m. To see this one observes, by referring to either Eq. 3.45 with $O = m\{S\}$ or to Eq. 3.48, that

$$m = \frac{T}{N}\frac{\partial\ln Z(h,T)}{\partial h} = -\frac{\partial f}{\partial h}. \tag{3.57}$$

When this is applied to Z expressed as the integral Eq. 3.53, one finds

$$m = \frac{1}{Z}\int_{-\infty}^{\infty} x dx \exp\left[-N\frac{x^2}{2T}+\ln\cosh\left(\frac{x+h}{T}\right)\right].$$

[7] Physics tends to lump together three methods – steepest descent, saddle-point and Laplace's method, unless distinctions are absolutely required.

In other words, m is the mean of the variable x with the probability distribution given by the integrand. But, at the *saddle-point* the integral in the above equation is dominated again by the value of the integrand at $\bar{x}$, namely by

$$\bar{x} \cdot e^{-Nf(\bar{x},h,T)},$$

with f given by Eq. 3.55. The denominator is the same expression without the factor $\bar{x}$. Hence,

$$m(h,T) = \bar{x}(h,T). \tag{3.58}$$

The important lesson to be drawn here is that the solutions, $m(h,T)$, of the equation

$$\frac{\partial f(m,h,T)}{\partial m} = 0,$$

which are also **minima** of f, Eq. 3.55, are equilibrium mean magnetizations of the system. Note that Eq. 3.58 allows the replacement of x by m. The free-energy f can therefore be used to provide landscapes for the dynamics.[8]

Non-equilibrium free-energy

From the expression for the partition function it transpires that f, of Eq. 3.46, is a function of h and T, which are the natural variables to be kept constant. After all, the Glauber process, for example, does not preserve E or m in its transitions, but rather T and h. This is also the free-energy which was proved to be monotonically decreasing in our dynamical process, in Section 3.4.1, as is explicitly shown by an equation like 3.43.

The conventional free-energy which is expressed in terms of the variables m and T, can be derived from $f(h,T)$, calculated in Eq. 3.55, by what is known in the jargon as a Legendre transform. See e.g., Appendix 3.6.4. It should be emphasized that this transformed free-energy, expressed in terms of m and T, is minimized to give the equilibrium values of m only when the field $h = 0$. In general one has

$$\frac{\partial g(m,T)}{\partial m} = h(m,T)$$

[8]The physicist should assimilate the fact that the above statement holds although f is, strictly speaking, the free-energy of the Helmholtz type – a function of T and h, as is explained below. See e.g., ref. [20].

which becomes an equation for a minimum only for $h = 0$. On the other hand the hybrid construct $f(m,h,t)$ is minimized to give m for any value of h. But, as was already indicated above, the different values of m for the same field h, do not represent equilibrium situations. For the fully connected model

- the slope of f, as a function of m, gives the rate at which m drifts toward its equilibrium value for the given value of h.

Degeneracy, symmetry and symmetry breaking

This leads to a point intimately connected to the breaking of ergodicity. In general one may expect to have more than one minimum for the function $f(m,h,T)$. It is rather exceptional that a few of these minima will give exactly equal values to f. The appearance of such degeneracies is, quite generally, the manifestation of an *underlying symmetry* of the system. For example, the fact that the free-energy of curve number 3 in Figure 3.7 had two minima for $h = 0$, with identical free-energies, was a consequence of the spin reversal symmetry possessed by the Ising system, as was remarked in Section 3.2.3. The symmetry of the energy brings about the equality of the probabilities of two states with all spins reversed. This necessarily leads to equal free-energies. The presence of the smallest field breaks the symmetry and lifts the degeneracy of the minima of the free-energy.

As far as the dynamics of the system is concerned, the symmetry and the resulting degeneracy are irrelevant. As we have seen in analyzing the dynamical behavior of the ferromagnetic phase, in Section 3.2.4, the initial conditions determine into which of the two symmetrically placed attractors the system will actually flow. In this sense the initial conditions *break the symmetry*. Since such accidents of degeneracy do not affect the dynamics, they should not be allowed to complicate our calculations. This is prevented by adding a small external field.

As soon as two minima differ in their free-energy, only one contributes to the integral, computed by the method of steepest descent. The reason is that if N is large enough, the contribution of a minimum whose free-energy is higher than that of the absolute minimum by an amount Δf, will be down by a factor $\exp(-N\Delta f)$ relative to the leading contribution from the lowest minimum. The addition of a small external field explicitly *breaks the symmetry* and lifts the degeneracy of the minima connected by the symmetry. Compare, for example, the

free-energies of curve (3) and the dotted curve in Figure 3.7. The degeneracy is lifted by a difference which vanishes when $h \to 0$. But it is proportional to N. If N is made very large, a single minimum is selected by the computation of the saddle-points. Then one can take the limit $h \to 0$ without mixing the various minima.

The procedure for unravelling all the minima, which are attractors for the dynamics, is to treat the saddle-point computation as if there were a single absolute minimum. That minimum can be any one of the absolute minima which are found by the minimization of $f(m, h, T)$. The understanding being that whatever the list of degenerate minima that may be thus uncovered, they may be treated one by one if a tiny external field is added to lower the free-energy of that particular phase. The same breaking of symmetry, and selection of a single attractor, will be affected dynamically by the initial conditions.

3.4.3 Attractors at metastable states

Let us consult again the dotted curve in Figure 3.7, which represents the free-energy in the presence of an external field. The field lowers one of the two minima relative to the other, but when $T < J$ the higher minimum remains a *local minimum*, and may even be surrounded by high *barriers*. It is the effect of such minima, which are local minima, on the dynamics of the system which will be briefly discussed in the present section, rather than the effect of external fields. As we shall see in much of the forthcoming analysis, when a neural network becomes interesting as a storage device, it develops local minima, sometimes referred to as *metastable states*, even in the absence of *symmetry breaking* fields. Despite the fact that these local minima will never dominate the steepest descent integral, they will significantly affect the dynamical behavior of the system for large classes of initial conditions. The Ising model is again the simplest meaningful terrain for clarifying such ideas.

According to Eq. 3.43, the dynamical development of the mean magnetization is governed by the relaxation equation

$$\Gamma \frac{dm}{dt} = -\frac{\partial f(m, h, T)}{\partial m}.$$

If the system is started at any point on the marked curve in Figure 3.7, to the left of the local minimum, the slope of f is negative and m will increase with time. If it reaches the minimum it will stay there, since dm/dt vanishes there. Similarly, if the initial conditions place the system anywhere between the local minimum and the maximum to its right, then the slope of f is positive and m will decrease with time, drifting again to the local minimum.

Any local minimum will be an attractor and a stable one. Therefore, in analyzing the free-energy, attention must be paid to all minima, including the local ones. As a matter of fact, in Chapter 6 we will find that it is the local minima that are of primary interest. This is unlike the situation in thermodynamics, where the system's equilibrium is determined by the absolute minima only. The origin of the difference lies in the fact that here we have prescribed a specific dynamical process which makes metastable states dynamically stable. In thermodynamics, the process is undefined and thus only absolute minima can be counted on.

3.5 Synaptic Symmetry and Landscapes

Returning to the general Ising system, we shall now establish that

- a sufficient condition for the existence of an *energy function* – noiseless *Lyapunov function* – or of detailed balance and hence of a noisy *landscape*, is the symmetry of the interactions.

The cases of synchronous and asynchronous dynamics will be discussed separately.

3.5.1 Noiseless asynchronous dynamics – energy

In a network of N discrete variables $S_i = \pm 1$, consider the function

$$E = -\frac{1}{2} \sum_{i,j,i \neq j} J_{ij} S_i S_j \tag{3.59}$$

with arbitrary J_{ij}'s. Let S_k, for a particular k, change to $-S_k$ at a given step of the dynamical evolution, while all the other S_i's remain fixed. The resulting change in E would be

$$\Delta E = S_k \sum_{j,j \neq k} J_{kj} S_j + S_k \sum_{j,j \neq k} J_{jk} S_j. \tag{3.60}$$

The 'firing rule' of Eqs. 2.6 and 2.5 implies that S_k changes sign only when the first term on the right hand side of Eq. 3.60 is negative. Otherwise S_k is in the direction of the local field and will not change. For a general matrix J the magnitude and the sign of the second term on the right hand side of Eq. 3.60 are unrelated to those of the first term. For a symmetric J_{ij} the two terms are identical and if a move is made, i.e., S_k is flipped, both terms are negative and E is necessarily reduced.

Symmetry has therefore been shown to be a *sufficient* condition for the existence of a landscape function of the noiseless asynchronous process. The function E is, of course, the energy of a corresponding spin system for which the symmetry is ensured by the laws of physics. The existence of such a landscape function ensures, in turn, that the network will asymptotically drift to a fixed point attractor.

It is important to note here, as well as in the following discussion of the noisy system, that the argument depends crucially on the absence of the *self-interaction* term J_{kk} from the expression for the field. The reason is that because $S_k^2 = 1$, the term with S_k is absent from Eq. 3.60, which expresses the energy change on flipping that particular spin. Thus the terms on the right hand side of 3.60 can be related to the local field at that site, **only if** S_k is absent, i.e., if $J_{kk} = 0$.

3.5.2 Detailed balance for noisy asynchronous dynamics

In the presence of synaptic noise, represented by the probabilistic firing rule of Eqs. 2.19, the network will never relax into a fixed point. Instead, under favorable conditions, its temporal evolution will sample a restricted set of states – breakdown of ergodicity – and the different sets will be characterized by different values of the averages of some observable quantities. See e.g., Section 3.3.4. As was indicated in Section 3.2.2 the condition which ensures the convergence of the distribution of states into attractors is *detailed balance*, Eq. 3.21, i.e.,

$$\frac{W(I \mid J)}{W(J \mid I)} = \frac{F(I)}{F(J)}. \tag{3.61}$$

- Synaptic symmetry is a sufficient condition for detailed balance.

This is the noisy equivalent of the existence of an energy function in the noiseless case[21].

The argument goes as follows: Suppose that in the transition between state J and state I it is neuron k which changes from S_k to $-S_k$. Eq. 2.19 implies that

$$\frac{W(I \mid J)}{W(J \mid I)} = \frac{\Pr(-S_k)}{\Pr(S_k)} = \frac{\exp(-h_k S_k/T)}{\exp(h_k S_k/T)} = exp\left(-\frac{2h_k S_k}{T}\right)$$

For detailed balance to hold, this last expression must be a ratio of a quantity which depends only on network state I to the same quantity for network state J. If the matrix J is symmetric then $2h_k S_k$ is just the energy difference, ΔE of Eq. 3.60, between the state with S_k and the one with $-S_k$ (compare Eq. 3.24). The function F is, in that case,

$$F(I) = \exp\left(-\frac{E(I)}{T}\right).$$

One can now carry over the entire discussion of Section 3.3.3.

- The distribution of the configurations relaxes after a long time to the Gibbs distribution, F (Eq. 3.26),

$$\Pr(\{S\}) \propto \exp\left(-\frac{E(\{S\})}{T}\right), \tag{3.62}$$

  with E given by Eq. 3.59.

- The attractors of the dynamical process are the global and local minima of the free-energy.

The landscape metaphor of Sections 2.3.1 and 3.3.3 now reflects a true downhill flow on the energy, or free-energy, surface. The dynamics of a neural network is thus reduced to thermodynamics and the noise parameter β^{-1}, defined by the dynamics, becomes a *bona fide* analog of a temperature.

3.5.3 Noiseless synchronous dynamics – Lyapunov function

The discussion which led to Eq. 3.59 does not apply in the case of synchronous dynamics, because one cannot assume that only one of the S_i's changes in a single dynamical step. However, also in this case it is possible to define a landscape (Lyapunov) function both in

the absence and in the presence of noise[21,22]. First, we discuss the noiseless situation.

Consider the following candidate for an energy:

$$E = -\sum_i |\sum_{j,j\neq i} J_{ij}S_j|. \tag{3.63}$$

Note that the diagonal elements of J are not included. The absolute value of a quantity can be expressed as

$$|x| = x\ \mathrm{sign}(x).$$

Hence, E can be rewritten as

$$E = -\sum_i \mathrm{sign}\left(\sum_{j,j\neq i} J_{ij}S_j\right)\sum_{j,j\neq i} J_{ij}S_j \tag{3.64}$$

If the S_j's represent the network state at time t–1, then the sign-factor is just the network state reached by synchronous updating at the subsequent instant t (Eq. 2.22). Thus, E can be read as a function of a single state, as in Eq. 3.63, or as a function of neural activities belonging to two consecutive states, i.e.:

$$E(t) = -\sum_{i,j,j\neq i} J_{ij}S_i(t)S_j(t-1). \tag{3.65}$$

It takes a few elementary manipulations of indices to show that for a symmetric J_{ij}, the change in $E(t)$, Eq. 3.63, is

$$\Delta E = E(t+1)-E(t) = -\sum_i[(S_i(t+1)-S_i(t-1)]\sum_{j,j\neq i} J_{ij}S_j(t). \tag{3.66}$$

When $S_i(t+1)$ is not equal to $S_i(t-1)$ (note the double time step), then the sign of their difference is the sign of $S_i(t+1)$. But the sign of the second factor (the sum over j) is also equal to the sign of $S_i(t+1)$, by virtue of the dynamical equations, Eq. 2.22. Thus, the right hand side has a negative contribution from every site at which $S_i(t+1) \neq S_i(t-1)$. This proves that E is a type of Lyapunov function – it is bounded from below and it decreases almost monotonically with every step of the dynamics. The function E will not decrease and remain unchanged either when the consecutive states are identical, or when two states alternate, i.e., when the system has either reached a *fixed point* or a *2-cycle*. The existence of such cycles, even for symmetric interactions, is a unique feature of synchronous dynamics[22].

3.5.4 Detailed balance for noisy synchronous dynamics

The transition probability in the noisy synchronous case is given by Eq. 2.21. The ratio of $W(I \mid J)$ to $W(J \mid I)$, which determines detailed balance is a product of two factors. The first is the ratio of the numerators in Eq. 2.21 and the second is the ratio of the denominators. The second factor is clearly a ratio of a quantity which depends *only* on I and a quantity which depends only on J. This term alone would have provided detailed balance unconditionally. The first term, the ratio of the numerators, is more problematic. It is the ratio of two exponentials, and hence an exponential of the difference

$$\sum_i h_i^J S_i^I - \sum_i h_i^I S_i^J.$$

But this difference becomes

$$\sum_{i,j,j\neq i} J_{ij}S_j^J S_i^I - \sum_{i,j,j\neq i} J_{ij}S_j^I S_i^J = \sum_{i,j,j\neq i}(J_{ij}-J_{ji})S_j^J S_i^I$$

when the expressions for h_i are substituted. It, therefore, clearly vanishes if the synaptic matrix is symmetric. One is left with

$$\frac{W(I \mid J)}{W(J \mid I)} = \frac{F(I)}{F(J)},$$

where

$$F(I) = \prod_i[2\cosh\beta h_i] = \exp\left(\sum_i \ln\left[2\cosh\left(\beta\sum_{j,j\neq i} J_{ij}S_j\right)\right]\right). \tag{3.67}$$

This proves that the symmetry is a sufficient condition for detailed balance and at the same time provides an explicit expression for the asymptotic distribution function[21]:

$$\Pr(\{S\}) \propto \exp\sum_i \ln\left[2\cosh\left(\beta\sum_{j,j\neq i} J_{ij}S_j\right)\right]. \tag{3.68}$$

From this expression one can read off the effective energy – the analog

of Eq. 3.59, which determines this distribution in the Gibbs formalism. It reads:

$$\bar{E} = -\frac{1}{\beta}\sum_i \ln\left[2\cosh\left(\beta\sum_{j,j\neq i} J_{ij}S_j\right)\right], \tag{3.69}$$

which is a rather unusual expression for the energy of a physical system. Note that in the zero temperature limit, this expression reduces to Eq. 3.63.

Finally, once detailed balance has been established and the form of the Gibbs distribution determined, either as Eq. 3.62 or as 3.68, the free-energy in both cases can be obtained using Eqs. 3.46 and 3.44. To recapitulate:

- The *partition function* is given as

$$Z = Tr_S \Pr(\{S\})$$

  with Tr_S standing for a sum over all network configurations.

- The free-energy, or the landscape function, is given in terms of the partition function as:

$$f = -\frac{T}{N}\ln Z.$$

3.6 Appendix: Technical Details for Stochastic Equations

3.6.1 The maximal eigen-value and the associated vector

We prove three assertions

1. All eigen-values of the stochastic matrix $\mathcal{W}$ are of magnitude smaller than or equal to unity.
2. The matrix $\mathcal{W}$ always has an eigen-value equal to unity.
3. There is a left eigen-vector, belonging to the eigen-value unity, whose components are all equal to one.

Consider the equation for a left *eigen-vector*

$$\sum_I V^L(I)W(I \mid J) = \lambda V^L(J). \tag{3.70}$$

The matrix elements of $\mathcal{W}$ are all positive and their sum over I is unity. Therefore, if $V^L(J_M)$ is the largest component of V, in absolute value, then

$$|\sum_I V^L(I)W(I \mid J)| \leq |V^L(J_M)|\sum_I W(I \mid J) = |V^L(J_M)|.$$

Taking absolute values on both sides of Eq. 3.70, for $J = J_M$, one can write

$$|\lambda||V^L(J_M)| = |\sum_I V^L(I)W(I \mid J_M)| \leq |V^L(J_M)|,$$

from which it follows that

$$|\lambda| \leq 1.$$

Next we note that, because of the normalization of the sum of $W(I \mid J)$ over I, setting $V^L(I) = 1$ in Eq. 3.70 the equation is satisfied for all I, provided $\lambda = 1$. Consequently, this vector is a left eigen-vector with eigen-value unity. This proves both assertions 2 and 3.

3.6.2 Differential equation for mean magnetization

Here we fill in the details necessary for the transition from Eq. 3.34 to the differential equation Eq. 3.35. First we write

$$\Gamma\frac{d\langle m\{S\}\rangle(t)}{dt} = \sum_{\{S\}} m\{S\}\Gamma\frac{d\rho(\{S\},t)}{dt}.$$

Now we substitute the equation for the dynamical development of the probability distribution ρ, under a microscopic Glauber dynamics, Eq. 3.33. This gives

$$\Gamma \frac{d\langle m\{S\}\rangle(t)}{dt} = -\sum_{\{S\}} m\{S\} \sum_{i=1}^{N} \sum_{\bar{S}_i=-1}^{1} S_i \bar{S}_i \Pr(-\bar{S}_i) \rho(S_1, \ldots, \bar{S}_i, \ldots, S_N, t), \tag{3.71}$$

For the transition probability we write

$$\Pr(-\bar{S}_i) = \frac{\exp(-Jm\{S\}\bar{S}_i)}{\exp(Jm\{S\}) + \exp(-Jm\{S\})} = \frac{1}{2}\left[1 - \bar{S}_i \tanh(Jm\{S\})\right].$$

The last equality can be verified directly be substituting the two possible values of $\bar{S}_i$. The magnetization operator $m\{S\}$ is just

$$m\{S\} = \frac{1}{N}\sum_{i=1}^{N} S_i.$$

When Pr and m are substituted in Eq. 3.71 it reads

$$\Gamma \frac{d\langle m\{S\}\rangle(t)}{dt} = -\frac{1}{2N}\sum_{\{S\}}\sum_{i,k} S_k S_i \times \sum_{\bar{S}_i=-1}^{1} \bar{S}_i \left[1 - \bar{S}_i \tanh(Jm\{S\})\right] \rho(S_1, \ldots, \bar{S}_i, \ldots, S_N, t).$$

This expression can be divided into two parts, according to the two terms in the square parentheses. The first, proportional to the 1, can be calculated as follows: writing explicitly the terms in the sum over $\bar{S}_i = \pm S_i$, this term reads

$$-\frac{1}{2N}\sum_{\{S\}}\sum_{i,k} S_k[\rho(S_1, \ldots, S_i, \ldots, S_N, t) - \rho(S_1, \ldots, -S_i, \ldots, S_N, t)],$$

but this term is clearly zero unless $k = i$ and when $k = i$ it reduces to

$$-\frac{1}{N}\sum_{\{S\}}\sum_{i} S_i \rho(S_1, \ldots, S_N, t) = -\langle m\{S\}\rangle.$$

Recalling that $S_i^2 = 1$, the second term is

$$\frac{1}{2N}\sum_{\{S\}}\sum_{i,k} S_i S_k \sum_{\bar{S}_i} \tanh(Jm\{S\})\rho(S_1, \ldots, \bar{S}_i, \ldots, S_N, t) = \frac{1}{2N}\sum_{\{S\}} \tanh(Jm\{S\}) \sum_{i,k} S_k S_i [\rho(S_1, \ldots, S_i, \ldots, S_N, t) + \rho(S_1, \ldots, -S_i, \ldots, S_N, t)],$$

where again the sum over $\bar{S}_i$ has been written explicitly. For terms in the sum with $i \neq k$ the dummy variable $-S_i$ in the second term can be replaced by S_i with a corresponding change of sign which leads to a cancellation of the two terms in the square brackets. For $k = i$ the two terms in the square brackets add and the result is

$$\sum_{\{S\}} \tanh(Jm\{S\})\rho(S_1, \ldots, S_N, t) = \langle \tanh(Jm\{S\})\rangle.$$

When it is all reinserted in Eq. 3.71 one finds,

$$\Gamma \frac{d\langle m\{S\}\rangle(t)}{dt} = -\langle m\{S\}\rangle(t) + \langle \tanh(Jm\{S\})\rangle. \tag{3.72}$$

But this is not quite Eq. 3.35. The step which leads from $\langle \tanh(Jm\{S\})\rangle$ to $\tanh(J\langle m\{S\}\rangle)$ is non-trivial. In fact, while the rest of the derivation thus far has been completely general, this step is not. What will permit it here is the fact that we are considering a system with an interaction of infinite range – a mean-field theory. In such a theory, fluctuations are negligible and consequently

- the mean of any function of an extensive variable equals the function of the mean of the variable:

$$\langle f(m)\rangle = f(\langle m\rangle). \tag{3.73}$$

Within the mean-field theory, the dependence of the probability distribution on the spin variables is only via $m\{S\}$. Moreover, it is very sharply peaked about an equilibrium value of m and as $N \to \infty$ it becomes essentially a delta function. Rather than embark on a detailed confirmation of these assertions, whose proper place is in a text on statistical mechanics, we shall restrict ourselves to a few remarks of relevance:

- In Eq. 3.31 we read that the energy is just $-\frac{1}{2}Nm^2$, i.e., a function of m only. Hence the Gibbs weight, $\exp(-\beta E)$, will also be a function of m only.

- Any sum over the configurations of the system of a function which depends only on m can be written as

$$\sum_{\{S\}} f(m) = \sum_{m} f(m) \exp[Ns(m)] \tag{3.74}$$

  where $\exp[Ns(m)]$ is the number of states with magnetization Nm – $Ns(m)$ is the system's entropy.

- The mean of any function of m will be computed as

$$\langle f(m) \rangle = \sum_{\{S\}} f(m) \exp(-\beta E) = \sum_{m} f(m) \exp\left(-N\beta[m^2 - Ts(m)]\right). \tag{3.75}$$

  This last sum is computed by the method of the saddle-point, see e.g., Section 3.4.2.

- If f has a finite limit as $N \to \infty$, then the sum over m is simply $f(\bar{m})$, where $\bar{m}$ is the minimum of the exponent in Eq. 3.75.

- In particular, $\bar{m}$ is the mean of m, which is just Eq. 3.73.

3.6.3 The minimization of the dynamical free-energy

The question is about the sign of the time derivative $\dot{F}$, in which the time dependence enters through ρ. Note first that

$$\dot{F} = \sum_I \dot{\rho}_I E(I) + T \sum_I \dot{\rho}_I (\ln \rho_I + 1). \tag{3.76}$$

Since ρ is normalized, the last term on the right hand side vanishes.

Inserting the master equation, Eq. 3.41, for $\dot{\rho}$ one arrives at the following chain of equations:

$$\begin{aligned}
\dot{F} &= \sum_I \dot{\rho}_I [E(I) + T \ln \rho_I] \\
&= \sum_I \sum_{J \neq I} [w(I \mid J)\rho_J(t) - w(J \mid I)\rho_I(t)][E(I) + T \ln \rho_I] \\
&= \sum_I \sum_J [w(I \mid J)\rho_J(t) - w(J \mid I)\rho_I(t)][E(I) + T \ln \rho_I] \\
&= \sum_I \sum_J [w(J \mid I)\rho_I(t) - w(I \mid J)\rho_J(t)][E(J) + T \ln \rho_J]
\end{aligned}$$

$$= \frac{1}{2} \sum_{I,J} [w(I \mid J)\rho_J(t) - w(J \mid I)\rho_I(t)]\{[E(I) + T \ln \rho_I] - [E(J) + T \ln \rho_J]\}. \tag{3.77}$$

The transition from the second to the third line is justified by the fact that the term with $I = J$ vanishes anyway; from the third to the fourth it is a simple renaming of the dummy variables and from the fourth to the fifth we simply take the average of the third and fourth expressions, which are equal to each other.

Next we introduce the only element of substance about the dynamics, *detailed balance*:

$$w(I \mid J) = w(J \mid I) \exp\left(-\frac{E(I) - E(J)}{T}\right).$$

When this is substituted into the right hand side of Eq. 3.77, one arrives at:

$$\begin{aligned}
\dot{F} = \; & \tfrac{1}{2} \textstyle\sum_I \sum_J w(J \mid I) e^{-E(I)/T} \{e^{E(J)/T} \rho_J(t) - e^{E(I)/T} \rho_I(t)\} \\
& \{[E(I) + T \ln \rho_I] - [E(J) + T \ln \rho_J]\}.
\end{aligned} \tag{3.78}$$

Note that the first term in the first curly brackets is equal to the exponential of the second term in the second curly brackets and similarly for the last two terms. The first two factors in each term in the double sum are positive. Hence, the sign of each term is determined by the product of the last two factors, which is either negative or zero.

We have proved, therefore, that F monotonically decreases and that it can stop decreasing only if the distribution reaches equilibrium at the Gibbs distribution.

3.6.4 Legendre transform for the free-energy

The beauty and power of the Legendre transform in very useful reformulations of thermodynamic information are exposed in Callen's monograph[18] and in the articles of DeDominicis and Martin[20]. Here we will present a bare minimum for local completeness.

The usual computation of the free-energy, Eq. 3.46, provides a free-energy which is a function of the externally controlled variables h and T, i.e., $f(h,T)$. This function contains all the thermodynamic information related to the system. In particular, given this function the magnetization per spin m is directly derived as

$$m(h,T) = -\frac{\partial f}{\partial h}. \tag{3.79}$$

If the information about the system is expressed in terms of the quantities m and T, for example, then f is no longer as useful, despite the fact that h can be eliminated from Eq. 3.79 in terms of m and T, and substituted in the free-energy f. But because this equation is a differential equation, information gets lost in the process.

A technique for expressing a function in terms of its tangents was developed by Legendre. It consists of defining a new function

$$g(m,T) = f(h(m,T),T) + mh(m,T), \tag{3.80}$$

where $h(m,T)$ is indeed derived from Eq. 3.79. This is manifestly a function of m and T. What remains to be shown is that in this process no information has been lost. To see this, note that

$$\frac{\partial g}{\partial m} = h,$$

and the other terms cancel each other by virtue of Eq. 3.79. The function f can now be reconstructed from g by the same process, namely

$$f(h,T) = g(m(h,T),T) - hm(h,T).$$

In other words, if $g(m,T)$ is given one can eliminate m as function of h and T and arrive back at $F(h,T)$.

For example, if

$$f(h,T) = -\frac{J}{2T}h^2$$

then

$$m(h,T) = \frac{J}{T}h.$$

Solving for h one finds

$$h(m,T) = \frac{T}{J}m.$$

Inserting in Eq. 3.80 one finds:

$$g(m,T) = \frac{T}{2J}m^2.$$

Bibliography

[1] K. Huang, *Statistical Mechanics* (John Wiley & Sons, NY,1963).

[2] F.R. Gantmacher, *The Theory of Matrices* (Chelsea Publishing Company, NY, 1959), Vol. II.

[3] P. Peretto and J.-J. Niez, Stochastic dynamics of neural networks, *IEEE Transactions: SMC* **16**, 73(1986).

[4] S. Amari, Learning patterns and pattern sequences by self-organizing nets of threshold elements, *IEEE Trans. Comput.*, **21**, 1197(1972).

[5] W.A. Little, The existence of persistent states in the brain, *Math. Biosci.*, **19**, 101(1974).

[6] E. Ising, Bertrag zur Theorie des Ferromagnetismus, *Zeitschrift Für Physik*, **31**, 253(1925).

[7] D.C. Mattis, *The Theory of Magnetism* (Harper & Row, NY, 1965).

[8] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.* **21**, 1087(1953).

[9] K. Binder ed., *Monte Carlo Methods in Statistical Physics* Vol. I and II (Springer-Verlag, Heidelberg, 1979 and 1984)

[10] O.G. Mouritsen, *Computer Studies of Phase Transitions and Critical Phenomena* (Springer-Verlag, Heidelberg, 1984).

[11] M. Creutz, *Quarks, Gluons and Lattices* (Cambridge University Press, Cambridge, 1983).

[12] R.J. Glauber, Time dependent statistics of the Ising model, *J. Math. Phys.* **4**, 294(1963).

[13] R. Peierls, On Ising's model of ferromagnetism, *Proc. Camb. Phil. Soc.* **32**, 477(1936).

[14] G.H. Wannier, *Elements of Solid State Theory* (Cambridge University Press, London, 1960).

[15] E. Bienenstock, F. Fogelman Soulié and G. Weisbuch, *Disordered Systems and Biological Organization* (Springer-Verlag, Berlin, 1986)

[16] M.A. Cohen and S. Grossberg, Absolute stability of global pattern formation and parallel memory storage by competing neural networks, *Transactions IEEE*, **SMC-13**, 815(1983)

[17] J.J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, *Proc. Natl. Acad. Sci. USA*, **81**, 3088(1984)

[18] H.B. Callen, *Thermodynamics* (John Wiley & Sons, NY, 1960).

[19] E.T. Copson, *Asymptotic Expansions* (Cambridge University Press, London, 1965).

[20] C.T. DeDominicis and P.C. Martin, Stationary entropy principle and renormalization in normal and superfluid systems I. Algebraic formulation, *J. Math. Phys.* **5**, 14(1964).

[21] P. Peretto, Collective properties of neural networks, *Biol. Cybern.*, **50**, 51(1984)

[22] E. Goles, Positive automata networks, in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman Soulié and G. Weisbuch eds. (Springer-Verlag, Berlin, 1986) and E. Goles and G.Y. Vichniak, Lyapunov functions for parallel networks, in J.S. Denker ed. (AIP Conf. 151, NY, 1986).

4

Symmetric Neural Networks at Low Memory Loading

4.1 Motivations and List of Results

4.1.1 Simplifying assumptions and specific questions

By now the reader may feel that the issue of attractors has been somewhat belabored. This is unavoidable, given that attractors and their close relatives are the main message we bring from physics. The accompanying taste and style will ultimately be judged by the results and the clarity that can be produced as well as by qualitative novelty. This task is undertaken in this chapter and in Chapter 6, which are devoted to an exposition of tools and results. Admittedly, the results are transparent in situations which are not fully realistic. But once the model is formulated, with its extreme simplifications, it becomes clear that the qualitative nature of the difficulties involved in disentangling its properties are of a similar order to those which one would expect in a highly interactive network of realistic neurons. Whether cognition can be accounted for by either the dynamics of the realistic or that of the simplified network is, of course, a problem of a different magnitude. But if an aesthetic criterion is involved in selecting a mechanism for higher mental functions, this chapter should go a long way toward reinforcing ANNs' claim for the role.

Let us recapitulate the simplifying assumptions which will be most pertinent for the technical manipulations of the rest of this chapter:

- Neurons are discrete two-state elements.
- The dynamics of a neuron is stochastic threshold dynamics – Eq. 2.19, with a noiseless limit given by Eq. 2.12.
- The network is fully connected, i.e., any two neurons in the network are equally likely to be connected by a synapse.
- The network is comprised of a very large number of neurons. In the following analysis this number will be infinitely large. The results approximate well even networks with a couple of hundred neurons. They are indistinguishable from simulations on networks with a couple of thousand neurons.

The above is a list of general simplifying features. It should be appreciated that under these conditions the network is still

- a stochastic dynamical system of a large number of highly nonlinear elements which strongly interact with each other with full feedback.

Such a system is not usually expected to behave in a controllable way. In our context, controllable means that

- the dynamics has simple asymptotic properties;
- it be relatively insensitive to initial conditions, to allow for associative recall;
- the set of attractors be simply related to the synaptic matrix, to allow for an intuitive identification of memory with network organization and learning.

As the network stands, with the above list of simplifications, one could expect chaotic, runaway trajectories, amplifying initial differences beyond recognition.

A priori the analytic task appears essentially as awesome when the following feature is added:

- The synaptic connectivity is symmetric

$$J_{ij} = J_{ji}. \tag{4.1}$$

  i.e., the influence of neuron j on neuron i is equal to the influence of neuron i on j.

This last condition is perhaps the most instrumental single element in the recent wave of interest and progress in the field of neural networks[1].

This modeling feature has been frequently criticized by neurobiologists, casting doubt on the entire modeling enterprise of ANN's. Indeed, all the channels of communication in biological neural networks are unidirectional and there is no reason to assume that neuron i is affected by neuron j in the same way as neuron j is affected by neuron i. It is not even guaranteed that if there exists a connection from neuron i into neuron j, then the back propagating connection exists at all.

On the other hand, it is this assumption which ushered in the analogies between neural networks and the *statistical mechanics* of *frustrated disordered systems* at a time when the investigations of such systems in physics reached a certain maturity. It made available a complete arsenal of concepts, insights and techniques. Those, in turn, have allowed a full analysis of the model, leading to results and intuitions which go way beyond the symmetric constraint. In fact we shall find in Section 7.1.1 a remarkable robustness of most of the essential features to the relaxation of the synaptic symmetry. We will, therefore, adopt the attitude that symmetric J_{ij}'s are 'a clever evolutionary step backwards'.[1] Finally, in the context of artificial intelligence, where the construction of the connectivity matrix is at our disposal, models with symmetric synaptic connections are perfectly legitimate.

Synaptic symmetry has very dramatic consequences in simplifying the symptotic dynamical behavior – asynchronous deterministic dynamics of a symmetric system has only fixed points. Noisy dynamics is also governed by attractors in the space of *probability distributions*, as was discussed in detail in the previous chapter. See in particular Section 3.5. The question then reduces to the control over the relation between the synaptic matrix and the collection of attractors (attractor states, or attractor distributions). It is reasonable to ask:

- Can a synaptic matrix be designed to ensure that a prescribed set of network states be attractors?
- If the answer to the first question is positive, can one classify the accompanying *spurious attractors*, i.e., those that parasitically accompany a set of stored attractors (see e.g., Section 2.3.1)?

[1] This is a paraphrase of a comment by Dr. G. Toulouse.

- If spurious states are considered undesirable, can they be controlled?

4.1.2 Specific answers for low loading of random memories

We do not have general answers to these questions. General answers will have to deal with worst cases which in turn will be related to specific properties of particular sets of memories. In the absence of information on any particular form of imprinted memories (attractors) and due to the communal inclination of physicists to deal with *typical* situations rather than with isolated worst-cases, the scope is restricted slightly further. Then, a surprising wealth of detailed answers can be provided[2] and the intuitions gained thereby go way beyond the restricted framework. The additional restrictions are:

- The memories (attractors) to be stored are **random**. (Some effects due to correlations among memories will be discussed in Chapter 8).
- The number of memories **intentionally** stored in the network is small compared to the number of neurons in the network. (Memory saturation effects will be deferred to Chapter 6).

There are storage prescriptions – expressions for synaptic coefficients in terms of memorized patterns – which are less sensitive to these last restrictions[3,4]. We discuss them in Section 4.2.3, below. Most of our discussion will be restricted to *local* prescriptions in which the synaptic efficacy connecting two neurons depends only on the expected activity of the connected neurons. The first result is that

- for a typical set of randomly chosen N-bit words, there exists a storage prescription – a matrix of synaptic efficacies – which ensures that these words will be attractor states of the network in the absence of noise;
- in the presence of noise, provided it is not too high, the same prescription provides *attractor distributions* in which only states *close* to one of the stored patterns have non-zero probabilities.

Usually, however, there will be many other attractors. These are the *spurious states*. They will be discussed in due course.

Attractors can be effectively observed by following the temporal evolution of the *overlaps* of the instantaneous network state with all the memorized patterns. As has been indicated in Section 1.4.2, overlaps are linearly related to distances between patterns. A system storing p patterns will therefore be characterized by p overlap parameters defined as:

$$m^\mu(t) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu S_i(t). \tag{4.2}$$

These are variables with values between −1 and +1. If the state of the network is uncorrelated with a given pattern, the corresponding m^μ will vanish.[2] If it is fully correlated with some pattern, that particular overlap will be unity – a macroscopic value. As the network enters an attractor the values of the overlaps become constant.

> The statement about the noisy case may not be quite lucid to the reader who has chosen to skip Chapter 3. It implies that since in the presence of noise there are no fixed states, because the network can make a step in the 'wrong' direction with a finite probability, the attractor concept has to be extended. What replaces the fixed points are dynamic trajectories that remain trapped near one of the stored patterns which is selected by the initial conditions. Following a short transient, the network will hover around the attractor state. This phenomenon can be characterized by the temporal average of the *overlap*, which measures similarity, of the consecutive network states with the stored pattern.
>
> For an *attractor distribution*, the average overlaps will converge to some non-zero asymptotic values. In particular, the result stated above implies that there will be trajectories which will make small fluctuations about each one of the *individual* stored patterns as well as about the spurious states. Which one it will be depends again on the initial conditions.

The specific form of the storage prescription that we shall concentrate on will be:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu, \qquad J_{ii} = 0. \tag{4.3}$$

[2] It will strictly vanish only in the limit of infinite N. Typically, it will be of order $N^{-1/2}$.

The coefficient in front of the sum is there for technical reasons. See e.g., Section 3.2.4. The zeroes in the diagonal of the matrix J express the fact that neurons do not usually synapse onto themselves. Its technical implications have been explained in Section 3.5.1. There are p patterns that are memorized by this prescription. These are ξ_i^μ for $\mu = 1, \ldots, p$. We will assume that the performance of the network can be analyzed keeping the synaptic values fixed, or *quenched*. This implies that during a typical retrieval time the changes that may occur in synaptic values are very small. It should be kept in mind that this is a particular option and the opposite attitude has been also proposed. See e.g., ref. [5].

The patterns are memorized in the sense that in the noiseless situation every one of the network configurations

$$S_i = \xi_i^\mu \quad \text{for} \quad i = 1, \ldots, N \tag{4.4}$$

for every one of the p patterns labelled by μ, is a fixed point of the dynamics. This is the case for a typical set of N-bit patterns $\{\xi^\mu\}$ in which each bit is chosen at random, equal probability, to be either $+1$ or -1. Note that the *locality* mentioned above is expressed by the fact that each pattern contributes to synapse ij a term which is the product of the corresponding ξ_i and ξ_j. These are exactly the activities of the neurons i and j when the network is in a state identical to the pattern.

It is this last feature which explains the nomenclature 'Hebbian learning rule' attached to the synaptic prescription Eq. 4.3. Hebb has suggested that learning may take place by synaptic modifications which reflect the activities of the pre- and post-synaptic neurons under the influence of the stimulus to be learned. According to Hebb[6]

> Let us assume then that the persistence or repetition of a reverberatory activity (or "trace") tends to induce lasting cellular changes that add to its stability.... *When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.* [p. 62]
>
> When one cell repeatedly assists in firing another, the axon of the first cell develops synaptic knobs (or enlarges them if they already exist) in contact with the soma of the second cell. [p. 63]

The translation of this wise speculation into the formula 4.3 is a living proof of associative memory.

The prescription 4.3 has some attractive features and some drawbacks. Briefly,

- It is local.
- It is additive, which is reminiscent of learning.
- It is generic, in the sense that it involves no knowledge about the memories.
- It is easily generalizable.

The main limitation, perhaps, is the equal intensity with which all the patterns are imprinted. It leads to excessive symmetry between all the stored memories. This degree of freedom has been put to creative use in describing short term memory which learns and forgets. See e.g., Section 6.6.1.

Synaptic example

To store the two patterns $\{\xi^1\}$ and $\{\xi^2\}$ in a network of five neurons, where

$$\xi^1 = \begin{pmatrix} +1 \\ -1 \\ +1 \\ +1 \\ -1 \end{pmatrix} \quad \xi^2 = \begin{pmatrix} +1 \\ +1 \\ +1 \\ -1 \\ -1 \end{pmatrix}, \tag{4.5}$$

according to the prescription 4.3, one could use the matrix:

$$J_{ij} = \begin{pmatrix} 0 & 0 & 2 & 0 & -2 \\ 0 & 0 & 0 & -2 & 0 \\ 2 & 0 & 0 & 0 & -2 \\ 0 & -2 & 0 & 0 & 0 \\ -2 & 0 & -2 & 0 & 0 \end{pmatrix}, \tag{4.6}$$

where the prefactor $\frac{1}{5}$ has been omitted.

If the network is in either of the two network states $S_i = \xi_i^1$ or $S_i = \xi_i^2$, the matrix 4.6 induces, respectively, the following PSP's in each neuron

$$h^1 = \begin{pmatrix} 4 \\ -2 \\ 4 \\ 2 \\ -4 \end{pmatrix} \quad h^2 = \begin{pmatrix} 4 \\ 2 \\ 4 \\ -2 \\ -4 \end{pmatrix}. \tag{4.7}$$

If these PSP's enter into the dynamics, Eq. 2.12, they reproduce exactly the two states ξ^1 and ξ^2.

One should exercise some restraint with this type of experiments because unless the number of neurons is relatively large the probability of encountering atypical situations becomes appreciable. The pitfalls should become clear as we proceed with the analysis.

The results will now be spelled out in more detail. First for the low-noise situation and then for substantial noise levels. Derivations are postponed to later sections.

4.1.3 Properties of the noiseless network

The results to be listed below are for a **large** network storing **few** patterns with synaptic efficacies given by Eq. 4.3. The dynamical process followed by single neurons is given by Eq. 2.12. The updating sequence for the network is either asynchronous and random, or synchronous, see e.g., Sections 2.2.2 and 2.2.3.

- Both synchronous and asynchronous dynamics are governed by a landscape function. See e.g., Section 2.3.1.
- The exact p patterns are fixed points of synchronous or asynchronous dynamics.
- Each memorized pattern is doubled.

With each stored pattern, the state for which each neuron is in the opposite state to its activity in the pattern is also a fixed point.

- In both dynamical processes the stored patterns, as well as their reversed twins, are **absolute** minima of an energy that is the (Lyapunov) landscape function for the asynchronous dynamics.
- Each memory has an enormous basin of attraction.

This is expressed in Figure 4.1. We plot the temporal development of the overlaps of the instantaneous network state, Eq. 4.2, with some of seven stored random patterns, in a network of 400 neurons. The initial overlap with pattern number one is 0.33, which implies that initially some 33% of the neurons are out of alignment with the pattern. Yet the pattern is quickly retrieved. Recall that at most 50% can be misaligned before the two patterns become totally uncorrelated. The overlap then is zero.

- On storing p random patterns by means of the matrix 4.3 the network is found to have a large number of *spurious attractors*, often referred to as *spurious states*.

A spurious state has large overlaps with several memorized patterns. In Figure 4.2, we show the simulation of the dynamical development of a network state which started with a significant overlap on three memorized patterns. There are N=400 neurons and p=7 patterns, memorized via Eq. 4.3. The stable state has large and equal overlaps, about 0.5, with three patterns. The overlaps with the other 4 patterns are seen to fluctuate about 0. The equality of the three large overlaps has lead to the nomenclature *symmetric mixture* for such attractors. Mixture states of this type are the states which have been invoked by Hoffman to capture the schizophrenic disturbance, Section 2.3.4.

- There is an attractor for **every** symmetric mixture of the embedded patterns.

This implies that there are attractors for which any subset of n, out of the p, overlaps will be non-zero and of equal magnitude. The sign of each of the n overlaps can be chosen arbitrarily. This is due to the fact that for every pattern attractor the reversed state is also an attractor. Such a mixture can be expressed as

$$\begin{aligned} m^\mu &= \pm m_n \quad \text{for} \quad \mu = 1, \ldots, n \\ m^\mu &= 0 \quad\;\; \text{for} \quad \mu = n+1, \ldots, p. \end{aligned} \tag{4.8}$$

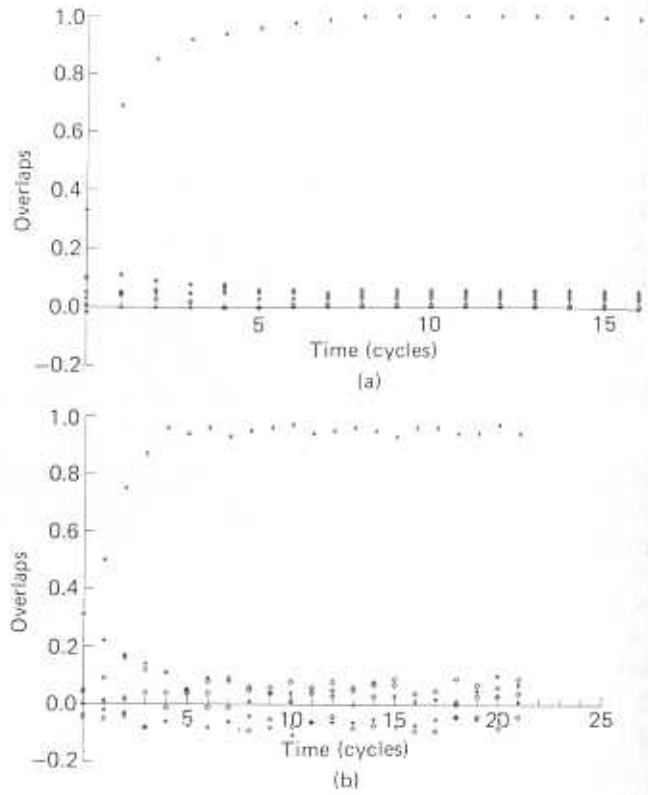


Figure 4.1: Retrieval attractor: Time development of overlaps of the network state with p=7 stored patterns. Asynchronous dynamics and N=400. (a) T=0, (b) T=0.5. The initial state has a large overlap only with memory no. 1, m^1=0.33. The overlap m^1 increases rapidly to 1. The overlaps with other patterns fluctuate around 0.

Note that the n non-vanishing overlaps can be chosen as the first of the p overlaps due to the complete symmetry between the stored memories.

In Figure 4.3 we present the temporal stability of a symmetric 5-mixture in a network of 400 neurons storing 7 patterns. In this attractor, the overlaps with the first three patterns are positive and with the last two negative. They are all approximately equal in magnitude to $\frac{3}{8}$. See e.g., ref. [2].

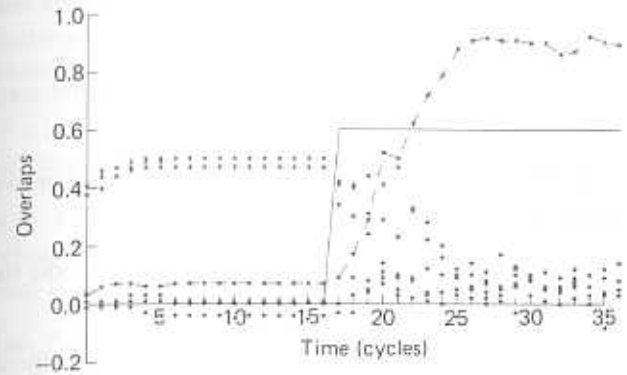


Figure 4.2: *Spurious attractor* – symmetric 3-mixture. Temporal development of overlaps from an initial state with three large overlaps. N=400, p=7, asynchronous dynamics. For t=0 through 15, T=0, and one can see 3 large stable overlaps. All overlaps with 4 other patterns are near 0. Then T=0.6 and 3-mixture is destabilized. See discussion in Section 4.1.4, below. Full curve – temperature, dashed curve – winning pattern.

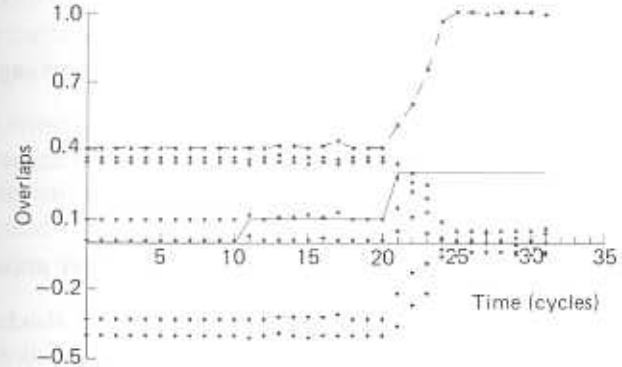


Figure 4.3: Spurious attractor – a symmetric 5-mixture with three positive and two negative overlaps. Simulation of N=400, p=7 asynchronous dynamics. Full curve – temperature, dashed – winning pattern, see e.g., Section 4.1.4.

- There are 2^n symmetric mixture fixed points with $n(\leq p)$.

This is due to the fact that every overlap can come in two signs.

- A network storing p patterns has a total of 3^p spurious state which are symmetric mixtures.
- Fixed points corresponding to even symmetric mixtures are unstable.

Any departure from an unstable fixed point will drive the network away from that fixed point.

- Fixed points which are symmetric mixtures of an odd number of patterns are all stable, and are hence attractors.
- The even symmetric mixtures are two-cycles in synchronous dynamics[7].

Such states have one group of neurons which flip over in every updating cycle while all other ones remain fixed[7,8].

- Stable attractors are the same for synchronous and asynchronous dynamics.
- The spurious states are all metastable states, i.e., their energies are higher than those of the pure patterns and their reversed companions.
- In addition to the symmetric spurious states there are asymmetric ones.

Asymmetric spurious states are more difficult to classify than the spurious ones. Example of such states will be presented in Section 4.5.3.[3]

4.1.4 Properties of the network in the presence of fast noise

Fast noise is represented by a Glauber type (heat-bath) stochastic dynamical process. This process is defined by Eq. 2.19, and is discussed in greater detail in Section 3.2.2. Updating can, of course, be either synchronous or asynchronous.

- Both synchronous and asynchronous dynamics obey *detailed balance* and hence both relax the network toward a Gibbs distribution. Hence both types of dynamics will be governed by a noisy landscape function – the *free-energy*.

[3]These have been first noticed by M.A. Virasoro and N. Parga (unpublished).

- Above a critical noise level $T=T_c=1$, the dynamical behavior of the network is ergodic and consequently the average overlaps with all stored patterns vanish. This is a *paramagnetic* phase.
- At $T=1$ the network undergoes a *phase transition*, provoking a qualitative modification in the shape of the landscape.

While for $T > 1$ there is but a single minimum, at which all overlaps vanish, below this temperature many additional minima appear. Every one of them is, of course, an attractor.

- Below $T=1$ the distribution with all vanishing overlaps (*paramagnetic state*) becomes a top of a hill and is unstable.
- Slightly below $T=1$ the only new valleys are *attractor distributions* which have a single non-vanishing average overlap.

In each of these valleys there is a non-vanishing overlap either with one of the stored patterns or with one of the reversed patterns. The landscape takes on the form of a mount surrounded by $2p$ valleys, beyond which it rises again indefinitely.

- As the noise decreases, the valleys deepen and move away from the central mount, which is the state of zero overlaps – the *paramagnetic* phase.

In other words, the average overlaps become greater and the attractors more stable. Just below $T=1$ these valleys are very shallow and very close to the center, where all overlaps vanish. (The transition is of 'second order' in thermodynamic parlance.) Phenomena of this type can be observed in Figure 3.7, in the transition from curve 2 to curve 3.

- Below T_c, all the symmetric spurious states of the noiseless case become *extrema* in the free-energy landscape. They are either minima, maxima or saddle-points.
- For $T > 0.461$ only the distributions corresponding to pure and reversed patterns are stable, and hence attractors.

Convergence into such an attractor implies that the network keeps wandering, but in a restricted manner: one overlap fluctuates about a non-zero average, while all the others fluctuate about zero. An example

of such behavior is shown in Figure 4.1(b), where a network of 400 neurons storing seven patterns is simulated at T=0.5. The network is initialized in a state with a larger overlap on one of the patterns and one observes the overlap with this pattern increase and then fluctuate about $m \approx 0.9$, while all other overlaps remain very small.

- As T keeps decreasing below T=0.461 spurious states become successively stable.

First, the symmetric three mixtures become stable and begin to attract. This is shown in a simulation in Figure 4.2. If the temperature is raised again, the symmetric 3-mixture becomes destabilized and one overlap takes over. The others decay to zero. This also is shown in Figure 4.2, when after 15 time steps the temperature has been raised to 0.6.

- As the temperature is lowered further, more and more of the symmetric odd mixtures become attractors. Lower mixtures become stable at higher temperature. For example, the 5-mixture becomes stable below T_5=0.385 and the 7-mixture below T_7=0.345.

- The pure pattern attractors remain the absolute minima in the landscape all the way down to T=0. They always have the largest basins of attraction.

- At T=0.452 asymmetric spurious states first show up.

- In the limit $T \to 0$ the very rich picture described in the previous section is recovered.

- When all spurious states are eliminated by raising the noise level to above 0.461, the *retrieval quality*, the magnitude of the overlaps with the pure patterns, is m=0.97, i.e., 1.5% misaligned neurons. This is what makes noise so useful and provokes metaphors such as the one of Section 2.3.4.

The rest of this chapter will be devoted to the technical aspects related to the results listed in the last two sections. In the process, many of the concepts that have been introduced will be clarified.

4.2 Explicit Construction of Synaptic Efficacies

4.2.1 Choice of memorized patterns

We are now ready to complete the last part in the construction of a model of the neural network as an associative memory. Memory in this context is the ability to recreate a firing pattern, which has previously propagated through the network. This firing pattern is the *network representation* of some *external stimulus*, and it has been *stored* in the network by some, yet unspecified, process of *learning*. Since the stored patterns are *network states* they are described by N-bit words which specify the distribution of neural activities when the network enters an attractor corresponding to the stored pattern, or when it retrieves a memory. The N-bit word representing a pattern will denoted by

$$\{\xi^\mu\} \equiv (\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu). \tag{4.9}$$

where μ labels the different memories and $\xi_i^\mu = \pm 1$, chosen randomly, independently with equal probability, namely,

$$\Pr(\xi_i^\mu) = \tfrac{1}{2}\delta(\xi_i^\mu - 1) + \tfrac{1}{2}\delta(\xi_i^\mu + 1) \tag{4.10}$$

where the δ-function is equal to one when its argument is zero, and to zero otherwise. This choice of the ξ_i^μ's implies that:

- The stored patterns are *unbiased*, namely, in a large network

$$\langle\langle \xi_i^\mu \rangle\rangle \equiv \frac{1}{N}\sum_i \xi_i^\mu = 0. \tag{4.11}$$

  - When the network is in a state corresponding to a pattern, $S_i = \xi_i^\mu$, which is when that pattern is retrieved, about half of the neurons in the network are active.

Such high mean activity level is incompatible with the much lower levels of activity found empirically in real networks[9]. The modifications which are required in order to remedy this apparent discrepancy will be discussed in Section 8.1.

- Different stored patterns are *uncorrelated*, namely in a large network

$$\langle\langle \xi_i^\mu \xi_i^\nu \rangle\rangle \equiv \frac{1}{N}\sum_i \xi_i^\mu \xi_i^\nu = 0, \qquad \mu \neq \nu. \tag{4.12}$$

This excludes from the discussion similarities between the memorized stimuli that may show up as correlations between their network representations. The generalization to the storage and retrieval of correlated patterns is the subject of Chapter 8.

4.2.2 Storage prescription – "Hebb's rule"

Hebb's hypothesis, as cited in Sec. 4.1.2, can be given the following formal interpretation[10,11]:

- Past network activity modifies the synaptic efficacy

$$\Delta J_{ij} \propto \langle (S_i + 1)(S_j + 1) \rangle_t, \tag{4.13}$$

  where $\langle \ldots \rangle_t$ denotes a time average over some period prior to time t.

We will not be more specific about the origin of this temporal modification because the whole issue of learning is still in an embryonic stage. The present discussion is more a suggestive motivation for the final expression, Eq. 4.3, than a phenomenological description of a learning process.

This formulation indicates that J_{ij} is enhanced when the network is in a persistent state in which neurons i and j are active in a correlated manner. It remains unchanged if at least one of them is silent. There is also empirical evidence that the efficacy of excitatory synapses is reduced when the pre-synaptic neuron j is silent and the post-synaptic neuron i is active. This effect is included if Eq. 4.13 is modified to read

$$\Delta J_{ij} \propto \langle (S_i + 1) S_j \rangle_t. \tag{4.14}$$

This rule would lead to non-symmetric synaptic efficacies. It is therefore mimicked by the symmetric modification rule

$$\Delta J_{ij} \propto \langle S_i S_j \rangle_t, \tag{4.15}$$

which treats excitatory and inhibitory synapses on an equal footing. This rule implies that

- J_{ij} is enhanced when the two neurons are active.
- J_{ij} is enhanced when the two neurons are silent.

- J_{ij} is reduced when the pre-synaptic neuron is active and post-synaptic neuron is silent or vice versa.

There is no experimental evidence for the last two possibilities. It seems that there may even be evidence to the contrary. Experiments on the visual cortex of cats[12] indicate that the strength of excitatory synapses is unchanged when the post-synaptic neuron is silent, whatever the pre-synaptic activity.

The preceding discussion is intended to serve as an introduction to the *storage prescription* suggested by several authors[13,11,1]. When a new pattern ξ^μ is 'learned' (added to memory) the synaptic efficacies are linearly modified by the addition of a term

$$\Delta J_{ij} = \frac{1}{N} \xi_i^\mu \xi_j^\mu \tag{4.16}$$

to the pre-existing efficacies.

- Learning is therefore a process in which the network adjusts *dynamically* its synaptic efficacies to accommodate a certain pattern ξ^μ as an attractor.

The discussion of learning is deferred to Chapter 9. Eq. 4.16 is merely a statement about the change in network parameters required for the addition of a given pattern ξ^μ as an attractor.

It is interesting to point out that the synapses with fixed i and $j = 1, \ldots, N$ are all on the same, post-synaptic neuron. Eq. 4.16, together with 4.11, implies that in a large system,

$$\sum_j \Delta J_{ij} = 0.$$

In neurophysiological terms this would mean that the *total* efficacy of synapses on a given neuron i is constant. This is a welcome property that fits suggestions that the modification of synapses takes place by a local redistribution of receptors, whose total number (per neuron) is conserved[14].

A network that has stored p patterns will be connected by the matrix:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu. \tag{4.17}$$

We conclude this section with a few additional remarks concerning this storage prescription:

- All the patterns are stored with the same amplitude. Interesting new features emerge when structure is introduced into the relative weights of different memories. One such example will be discussed in Section 6.6.1. But each generalization of this simple storage prescription involves additional commitments and trades generic status for structured performance.

- Each neuron has, on the average, an equal number of excitatory and inhibitory synapses emerging from its axon. This is a result of the fact that Eq. 4.11, in addition to implying $\sum_j \Delta J_{ij} = 0$ for a large network, implies also

$$\sum_i J_{ij} = 0, \tag{4.18}$$

which is the vanishing of the sum of all pre-synaptic efficacies of the synapses of neuron j. There seems to be a wide empirical consensus, Dale's law[15], that typically neurons have either one or the other type of *efferent* (outgoing) synapses. A modification of the model 2.18, in which neurons have synaptic specificity, will be described in Section 7.3.3.

- The synaptic efficacies of Eq. 4.3 can, in principle, have $2p+1$ different values, from $-p$ to $+p$. This is sometimes referred to as the *analog depth* of the synapses. It is quantified by the number of bits required for storing this spectrum, in the present case $\log_2(2p+1)$. It is hard to imagine that there exists a neurochemical mechanism which enables such a detailed distinction between so many values of the J_{ij}'s. We return to this issue in Section 7.2.3, where we show that the network continues to perform well even under extreme coarse graining of the values of the synaptic parameters.

4.2.3 A decorrelating (but nonlocal) storage prescription

One major disadvantage of the storage prescription of Eq. 4.3 is the sensitivity of the model to correlations between the memorized patterns. A network connected by synapses of the form given by Eq. 4.3 performs best with memorized patterns which are mutually orthogonal. When the stored patterns are random, as they were chosen above, there are *on the average* no correlations. But fluctuations occur and multiply with the number of stored patterns, and eventually degrade

the performance of the network completely. Even before correlations between random patterns become a problem, this storage prescription, as it stands, does not allow for a memorization of structured data sets. For specific data structures, modifications of the local storage prescription have been proposed. These will be discussed in Chapter 8. None of these modifications can cope with a general collection of patterns. The main attraction of the storage rule, Eq. 4.3, is its locality, which has a biological flavor. But where technological applications are concerned, the storage prescription can be optimized and a general decorrelating prescription has been suggested by Kohonen et al[3], and adopted into the context of ANN's by Personnaz et al[4]. For a systematic analysis of this model see e.g., ref. [16].

In this approach, the J_{ij}'s are chosen to be solutions of the equations

$$\sum_j J_{ij}\xi_j^\mu = \lambda\xi_i^\mu, \qquad (i = 1, \ldots, N;\ \mu = 1, \ldots, p), \tag{4.19}$$

where λ is a positive constant. These equations guarantee the stability of an arbitrary set of stored patterns. One possible solution, for $\lambda = 1$, is

$$J_{ij} = \frac{1}{N}\sum_{\mu,\nu=1}^{p} \xi_i^\mu \xi_j^\nu A^{-1}_{\mu\nu} \tag{4.20}$$

where the matrix $\mathcal{A}$, whose matirx elements are $A_{\mu\nu}$, is the matrix of mutual correlations between the patterns, is defined by

$$A_{\mu\nu} = \frac{1}{N}\sum_i^N \xi_i^\mu \xi_i^\nu, \qquad \mu, \nu = 1, \ldots, p. \tag{4.21}$$

Clearly, for this solution to exist, the stored patterns must be linearly independent. This limits the number of patterns that can be stored this way to N, the number of neurons, which is much higher, though, than the number of random patterns that can be stored by the local prescription. As a matter of fact, while the inverse exists up until $p = N$, the basins of attraction of the patterns disappear at $p = N/2$[16]. The remedy is in the removal of the self-interactions.

It is the expression of the matrix $\mathcal{A}$ as the inverse of the correlation matrix that introduces the dependence of every synaptic efficacy on the activities of all other neurons in the memorized patterns. This

is the source of non-locality. The resulting synaptic matrix is manifestly symmetric. But, as we have explained in Section 3.5, this is not sufficient to ensure the existence of a landscape function, or of a thermodynamic description. To ensure those one has to impose the additional condition, namely

$$J_{ii} = 0.$$

It complicates the analysis of the network somewhat but allows it to become truly systematic[16].

This storage prescription, sometimes referred to as *pseudo-inverse* or *projection matrix*, has several advantages over the local rule (Eq. 4.3):

- It is not restricted to unbiased and uncorrelated patterns.
- It has a larger storage capacity.
- It has a better quality of retrieval – no errors.
- The local fields (PSP's) on all neurons are equal when the network is in a pattern.

4.3 Stability Considerations at Low Storage

4.3.1 Signal to noise analysis – memories, spurious states

The first question to ask about the storage prescription Eq. 4.3 is: Does it, in the absence of noise, actually stabilize the inscribed patterns – ξ^μ? Would a network in a state that is a stored pattern, $S_i{=}\xi_i^1$, be dynamically stable?[1]. The condition that a certain network state $\{S_i\}$ be dynamically stable is

$$S_i h_i > 0 \qquad (i = 1, 2, \ldots, N), \tag{4.22}$$

namely the local field must be of the same sign as the local value of the spin – the neuronal activity variable.

With Eq. 4.3, the local field at neuron i, is

$$h_i\{S\} = \frac{1}{N} \sum_{j,j\neq i} \sum_{\mu} \xi_i^\mu \xi_j^\mu S_j \tag{4.23}$$

Next, S_i is replaced by $\{\xi_i^1\}$, where the choice of the pattern is arbitrary since the patterns enter the synaptic matrix symmetrically and what holds for one pattern will hold for any other. Substituting ξ^1 for S and Eq. 4.23 for h in Eq. 4.22, the condition for the stability of bit number 1, becomes

$$\xi_1^1 h_1 = \frac{N-1}{N} + \frac{1}{N} \sum_{j,j\neq i}^{N} \sum_{\mu=2}^{p} \xi_1^1 \xi_1^\mu \xi_j^\mu \xi_j^1 > 0 \tag{4.24}$$

where the sum over μ, implied in Eq. 4.23, has been separated into two parts:

- A *signal* term, the term with $\mu = 1$, corresponding to the pattern whose stability is being investigated.
- A *noise* term which includes the contribution of all the remaining stored patterns to the PSP.

In a large system (as $N \to \infty$), the signal term is equal to unity. The noise term contains a sum of $(N-1)(p-1) \approx Np$ bits of $+1$ and -1. Since the bits of different patterns at the same site and the bits of the same pattern at different sites are uncorrelated, the sum of the Np bits in the noise term is a one dimensional *random walk* of Np steps of size unity. In such a walk, steps are taken both forward and backward. The sum – the end point of the walk – will fluctuate about zero. Its *mean square* displacement from the origin is Np. Taking into account that the sum is divided by N, one can estimate the noise term by $\sqrt{p/N}$. The local field at a typical neuron can now be written as

$$h_i \xi_i^1 = 1 + R$$

with

$$|R| \approx \sqrt{\frac{p}{N}}.$$

The conclusion is that if p is kept fixed as N is made very large, the noise becomes negligible in comparison with the signal. Since the sign of the field is equal to the sign of the spin, the subsequent state of the

network will be identical to the present state. This, in turn, implies that every pattern is a *fixed point*.

In fact, the patterns are very stable fixed points. Suppose that a finite fraction, d, of spins is flipped away from one of the patterns, randomly, then the expression for the signal will become $1-2d$ and the noise will remain unchanged. The local fields will be

$$h_i = m_0 + R$$

with

$$m_0 = 1 - 2d.$$

The noise, being of order $N^{-1/2}$ is still negligible compared to the signal which is of order unity. As a consequence, the network will immediately align itself with the pattern, even if it was rather far off to start with. In other words, the patterns have very large *basins of attraction*.

Spurious states

Despite the fact that the J_{ij}'s have been constructed to guarantee that certain specified patterns, ξ^μ, be fixed points of the dynamics, the non-linearity of the dynamical process induces additional attractors - *spurious states*. The simple *signal to noise* considerations can be extended to detect those *spurious states*. As an simple example, consider the network state

$$S_i = \mathrm{sign}(\xi_i^1 + \xi_i^2 + \xi_i^3) \tag{4.25}$$

This is a symmetric 3-mixture. It is a state that is formed out of three random patterns by a *majority rule*, namely each bit in the state is equal to the bits that form the majority among the three corresponding bits in the three patterns ξ.

First, one notes that the overlap of this state with the three patterns, for a large network, is

$$m^\nu = \langle\langle S_i \xi_i^\nu \rangle\rangle \equiv \frac{1}{N}\sum_{i=1}^{N} S_i \xi_i^\nu = 0.5, \tag{4.26}$$

for $\nu = 1, 2, 3$. This is the case because each bit in each of the three patterns has a probability of three quarters for being in the majority among the three bits which vote in the corresponding bit of S. To be

in the minority, both other bits must be of opposite sign. Thus, in S three quarters of the bits are the same as in any one of the three patterns ξ^ν and one quarter are of opposite sign.

To check the stability of this pattern, we consider the sign of the PSP on a typical neuron relative to the sign of that neuron in the *3-mixture*, i.e., we evaluate

$$S_i h_i = \text{Signal} + \text{Noise}.$$

The local field is calculated via Eq. 4.23 with S_j substituted from Eq. 4.25. The *signal* term is obtained when out of the sum over μ we select the first three values. These have a finite overlap with the 3-mixture. Using Eq. 4.26 one finds

$$\text{Signal} = S_i \sum_{\mu=1}^{3} \xi_i^\mu = 0.5\,(\xi_i^1 + \xi_i^2 + \xi_i^3)\mathrm{sign}(\xi_i^1 + \xi_i^2 + \xi_i^3). \tag{4.27}$$

The rest of the terms compose the noise which reads:

$$\text{Noise} = \frac{1}{N}\sum_j \sum_{\mu>3} \eta_1 \xi_1^\mu \xi_j^\mu \eta_j. \tag{4.28}$$

Stability is mostly threatened by the sites which have the lowest values of the *signal*. These are the sites for which in Eq. 4.27 two of the bits in parentheses are of one sign and the third of the opposite sign. In that case, the signal term is 0.5. The state of Eq. 4.25 is random relative to all $p-3$ patterns with $\mu > 3$. Hence, the noise is a random walk again, and the root mean square of its magnitude is $\sqrt{(p-3)/N}$. As $N \to \infty$ keeping p fixed, the ratio of the noise to the signal tends to zero, implying that the symmetric 3-mixture is stable. It has a rather large basin of attraction, but it is much reduced compared to the pure patterns. The reason is, of course, that now stability is guaranteed by signals of magnitude one half only. Moreover, in order to be in the *basin of attraction* of a particular 3-mixture, the initial state must have specially large overlaps with three patterns rather than with a single one. This causes a very significant reduction in the size of the basins. But they are there, as witnessed by the simulation presented in Figure 4.2.

4.3.2 Basins of attraction and retrieval times

A detailed discussion of the *basins of attraction* is deferred to Section 6.5.2. Here, a simple numerical simulation will serve as an indication that a network with synaptic connections defined in Eq. 4.3 has indeed rather large basins of attraction. In addition, it exhibits the dependence of the retrieval time on the initial distance of the stimulus from the memorized pattern.

We consider a network with N=400 and p=7. Seven patterns are chosen at random and the matrix J_{ij} is computed according to Eq. 4.3. The initial configuration (stimulus), $\{S_i^{in}\}$, is chosen to have a finite overlap with the first pattern $\{\xi^1\}$, of

$$m^1(t=0) = \frac{1}{N}\sum_i \xi_i^1 S_i^{in}.$$

The Hamming distance of the stimulus from the nearest memorized pattern is, Section 1.4.2,

$$d_H = \tfrac{1}{2}N[1 - m^1(0)].$$

The system is let evolve under noiseless sequential dynamics, and the overlaps with all the embedded patterns are followed at discrete time steps, after each updating sweep through the entire network. The results are plotted in Figure 4.4 for two initial stimuli (a) $m^1(0)$=0.35 and (b) $m^1(0)$=0.2. (Recall that an overlap of 0.3 implies 35% of misaligned spins). For both stimuli, the 'erroneous bits' in the stimulus are corrected dynamically and the network converges to a firing pattern with a large overlap with the memorized pattern with which it was initially most closely associated. The overlaps with all the other stored patterns are then negligibly small.

There is a clear difference between the equilibration (retrieval) times in the two parts of the figure. For $m^1(0)$=0.3 the network relaxes to the memory faster than for a stimulus which has only 0.2 for initial overlap. Any psychological, psychophysical or physiological restriction on this time is, *ipso facto*, related to the effective size of the basins of attraction or to the acceptable level of missing information in the input pattern. This continues a discussion that was initiated in the first chapter.

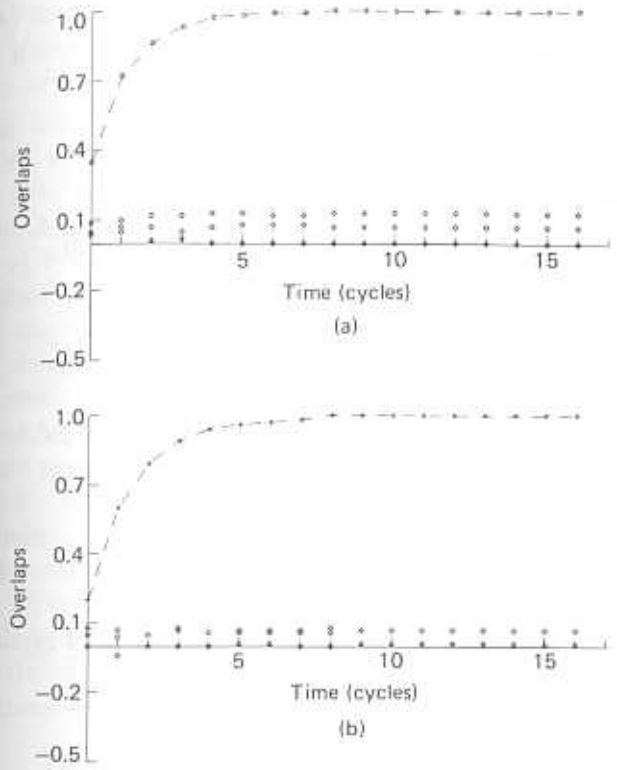


Figure 4.4: Time evolution of overlaps of a network state with the stored memories for two initial configurations. (a) m^1=0.35; (b) m^1=0.2. Simulation with N=400, p=7, and noiseless asynchronous dynamics.

Another factor which affects the time of retrieval is, of course, noise or temperature. This is illustrated in Figure 4.5. While noise increases the basins of attraction of the memories by eliminating spurious states (see e.g., the discussion in Sections 2.3.3 and 2.3.4), the price is both lower retrieval quality as well as longer retrieval times. Here again, cognitive psychological questions become intimately connected to the biological question of the degree of synaptic noise in real networks. The detailed discussion of the noisy situation will be the topic of the next few sections.

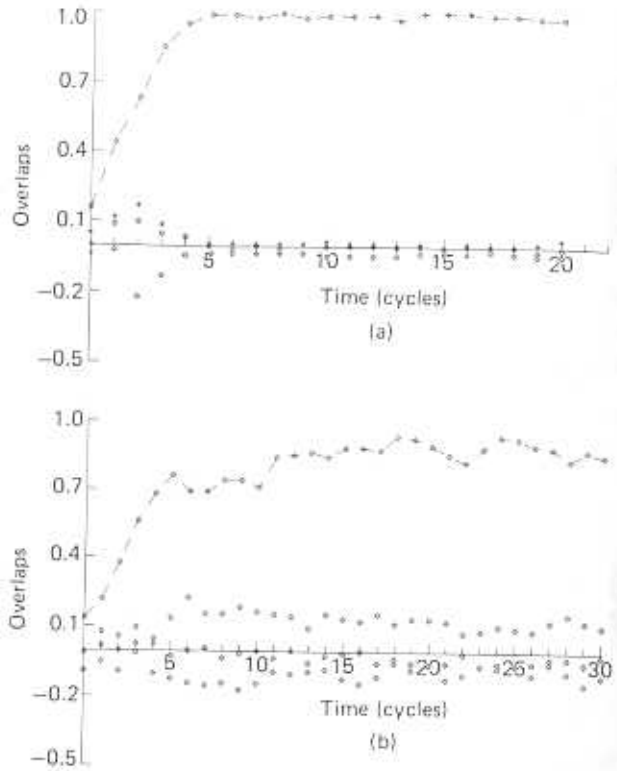


Figure 4.5: Retrieval times in the presence of synaptic noise. N=400, p=7. (a) $m^1(0)$=0.16, T=0.4; (b) $m^1(0)$=0.14, T=0.6.

4.3.3 Neurophysiological interpretation

The stability of a network state which is aligned with a stored pattern can be read in neurophysiological language. This is not to say that we have deciphered the workings of the cortex. Instead, since the model is attempting to capture some neuronal mechanisms, it is only natural that the consequences of the model be restated in the terminology of the modeled object, for perusal.

In a stable *network state* each neuron is receiving a total afferent PSP which is consistent with the reproduction of its present *neuronal* firing state. That is to say that the effect of the cooperative activity of the network feeds into every neuron that should be firing in the pattern a PSP which is above threshold and to every neuron which is *refractory* (quiescent) in the pattern a sub-threshold PSP. The stability implies that even if some of the neurons which should be firing in the pattern fail to fire, or fail to fire in time, the pattern will be reproduced.

Finally, the fact that at every step the PSP's support the firing pattern implies, if it were to be observed physiologically, that some of the neurons – the +1's in the pattern – must fire *bursts*, while others should be very quiet. At the present stage in the discussion, these bursts are expected at a rate of some 500 per second. This is because the basic cycle-time has been taken to correspond to the *absolute refractory period*. In the central nervous system of mammals, such high rates have not been observed. But bursts at a rate of 100 − 150 per second have been observed in the macaque monkey, for example refs. [17,18,19]. Histograms from a set of neurons in the parietal cortex of a macaque monkey are presented in Figure 4.6. The bursts are clearly correlated in time with visual stimuli arriving at special angles of gaze of the tested monkey. The stimuli are one second long and they start at the vertical dotted line. The fact that bursting decays before the stimulus is turned off may be taken as an indication that the bursting itself is not sustained by the stimulus, but is instead a network property. The fact that the rates are lower than the expected 500 per second is taken up in Section 7.4.

4.4 Mean Field Approach to Attractors

4.4.1 Self-consistency and equations for attractors

In this section we begin a systematic treatment of the basic model of ANN's, applying the methods and concepts of statistical mechanics developed in the preceding chapter. Let us recall that we have reduced the question of existence of correlated non-ergodic dynamical behavior of a model neural network, operating asynchronously, to the study of *equilibrium* statistical mechanics of a fully connected spin system, described by the energy function

$$E = -\tfrac{1}{2}\sum_{ij} J_{ij} S_i S_j, \tag{4.29}$$

where J_{ij} is given by Eq. 4.3, with patterns whose elements are independent random variables, which take the values ± 1 with equal probability.

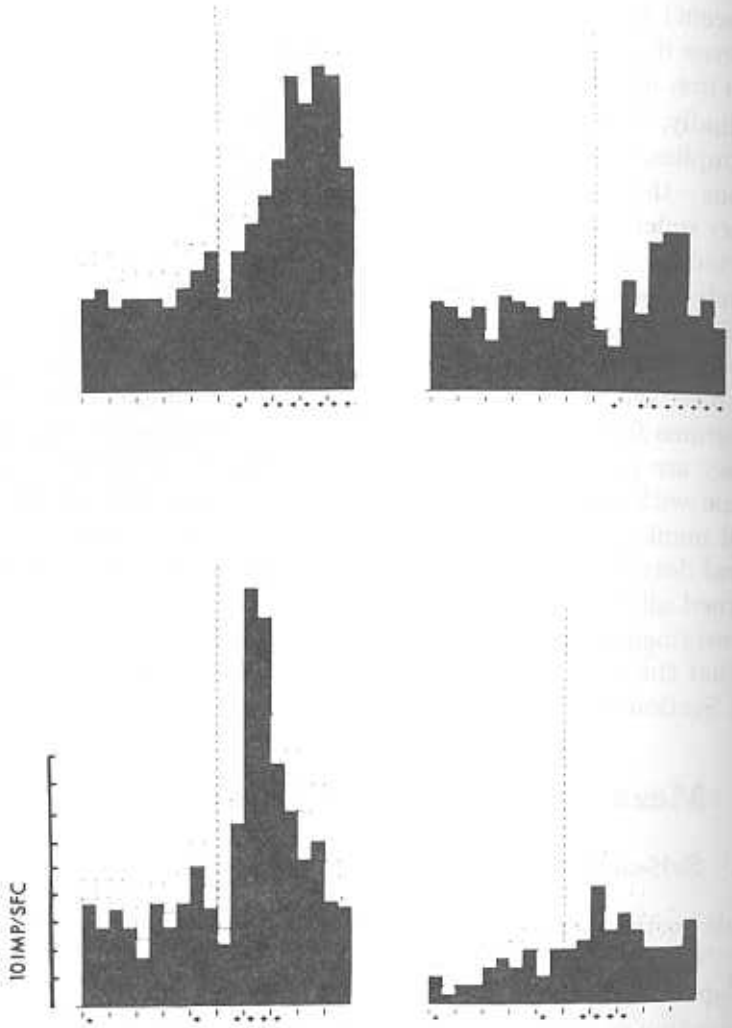


Figure 4.6: Spike histograms of two visual neurons in the parietal cortex of monkeys, exhibiting bursts of 150 spikes per second in response to visual stimuli arriving when monkey is gazing in special directions. The abcissa is a time axis, with half a second before and half a second after the stimulus which is one second long and starts at the vertical dotted line. The vertical bars are numbers of spikes per second within bins of 50 ms. (After ref. [17], by permission).

The new element which distinguishes this model from the Ising ferromagnet, discussed in the preceding chapter, is the inherent disorder introduced by the randomness of the parameters J_{ij}. In the extreme case, as the number of stored patterns becomes very large, the J_{ij}'s become totally uncorrelated Gaussian variables, except for the symmetry $J_{ij} = J_{ji}$. In this limit, Eq. 4.29 describes the *infinite range spin-glass*, known also as the Sherrington-Kirkpatrick, or SK-model[20]. In the present discussion, p/N will be small and the particular structure of the J_{ij}'s in Eq. 4.3 preserves some of the aspects of the Ising ferromagnet by way of the remaining correlations. Precisely those aspects endow the model with attributes of associative memory. It represents an interpolation between the simple, ordered ferromagnet and the extremely disordered spin-glass.

The competition between order and randomness can be envisaged as follows: Suppose that to the ferromagnetic Ising model, in which spins interact with constant strength J/N, one adds a random interaction, J_1 of zero mean and variance

$$\tilde{J}_1^2 = \frac{\tilde{J}^2}{N}.$$

The parameter which determines whether the system will preserve its ferromagnetic nature and order with all variables of equal sign, or become a spin-glass with uncorrelated spins, is

$$\Delta = \frac{J}{\tilde{J}}.$$

When this parameter is large, ferromagnetism prevails. If it is small, the system becomes a spin-glass[20].

The appropriate measure of the analog of the ferromagnetic interaction in our model is the contribution of a single neuron aligned with a pattern to the signal part of the PSP of another neuron. It is $1/N$ and so $J = 1$. The variance of the interaction between two neurons can be computed directly, using Eqs. 4.3 and 4.10. It is

$$\tilde{J}_{ij}^2 = \frac{p}{N^2}$$

which gives,

$$\tilde{J} = \sqrt{\frac{p}{N}}.$$

The ratio of the retrieval (ordering) strength to the spin-glass influence is $\sqrt{N/p}$, which implies that before p becomes of order N order prevails. When p becomes very large the system becomes a spin-glass with its enormous number of attractors[21]. We come back to these important issues in Chapter 6.

Before tackling the analysis of the model defined above, we make a short digression in order to call attention to what randomness and frustration are not. Moreover, the simple example that will be presented for this purpose will also provide us with some notational hints for the full analysis.

Digression on the Mattis model

Consider an extreme case of Eqs. 4.29, and 4.3 with $p = 1$. This is the infinite range Mattis model[22], which has been proposed as a simple solvable model of a spin-glass. The energy in this case reduces to

$$E = -\frac{1}{2N} \sum_{ij, i \neq j} \xi_i \xi_j S_i S_j. \tag{4.30}$$

In this model one has positive and negative J_{ij}'s and one might have expected *frustration* effects. Yet, the disorder is illusory, and so is the frustration. In fact, it is an instructive exercise to repeat the tabulation of the energy levels of Eq. 4.30 in the footsteps of the example of Section 2.3.6. The resulting table will have the form of columns (a) and (c) of Table 2.1, rather than that of column (b). The reason is that all the randomness and the fluctuating signs in the couplings can be absorbed in a redefinition of the spin variables. Defining

$$S'_i = \xi_i S_i, \tag{4.31}$$

the energy can be written as

$$E = -\tfrac{1}{2} \sum_{ij, i \neq j} S'_i S'_j,$$

which is just the energy of the Ising ferromagnet, Eq. 3.16. Such a transformation is called in the physics jargon a *gauge transformation*.

The properties of this model can be obtained by solving the fully connected Ising model, Section 3.2.4, and then transforming back to the original spin variables. In particular, there are two ground state configurations $S'_i = 1$ and $S'_i = -1$, which in the original variables are the two configurations: $S'_i = \xi_i$ and $S'_i = -\xi_i$. The entire discussion of the free-energy in Section 3.2.4 can be carried over to this case, provided that one identifies the magnetization parameter m as:

$$m = \frac{1}{N} \sum_i \xi_i S_i, \tag{4.32}$$

which is, of course, nothing but the mean value of S'_i. As in Section 3.4.2, the degeneracy between the two minima of the free-energy is lifted by a small external field. However, the field, which is uniform when coupled to the primed variables, now depends on the site i via:

$$h_i = h\xi_i. \tag{4.33}$$

As a spin-glass, or as an associative memory, the Mattis model is not very useful. It was extended to p=2, revealing no particularly interesting features[23], mainly because 2-mixtures are unstable, as was mentioned in Section 4.1.

Equations for self-consistency

Let us now consider a network with a specific realization of a general number ($p > 1$) of randomly chosen patterns $\{\xi^\mu\}$ ($\mu = 1, \ldots, p$). Suppose that the total local field at site i, due both to external inputs as well as to mutual influences between neurons, is h_i. Such a local field, if it persists, will produce an average temporal value for S_i equal to

$$\langle S_i \rangle = \tanh(\beta h_i). \tag{4.34}$$

This is simply because in the presence of noise the probability for $S_i = 1$ is, Eq. 2.19,

$$\Pr(S_i = 1) = \frac{\exp(\beta h_i)}{\exp(\beta h_i) + \exp(-\beta h_i)}$$

and $\Pr(S_i = -1) = 1 - \Pr(S_i = 1)$. Therefore,

$$\langle S_i \rangle = (+1) \times \Pr(S_i = 1) + (-1) \times [1 - \Pr(S_i = 1)],$$

which leads to 4.34.

But, for the field h_i to persist, the contributions to its internal part, that part which is contributed by all the other spins in the network,

must be reproduced by the average values of the all the spins connected to site i. That contribution is:

$$h_i^{int} = \frac{1}{N} \sum_{j,j\neq i}^{N} J_{ij} \langle S_j \rangle. \tag{4.35}$$

Introducing the explicit prescription for J_{ij} in terms of the memorized patterns, Eq. 4.3, the expression for the internal field can be rewritten as

$$h_i^{int} = \sum_{\mu=1}^{p} \langle m_\mu \rangle \xi_i^\mu - \frac{p}{N} \langle S_i \rangle \tag{4.36}$$

where the mean overlaps are just

$$\langle m_\mu \rangle = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \langle S_i \rangle. \tag{4.37}$$

One of the lessons of the Mattis model is that these overlaps are the natural extensions of the magnetization in the pure ferromagnet. Compare Eq. 4.32. The last term on the right hand side of Eq. 4.36 expresses the fact that $J_{ii} = 0$. For an arbitrary *network state* the overlaps with the patterns will tend to zero as N becomes very large. In fact, they will typically be of order $N^{-1/2}$. In situations of retrieval at least one of the overlaps will be of order unity. If the first term in Eq. 4.36 does not accidentally vanish, then the last term can be neglected, provided p remains fixed when $N \to \infty$. When this first term does vanish at some sites the last term becomes a source of instability, as it does for the symmetric even mixtures.

Self-consistency implies either of the following equivalent conditions

- That h_i^{int} be equal to the value of the internal field which gave rise to the distribution of $\langle S_i \rangle$'s.
- That the values of the average S_i's be equal to the values that entered into the field which produced them.
- Alternatively, since h_i depends only on the m^μ's, that the overlaps be reproduced by the local field that has given rise to them.

These are the conditions under which a distribution will persist.

We prefer to use the last condition which, substituting Eq. 4.34 in Eq. 4.37, is expressed as follows:

$$\langle m^\mu \rangle = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \tanh \left[\beta \left(\sum_{\mu=1}^{p} \xi_i^\mu \langle m^\mu \rangle + h^{ext} \right) \right]. \tag{4.38}$$

This is a set of p non-linear equations for the p variables m^μ. They are called, in physics, *mean-field* equations. In a common physical system they serve as useful approximations. Here, because the network is fully connected and p is held fixed when $N \to \infty$, they become exact. In particular, if the network has memorized a single pattern with all $\xi_i = 1$, then Eq. 4.38 reduces to the mean-field equation of the fully connected Ising model, Eq. 3.36.

4.4.2 Self-averaging and the final equations

A rather important point which has a simple manifestation for finite p arises here. On the right hand side of Eq. 4.38 we still have a sum over the N neurons. We are after the *thermodynamic limit*, namely the behavior for large values of N. At every site i there are p ± 1's, each of which is the i-th bit of one of the p patterns. There are altogether 2^p different possible combinations of the p bits at any site. Now, if p is finite so is 2^p, and when $N \to \infty$ each of these p-bit realizations will itself appear a very large number of times. The frequency with which each one will appear is, by the *law of large numbers*, simply 2^{-p} Therefore, if $G(\xi_i^\mu)$ is a function of the p bits that the p patterns have at site i, then as N tends to infinity

$$\frac{1}{N} \sum_{i=1}^{N} G(\xi_i^\mu) \to \sum_{\xi^\mu} \Pr(\xi^\mu) G(\xi^\mu) = 2^{-p} \sum_{\xi^\mu} G(\xi^\mu) \equiv \langle\langle G(\xi) \rangle\rangle, \tag{4.39}$$

which also serves to define the double brackets. The sums over ξ^μ are over the 2^p possible configurations of p bits that the patterns can produce at any site. In other words,

- The average over the values of a function of the quenched random variables, at all the network points, is equal to the average over the distribution of the values of the function at any particular point, as if there were many systems in which the function takes the different possible values at that point with a probability equal to the probability of the random variables.

This is an expression of self-averaging. The restriction imposed for its validity, namely that p remain finite as $N \to \infty$, is stronger than the actual necessary condition. For a more detailed technical discussion of this property one should consult ref. [24].

For our present purposes, two main conclusions should be drawn. The first is of a particular nature and reduces Eq. 4.38 to

$$\langle m^\mu \rangle = \left\langle\!\!\left\langle \xi^\mu \tanh\left[\beta\left(\sum_{\mu=1}^{p} \xi^\mu \langle m^\mu \rangle + h^{ext}\right)\right]\right\rangle\!\!\right\rangle . \qquad (4.40)$$

in which the double brackets indicate an average over the *quenched* variables ξ, as defined in Eq. 4.39. In the following, we shall omit the single brackets over the overlaps and denote by m^μ a thermal, or time averaged quantity.

The second conclusion is that

- Self-averaging is tantamount to the statement that the properties of a system with a given realization of random variables over its sites are the same as those that would have been obtained by averaging over different systems, each with a different realization of the ξ_i^μ.

This conclusion has a rather naive exterior. It says nothing much more than the familiar fact that if one averages the outcomes of throws of a dice over a very large number of trials the result should converge toward the *mean* outcome – which is the sum of all possible outcomes, each weighted by its probability. It will hold provided, of course, that the distribution of the computed values is very sharply peaked about some typical values. But, as is common in theoretical physics, very simple results are extended cavalierly and the risks taken often lead to important new ideas. In the present context it has become common practice, that whenever a system is described by **fixed** random parameters, i.e., *disordered system* with *quenched randomness*, then the computations are all done as if there were many systems each with different realizations of the random variables. This is of great technical importance because for any particular realization of the random variables the situation is extremely complex and non-uniform.

However, one must decide which are the quantities for which the replacement of a specific realization of the random variables by an ensemble average is correct, namely which are the *self-averaging* quantities. Here we have seen that if one can define local quantities, such as $\langle S_i \rangle$ or h_i, self-averaging is assured. Local quantities can be derived from *extensive* thermodynamic quantities, i.e., system observables which are proportional to the size of the system (or to N) as the size becomes very

large. Such quantities are, for example, the energy, the free-energy (see e.g., Appendix 4.7.1), the magnetization, the overlaps etc. The partition function, for example, which increases exponentially with N, is not. Some probability distributions may not be, see e.g., ref. [25,24].

In conclusion, one can replace the computation of extensive thermodynamic quantities, for a specific realization of the quenched random variables, by an average over an ensemble of systems which is distributed according to the probability distribution of the random variables. It is a simple consequence of the law of large numbers, if the number of values each of the random variables can take is much smaller than the size of the system. It becomes a deeper statement as the number of random variables increases, and then surprises may arise. Such surprises are best detected by computing the moments of the distribution, beyond the averages. If large fluctuations are found, then one has been dealing with a quantity that is not *self-averaging*. See e.g., ref. [25].

4.4.3 Free-energy, extrema, stability

The set of equations 4.40 describes which states (or distributions of states) are self-consistent in the sense that provided the network were to reach them, it would persist in those states. But nothing in these equations tells us whether the solutions will be stable against small perturbations. These equations are essentially like the equations for the extrema of a landscape. But, unless the landscape itself is given, we cannot know whether the solutions are minima, and therefore stable, maxima or saddle-points, which are both unstable.

The discussion in Chapter 3 pointed out that in order to answer such questions we must turn to the *free-energy*, which is the dynamical landscape function for the noisy situation. It also tends to the energy as the noise vanishes, which is the appropriate landscape function in that case. The *attractor distributions* are parametrized by the p overlaps m^μ, which are a straight-forward generalization of the magnetization of the Ising model. What we need is a generalization of Eq. 3.55. We are ensured that this function exists because the synaptic matrix of Eq. 4.3 is symmetric and with asynchronous Glauber dynamics (see e.g., Section 3.5) the energy is

$$E = -\frac{1}{2N} \sum_{ij} \sum_{i \neq j} \sum_{\mu} \xi_i^\mu \xi_j^\mu S_i S_j , \qquad (4.41)$$

as in Eq. 3.59. It can be used, see Appendix 4.7.1, to derive:

$$f = \tfrac{1}{2}\sum_{\mu=1}^{p}(m^\mu)^2 - \frac{1}{\beta}\left\langle\left\langle \ln\left\{2\cosh\left[\beta\left(\sum_\mu \xi^\mu m^\mu + h\right)\right]\right\}\right\rangle\right\rangle, \quad (4.42)$$

after the superscript has been taken off h^{ext}. Recall that the ξ^μ's appearing in this equation are not N-bit patterns but rather p single bits which the patterns may realize at some arbitrary site with the appropriate probability. Eqs. 4.40 are the variational equations of the free-energy $f(m^\mu, T, h)$, namely they are equivalent to the set of equations:

$$\frac{\partial f(m^\mu, T, h)}{\partial m^\mu} = 0. \quad (4.43)$$

Given the free-energy, the various solutions can be sorted out into the different classes of extrema. This will be done in the following sections. Here we will conclude by pointing out that ergodic behavior will prevail if the equations have but a single solution. When they have a unique solution all m^μ must be zero. Such will be the situation for high noise levels $T > 1$, or $\beta < 1$. If any $m^\mu \neq 0$, then the symmetry between the different m's and the symmetry of the equations under the change of sign of any m^μ immediately imply additional solutions and hence broken ergodicity. When ergodicity is broken, $m^\mu \neq 0$ for some μ's. The dynamics leads the network into restricted subspaces of states in each of which the states have *macroscopic (condensed)* overlaps which are of order unity, rather than the typical overlap of order $N^{-1/2}$. These attractor distributions correspond to phases, or states, with macroscopic magnetization. In particular, an attractor with a single $\mu \neq 0$ corresponds to unambiguous retrieval of that particular memory pattern. These states are called *retrieval states*, and at times *Mattis states*, since they are identical in structure to the magnetized phases of the Mattis model of Section 4.4.1. Usually, we shall find attractors with several non-zero μ's, which will be referred to as *mixture states*, or *spurious states*.

4.4.4 Mean-field and free-energy – synchronous dynamics

In Section 3.5.4 we have seen that synchronous dynamics obeys detailed balance but with a rather particular 'energy' – Eq. 3.69. Using this form

for the energy, one can proceed to evaluate the free-energy, to derive the mean-field equations for its extrema, the various attractors in the landscape, stabilities, etc.. When this free-energy is calculated, (see Appendix 4.7.2), there is a small surprise. It reads

$$f(\mathbf{m},\mathbf{t},T) = \sum_{\mu=1}^{n} m^\mu t^\mu - \frac{1}{\beta}\left\langle\left\langle \ln\left\{2\cosh\left[\beta\left(\sum_\mu \xi^\mu m^\mu\right)\right]\right\}\right\rangle\right\rangle - \frac{1}{\beta}\left\langle\left\langle \ln\left\{2\cosh\left[\beta\left(\sum_\mu \xi^\mu t^\mu\right)\right]\right\}\right\rangle\right\rangle, \quad (4.44)$$

where the external field has been set to zero. Note that in contrast to the single set of overlap order-parameters which was required to describe the landscape of the asynchronous dynamics, Eq. 4.42, here we have two such sets – $\mathbf{m}$ and $\mathbf{t}$. This doubling is intimately related to the potential appearance of *2-cycles*. The two sets of order-parameters can be related to overlaps at two consecutive cycle-times.

The mean-field equations are obtained by varying f with respect to m^μ and t^μ. They read:

$$m^\mu = \langle\langle \xi^\mu \tanh(\beta \mathbf{t}\cdot\xi)\rangle\rangle \quad (4.45)$$

$$t^\mu = \langle\langle \xi^\mu \tanh(\beta \mathbf{m}\cdot\xi)\rangle\rangle. \quad (4.46)$$

The dot products are between pairs of p-dimensional vectors of overlaps or pattern components. While these equations look rather different from Eqs. 4.40 with $h = 0$ they possess the surprising property of having the same stable solutions, namely all stable solutions satisfy

$$\mathbf{t} = \mathbf{m}.$$

The proof is provided in Appendix 4.7.2. The exceptional *2-cycles* are connected to the instabilities which accompany situations in which $\xi\cdot\mathbf{m}$ can vanish. These are discussed in Section 4.5.2.

In what follows, we will be mostly concerned with asynchronous dynamics for reasons which were explained in Chapter 2, so this discussion will be cut short.

4.5 Retrieval States, Spurious States – Noiseless

4.5.1 Perfect retrieval of memorized patterns

Consider the mean-field equation for the overlaps Eq. 4.40, with $h = 0$, in the noiseless limit $\beta \to \infty$. In this limit, for $x \neq 0$

$$\lim \tanh(\beta x) = \mathrm{sign}(x).$$

A retrieval state has a single non-zero m component which we choose to be m^1. Its amplitude will be denoted by m. The equation for $\mu = 1$ among the p equations 4.40 reads:

$$m = \langle\langle \xi^1 \mathrm{sign}(m\xi^1) \rangle\rangle = \mathrm{sign}(m) \langle\langle |\xi^1| \rangle\rangle,$$

where we have made use of the simple identity: $x\mathrm{sign(x)}=|x|$. Since ξ^1 is a single bit ± 1, $|\xi^1| = 1$ and hence its mean is also 1. The above equation can now be written as

$$m = \mathrm{sign}(m), \tag{4.47}$$

whose solutions are $m = +1$ or $m = -1$. For $\mu \neq 1$ the equations read:

$$m^\mu = \langle\langle \xi^\mu \mathrm{sign}(m\xi^1) \rangle\rangle = \mathrm{sign}(m) \langle\langle \xi^\mu \xi^1 \rangle\rangle = 0.$$

Recall that the double brackets denote averaging over the single bits in different memories. The single bits are uncorrelated and hence the last term in the above equation is zero.

The consequence is that a state with a single non-zero overlap is a solution of the full set of equations. Observe that:

- $m = 1$ means that the solution is an attractor state which is perfectly aligned with memorized pattern no. 1;
- $m = -1$ is an attractor state in which every single bit is anti-aligned with the same pattern;
- The index $\mu = 1$ was chosen arbitrarily. There are therefore $2p$ solutions with a single non-zero overlap.

Next we show that these retrieval states are the absolute minima of the energy of the network – the 'ground states'. To see this, note that the energy, Eq. 4.41, with J_{ij} given by Eq. 4.3, can be written as

$$E = -\tfrac{1}{2} N \sum_{\mu=1}^{p} \left(\frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu S_i \right)^2 + \tfrac{1}{2} p = -\tfrac{1}{2} N \sum_{\mu=1}^{p} (m^\mu)^2 + \tfrac{1}{2} p. \tag{4.48}$$

The first term increases linearly with N, while the second one remains fixed when the network is at low loading levels, i.e., for low values of p/N. The second term can therefore be neglected, leaving us with

$$E = -\tfrac{1}{2} N \mathbf{m}^2. \tag{4.49}$$

Each component of the p-dimensional vector $\mathbf{m}$ is essentially a cosine and satisfies $|m^\mu| < 1$. But, in fact, as is shown in Appendix 4.7.3, one has the stronger condition:

$$\mathbf{m}^2 \leq 1$$

and equality holds for vectors with a single non-zero component only. The energy attains the value

$$E = -\tfrac{1}{2} N$$

for the retrieval states and is higher for all other attractors.[4]

From the point of view of thermodynamics these lowest lying states exhaust the contributions to the free-energy in equilibrium. As far as storage and retrieval of patterns is concerned they are ideal, since each one of them is unambiguously and fully correlated with one single stored pattern. Their usefulness is exemplified in Figure 4.4, which demonstrates that even a small finite initial overlap draws the network to clearly identify the corresponding stored pattern. In each of these states

$$S_i = \xi_i^\mu$$

for one value of μ. But in a dynamical system every attractor may be effective and all possible solutions have to be investigated, as we proceed to do.

[4] It has recently been shown by B. Tirrozzi and C. Procesi that the retrieval states are absolute minima of the free-energy for all $T < 1$.

4.5.2 Noiseless, symmetric spurious memories

We now turn to look for solutions of the mean-field equations which have the form

$$\mathbf{m} = m_n(\underbrace{1,1,\ldots,1}_{n},\underbrace{0,0,\ldots,0}_{p-n}). \tag{4.50}$$

As was already mentioned, there is a complete symmetry of the solutions under permutations of the components of $\mathbf{m}$, as well as under the change of sign of any number of them. We can therefore choose the n non-zero components to be the first ones and to be all positive. Their common amplitude is m_n. Eq. 4.50 represents

$$2^n \begin{pmatrix} p \\ n \end{pmatrix} \tag{4.51}$$

solutions, where the first factor stands for the number of possible sign changes and the second one for the number of ways that the n non-zero components can be selected out of the p stored patterns. When a solution of the form of Eq. 4.50 is substituted into the mean-field equations 4.40, they reduce to a single equation for the amplitude m_n. It reads

$$m_n = \frac{1}{n}\langle\langle z_n \tanh \beta m_n z_n \rangle\rangle. \tag{4.52}$$

This equation is a summary of the n identical equations for the non-vanishing components. The remaining $p-n$ equations become identities (0=0), much like in the previous section, due to the lack of correlations between different ξ's. The new variable z_n introduced in this equation is defined by:

$$z_n^i = \sum_{\mu=1}^{n} \xi_i^\mu. \tag{4.53}$$

Note that when the ξ_i^μ are chosen with equal probability, the distribution of the values of the variable z_n^i is given by:

$$\Pr(z_n) = 2^{-n} \begin{pmatrix} n \\ k \end{pmatrix}, \tag{4.54}$$

where k is the number of positive entries in z, which can be expressed as

$$k = \tfrac{1}{2}(z_n + n).$$

This is, incidentally, the distribution of a one-dimensional random walk on a lattice.

Taking the noiseless limit in Eq. 4.52, it reduces to:

$$m_n = \frac{1}{n}\langle\langle |z_n| \rangle\rangle, \tag{4.55}$$

and the corresponding energy of these states is obtained when Eq. 4.50 is substituted in Eq. 4.49. It gives:

$$E_n = -\tfrac{1}{2} N n m_n^2,$$

which has the same value for all the solutions with n non-zero overlaps. The important point to notice is that Eq. 4.55 is an explicit solution of the mean-field equations at T=0, because its right hand side does not depend on m_n. It can be computed explicitly to give[2],

$$m_n = \frac{1}{2^{2k}} \begin{pmatrix} 2k \\ k \end{pmatrix} \tag{4.56}$$

where n=$2k$ for even n and n=$2k+1$ for odd n. An example of a 5-mixture attractor has been presented in Figure 4.3. Note, by the way, that the asymptotic behavior of the amplitudes as the number of admixed patterns becomes very large is[2],

$$m_n \approx \left(\frac{2}{n\pi}\right)^{1/2}. \tag{4.57}$$

This result will become interesting when in Chapter 6 we take up the question of networks near their memory saturation.

When the values of m_n are substituted in the expression for the energy, one finds the following ordering of energies:

$$E_1 < E_3 < E_5 \cdots < E_\infty < \cdots < E_6 < E_4 < E_2. \tag{4.58}$$

Some representative values are $E_1 = -0.5N$, $E_2 = -0.25N$, which are the lower and upper bounds for the energy sequence; $E_3 = -0.375N < E_2$. The energies of the odd mixture states increase with the number of admixed patterns, while the energy of the even mixtures decreases with n. As p becomes large so do the allowed values of n. As this

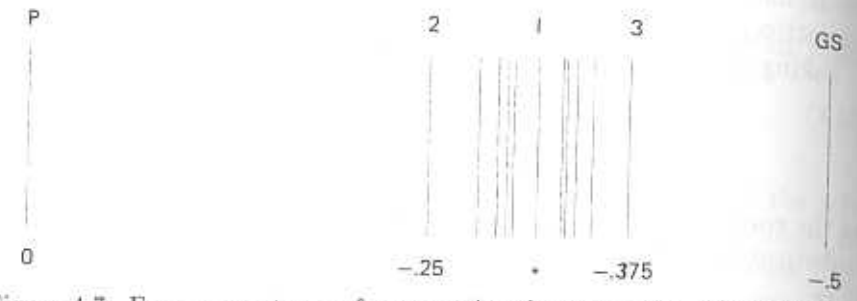


Figure 4.7: Energy spectrum of symmetric mixture states. Above the bars are state labels: P – paramagnetic state; I – $n = \infty$. Below are energy values, $* = -1/\pi$. Each level has a degeneracy depending on n, given by Eq. 4.51.

number becomes large the energies of the odd and even mixtures tend toward a common limit which is

$$E_\infty = -\frac{N}{\pi}.$$

This phenomenology is summarized pictorially in Figure 4.7.

All the mixture states described so far are solutions of the mean-field equations, but that does not make them necessarily attractors. One has to test their stability. The systematic way of doing this is by computing the second variations of the energy at each of the stationary points. Those that are positive definite – increasing in all directions – correspond to minima of the energy and hence are true attractors. Those extrema at which the second variation has decreasing directions are unstable. The full analysis[2] reveals that:

- all the odd mixtures are attractors;
- all the even mixtures are unstable – they are *saddle-points* of the energy.

In the noiseless case, one can arrive at this result by a more elementary route. The local field at any given site is, according to Eq. 4.36

$$h_i = \sum_{\mu=1}^{p} m^\mu \xi_i^\mu - \frac{p}{N} S_i.$$

Let us consider first the odd symmetric mixtures. For those, the field can be written as:

$$h_i = m_n z_n^i - \frac{p}{N} S_i \tag{4.59}$$

with z_n defined by Eq. 4.53. For odd values of n, z_n cannot be smaller than unity, in absolute value. One is therefore justified in neglecting the second term on the right hand side of the expression for the local field. This distribution of fields induces the following distribution of spins:

$$S_i(t + \delta t) = \mathrm{sign}(h_i) = \mathrm{sign}[z_n^i(t)]. \tag{4.60}$$

Each of the spins is in the same direction as its field, Eq. 4.59, and will consequently not change. If some spins deviate from z_n^i, they will not affect the macroscopic overlaps and will very rapidly be aligned with the symmetric mixture. Hence, the odd mixtures are stable. Note, by the way, that Eq. 4.60 implies that, in an odd symmetric mixture, each neuron assumes the state of the *majority* among the activities of the patterns at this site.

For the even mixtures, the situation is quite different. Considering the local field, Eq. 4.59, one must take into account that now there will be a finite fraction of the sites at which the first term in the field vanishes. Those are all the sites at which $z_i^n = 0$. Their number can be read off Eq. 4.54, with $k = n/2$ to be

$$N_{z=0} = \frac{N}{2^n} \binom{n}{n/2}.$$

At all these sites the second term of the field, the subtraction of the self-interactions, cannot be neglected. At all these sites

$$h_i = -\frac{p}{N} S_i,$$

which is always opposite to the direction of the spin at the site, and every such spin will flip, making the state unstable.

Having classified all the symmetric mixture states, one can proceed to count them. If one were to count all the symmetric solutions of the mean-field equations, their number would amount to

$$\#(\text{extrema}) = \sum_{n=0}^{p} 2^n \binom{p}{n} = 3^p. \tag{4.61}$$

This is nothing but the sum over the degeneracies of all the solutions, Eq. 4.51, including $n = 0$ which is also a solution. Out of the 3^p solutions, about one half (the even ones) are unstable. The total number of spurious symmetric attractors is, therefore, approximately $3^p/2$.

4.5.3 Non-symmetric spurious states

Despite the fact that there is a plethora of symmetric spurious states, they have the attractive feature that they can all be classified, counted, and as we shall see in the next section, systematically manipulated by noise. Unfortunately, the story is more complicated. There are spurious states that are not symmetric. This may at first glance appear as a great disappointment, but one should keep in mind that the aesthetic structure of the symmetric mixtures has been strongly contingent on the particular symmetric structure of the J_{ij}'s. This is a feature which is at odds with our stress on robustness. But the issue is not the particular structure of the spurious states. It is the ability to control their numbers and their stability. The number of spurious states is more robust than the form, because it is topologically preserved. In other words, small changes in the synaptic matrix may displace the positions of valleys on the energy landscape, but something more drastic has to happen before valleys disappear.[5]

For the sake of completeness, we present a few asymmetric, stable spurious states. The first one detected has been a 5-mixture of the form:

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

This was the asymmetric state mentioned in Section 4.1. It has an energy of $-0.344N$, which is higher than the energy of the symmetric 3-mixture, $-0.375N$. In Appendix 4.7.4 it is verified that it is a stable solution of the equations.

Another example is

$$\mathbf{m} = (\tfrac{3}{8}, \tfrac{3}{8}, \tfrac{3}{8}, \tfrac{3}{8}, \tfrac{3}{16}, 0, \ldots, 0)$$

with energy $E = -0.334N > E_3$. In both cases, as in all other asymmetric spurious states, one can perform all the symmetry operations

[5]The mathematician may abbreviate such considerations by reference to Morse inequalities connecting the indices of extrema of many dimensional surfaces.

among the components of $\mathbf{m}$ and change their signs, to produce additional solutions with the same energy. These examples should convince the reader that the situation is overly rich.

4.5.4 Are spurious states a free lunch?

Spurious states present a real temptation. Information is stored by specially sculpting synaptic connections to ensure the memorization of a given set of patterns and, lo and behold!, cognitive retrieval of extra patterns becomes possible. Those are patterns that have not been learned intentionally. How much closer can one approach creative thinking?

- Our attitude here is that spurious states are spurious!

There is a whole list of arguments in support of this dogmatic position. Let us start by reflecting on the fact that if a symmetric *5-mixture* is an attractor, the network will recognize a pattern which is created bit by bit, using the majority rule, from five patterns that have been actually memorized. Such a pattern will appear familiar to the network, while any inspection of such a pattern will lead to the intuitive conclusion that it should be unrecognizable. Moreover, the combination of **every** grouping of five patterns becomes retrievable. It is a rather general feature of information retrieval that, at best, one can recall out of a channel the entire information that had been inserted in it. Np bits are used to construct the synapses, in some process of learning, and at most Np bits can be independently read out.

Spurious states may inhabit the dynamics of the network, and as such will be considered a nuisance. Usually we will look for tools which will help in disposing of them. Noise will sometimes play this role and sometimes it will be played by asymmetry of synapses[26,27]. As a general attitude, we prefer the psychiatric metaphor of schizophrenia, discussed in Section 2.3.4, as a categorization of mixture, spurious states.

4.6 Role of Noise at Low Loading

4.6.1 Ergodicity at high noise levels - asynchronous

Turning back to the full mean-field equations 4.40 for the average *condensed* overlaps,

$$m^\mu = \left\langle\!\left\langle \xi^\mu \tanh\left(\beta \sum_{\mu=1}^{p} \xi^\mu m^\mu\right)\right\rangle\!\right\rangle,$$

we recall that the double brackets indicate an average over the discrete distribution of the quenched, randomly chosen ξ^μ. These equations always have a solution with

$$m^\mu = 0$$

for all $\mu = 1, \ldots, p$. Non-ergodicity, *symmetry breaking*, etc. can take place only if additional solutions appear.

Some conclusions concerning the absence of other solutions can be reached by a rather elementary argument. One notes that

$$|\tanh x| \leq |x|.$$

The right hand side of each mean-field equation, number 1 for example, is therefore smaller than

$$\left\langle\!\left\langle \xi^1 \left(\beta \sum_{\mu=1}^{p} \xi^\mu m^\mu\right)\right\rangle\!\right\rangle.$$

Now, the average over the ξ's will leave only one term in the sum, because the different patterns are uncorrelated, namely

$$\langle\langle \xi^1 \xi^\mu \rangle\rangle = 0,$$

unless $\mu = 1$, in which case it equals 1. The right hand side of each equation can now be bounded by

$$\beta m^\mu.$$

Consequently, if $\beta < 1$ – when the noise level is high – the right hand side of every equation is always smaller than the left hand side and no solution, other than the *paramagnetic* one with all $m^\mu = 0$, is possible.

When only a single solution exists, the dynamical behavior of the system is necessarily ergodic. After all, every initial condition can develop after sufficiently long time to the same distribution of states. See e.g., Chapter 3.

4.6.2 Just below the critical noise level

As we have just seen, for $\beta < 1$ or $T > 1$ there are no interesting attractors, no solutions with any $m^\mu \neq 0$. As the noise decreases, β increases, and the hyperbolic tangent will be able to intersect the straight line m^μ. This change is schematically depicted in Figure 3.5, which leads one to the expectation that when retrieval solutions appear, they do so in a continuous fashion, namely the amplitudes of these solutions vanish as the noise level approaches T=1 from below. The continuity of the transition provides a very useful tool for the analysis of the appearance of the retrieval states with noise decreasing from 1 down. The flavor of this analysis will be given below. The main features are that below, but near, the critical noise level:

- All stationary points – the solutions of the mean-field equations – are symmetric mixtures.
- The square amplitudes of the overlaps in these solutions vanish linearly as the noise level $T \to 1$, from below.
- The free-energies of all these solutions increase monotonically and linearly in the number n of admixed patterns.
- The only stable solutions are the 'retrieval' states, which have a single non-zero overlap. All other extrema are saddle-points, except for the ergodic solution which is a maximum.

We have put retrieval here between quotes because as one approaches $T = 1$ the amplitude of the single overlap tends to zero and it becomes less and less likely that such states can actually be used for retrieval.

In order to visualize the shape of the landscape, one can return to Figure 2.11. The free-energy near T=1 for p=2 when viewed in the m_1-m_2 plane has very much that shape. One can see the four retrieval state minima, placed symmetrically on the axes. At the origin is the paramagnetic state, which is a maximum. At high values of the two overlaps the free-energy increases, essentially quadratically. As one tries to tailor the rest of the surface, one soon finds that the decreasing contour lines, coming from the origin, and the increasing one coming from the four minima, must match at saddle points, somewhere between the minima. The saddle points are just the four symmetric 2-mixtures. What is significant is that their presence is imposed by topological

considerations of fitting continuous surfaces, which rise at infinity, and have four minima.

Digression: Analysis of retrieval near $T = 1$

Let us denote $1-T$ by t, which will be assumed to be small. Then, assuming that just below T=1 any m^μ which is not zero will be very small, one can expand the tanh on the right hand side of the mean-field equations, Eqs. 4.40, according to

$$\tanh(x) \approx x - \tfrac{1}{3}x^3.$$

The equations now read

$$m^\nu = \beta \sum_{\mu=1}^{p} \langle\langle \xi^\nu \xi^\mu \rangle\rangle - \tfrac{1}{3}\beta^3 \sum_{\mu,\rho,\sigma=1}^{p} m^\mu m^\rho m^\sigma \langle\langle \xi^\nu \xi^\mu \xi^\rho \xi^\sigma \rangle\rangle. \qquad (4.62)$$

This equation is much simplified when one evaluates the averages over the the ξ's, which for different ξ's are uncorrelated bits. Furthermore, since the second term on the right hand side of Eq. 4.62 is already of third order in the small quantities m^μ, one can set $\beta = 1$ in that term. In the first term, linear in m^μ, we must keep the correction t. We arrive at

$$tm^\nu = \tfrac{1}{3}m^\nu \left(3\mathbf{m}^2 - 2(m^\nu)^2\right).$$

Since we are after solutions with non-zero components, m^ν can be cancelled to give:

$$\mathbf{m}^2 - 2(m^\nu)^2 = 3t. \qquad (4.63)$$

This is a rather interesting result. It implies that:

- all non-ergodic solutions below but near the critical noise level are symmetric mixtures.

This is because the equation can be read as saying that

$$(m^\nu)^2 = \tfrac{3}{2}\left(\mathbf{m}^2 - t\right),$$

which implies that all components of any possible solution are equal in magnitude, since they all are equal to the sum of the squares of all the components participating in the solution.

- All retrieval solutions as well as all symmetric mixtures appear together, as T goes below 1.

Other spurious states may appear but only at a finite temperature below the critical T=1. To see this, let us take again the form of the symmetric mixture, Eq. 4.5.2, as a candidate solution. When substituted in 4.63 it gives:

$$m_n^2 = \frac{3t}{3n-2}. \qquad (4.64)$$

This shows that all the 3^p symmetric solutions start growing in amplitude as T decreases below 1, and all vanish continuously as $T \to 1$. In particular, the retrieval states, those with only a single non-zero overlap, have the amplitude:

$$m_1 = 3t.$$

The free-energy, Eq. 4.42, can also be evaluated explicitly near the critical temperature[2] and it reads:

$$f_n \approx -\frac{n}{2}t(\beta m_n)^2 + \frac{n(3n-2)}{12}(\beta m_n)^4 + \ln 2 \qquad (4.65)$$

where, unlike the energies listed in the previous section, this is the free-energy per spin. When the solutions for m_n are inserted in Eq. 4.65, one finds:

$$f_n = -\frac{3nt^2}{4(3n-2)} + \ln 2.$$

Here the situation is different from that at T=0. The spurious states are ordered monotonically in free-energy – higher n implies higher free-energy, irrespective of the parity of the mixture. It is probably this simple ordering of the solutions in energy that underlies the absence of additional spurious states. It also leads to the interesting feature that:

- the only stable states near the critical noise are the retrieval states.

Questions of stability

To investigate the stability of the extrema of a landscape function, f, one must turn to the second variations around these points. One must investigate the properties of the matrix:

$$A^{\mu\nu} \equiv \frac{\partial^2 f}{\partial m^\mu \partial m^\nu} \qquad (4.66)$$

at the solutions of the mean-field equations. When the second derivative is applied to the free-energy, Eq. 4.42, it leads to the following form for this stability matrix:

$$A^{\mu\nu} = (1-\beta)\delta^{\mu\nu} + \beta Q^{\mu\nu} \tag{4.67}$$

with

$$Q^{\mu\nu} = \langle\langle \xi^\mu \xi^\nu \tanh^2(\beta \mathbf{m}\cdot\xi)\rangle\rangle. \tag{4.68}$$

A solution will be locally stable only if all the eigen-values of the matrix $\mathcal{A}$ are positive.

Consider a symmetric mixture solution of the form Eq. 4.5.2. At such a solution the stability matrix $\mathcal{A}$ will have the following simple form:

- All diagonal elements are

$$A^{\mu\mu} = 1 - \beta(1-q),$$

with

$$q = \langle\langle \tanh^2(\beta m_n z_n)\rangle\rangle. \tag{4.69}$$

The physicist will recognize the Edwards-Anderson *order-parameter*, which measures the degree of freezing of the spins. It will be discussed in greater detail in Section 6.3.

- The off-diagonal elements with $\mu,\nu \le n$ are all equal to

$$A^{12} = \beta\langle\langle \xi^1\xi^2 \tanh^2(\beta m_n z_n)\rangle\rangle \equiv \beta Q. \tag{4.70}$$

- All other elements vanish.

This form applies to the symmetric solutions at all temperatures[2], and the form of the eigen-values can also be derived without recourse to expansions. There are, in all, three different eigen-values.

1. A single (*non-degenerate*) eigen-value, which expresses the stability against a uniform increase in the amplitude m_n of all the overlaps. It is

$$\lambda_1 = 1 - \beta[(n-1)Q - (1-q)]. \tag{4.71}$$

2. An eigen-value of degeneracy $n-p$

$$\lambda_2 = 1 - \beta(1-q) \tag{4.72}$$

which measures the stability against the appearance of additional overlaps.

3. An eigen-value of degeneracy $n-1$

$$\lambda_3 = 1 - \beta(1-q) - \beta Q \tag{4.73}$$

which is associated with fluctuations which tend to make the solution asymmetric, namely to change the amplitude of the overlaps relative to one another.

Specializing again to the neighborhood of the transition, the three eigen-values can be computed explicitly in an expansion in $t = (1-T)$. First one finds from Eq. 4.69, and then from Eq. 4.64 that

$$q \approx \frac{3nt}{3n-2} = nm_n^2. \tag{4.74}$$

Expanding 4.70 one finds

$$Q \approx \frac{2q}{n}. \tag{4.75}$$

Before proceeding with the analysis of the eigen-values we stop to observe the curious fact that on the one hand q, as defined by Eq. 4.69, is

$$q = \langle\langle \tanh^2(\beta m_n z_n)\rangle\rangle = \langle\langle\ \langle S_i\rangle^2\ \rangle\rangle.$$

On the other hand, m_n is, essentially

$$m_n = \langle\langle\ \langle S_i\rangle\ \rangle\rangle.$$

Note the two types of averages in the last two equations. One, the internal single brackets is a thermal, or temporal, average. The other, the double external brackets stand for an average over the random variables ξ. In an attractor, one might have expected

$$q = m_n^2,$$

as in a ferromagnet, for example. Now, for $n = 1$, this is true here as well. Compare Eqs. 4.64 with 4.74. This is a ferromagnetic state in the Mattis sense. For $n > 1$ this is no longer true. In fact, one finds that $q > m_n^2$, indicating that there is random freezing of spin orientations over and above the freezing that is associated with the correlation with one of the memories. These extra frozen spins are randomly oriented with respect to all the stored patterns. More on this topic will appear in Section 6.3.

Coming back to the stability analysis, we substitute q and Q in the expressions for the three eigen-values to find:

$$\lambda_1 \approx -t + q + (n-1)Q \approx 2t > 0.$$

Hence, in this direction all the solutions are stable. Also $\lambda_2 > 0$ because,

$$\lambda_2 \approx -t + q \approx \frac{2t}{3n-2}.$$

Finally,

$$\lambda_3 \approx -t + q - Q \approx \frac{-4t}{3n-2}$$

which is always negative! But, by its definition, this eigen-value does not exist if $n = 1$. The conclusion is that all stationary points with $n > 1$ are unstable in the direction that tends to induce anisotropy among the component overlaps. Actually, they all tend to flow to one of the $n = 1$ valleys, which are the only ones that are stable.

4.6.3 Positive role of noise and retrieval with no fixed points

As the temperature is lowered from T=1, we must arrive at a point where the stability properties of the various mixture states begin to change. After all, upon reaching $T = 0$ we must recover the full complexity of the situation described in the previous section. The even-n mixtures remain unstable in the entire range $0 < T < 1$. The odd-n ones must start crossing the even ones in energy and start to change their stability from saddle-points to local minima. It is in carrying out this complex reshuffling process that topology resorts to the asymmetric mixtures in order to make do. It turns out[2] that each odd asymmetric mixture becomes locally stable at its own temperature.

In fact, for stability of one of the extrema to change it is λ_3, Eq. 4.73, that must change sign.

The temperature at which the n-mixture will become stable is determined by the condition $\lambda_3 = 0$, with the eigen-value evaluated at that solution. This condition can be written, using 4.73, as:

$$T_n = Q - q = \langle\langle(\xi^1\xi^2 - 1)\tanh^2(\beta m_n z_n)\rangle\rangle \qquad (4.76)$$

which has to be solved numerically for each value of n separately. It turns out that the first change of stability, when the 3-mixture becomes an attractor, is at T_3=0.461. This is a very interesting result, mentioned in Section 4.1.4, above. It implies that noise has a positive role. It eliminates the spurious states, much like has been anticipated in the discussion in Section 2.3.4. In the noise window $0.461 < T < 1$ only pure retrieval states are attractors. But to actually improve retrieval it must also be the case that there be a window in which the overlaps are very close to unity so that there are not too many errors. That this is in fact the case is exhibited in Figure 4.2, where one can observe retrieval at T=0.6, after the 3-mixture has been destabilized. Similar qualitative conclusions about the role of noise in a more realistic neural network have also been reached in ref. [28].

Two final general comments are in place in closing this chapter. The first comment concerns the meaning of retrieval at finite noise level. Clearly, in the presence of noise there are no fixed points. The state of the network will never cease to fluctuate. Retrieval must therefore imply a temporal average of the network states over some time interval which is determined by the biological output, or readout, mechanism. Retrieval would be called if this average is high enough. The important point to realize here is that in our interpretation of a cognitive event there is no significant difference between the noisy and noiseless retrieval. If retrieval depends, in the noiseless case, on the arrival of the network at a fixed point, then the network must be able to ascertain that it is indeed at such a fixed point, which it can do only by means of a temporal average, just as in the noisy case. See e.g., Sections 2.3.2 and 1.4.4.

One can further argue that even in the noiseless situation there is, in the more realistic situation[28], an effective noise due to the stochastic way in which neurons emit spikes, also above threshold. Hence, the arrival of the network at a noiseless fixed point must be interpreted as representing a situation in which a certain group of neurons fire trains

of spikes while others do not. But those trains are unsynchronized and only a temporal average over a time of 30–40 milliseconds can meaningfully detect the special activity pattern.

The second comment concerns the status of the aesthetic detail described in this chapter, following ref. [2]. It should be clear that much of the detail is strongly dependent on the precise form of the synapses as given in Eq. 4.3. The interest in such a study, beside its manifest beauty which may be a mere cultural convention of physicists, is that it brings out a number of general features in a very precise form. These include:

- The necessary appearance of spurious states, as a topological consequence of the existence of a landscape function in many dimensions.
- The considerations of stability of extrema and the possible change in stability as a function of noise.
- The role of noise as a candidate for improved retrieval where spurious states are a problem.

4.7 Appendix: Technical Details for Low Storage

4.7.1 Free-energy at finite p – asynchronous

We proceed to the computation of the free-energy of the N neuron network in which the couplings are constructed according to Eq. 4.3 from a particular realization of p random patterns. In other words, we are after

$$f(\beta, m^\mu, J_{ij}, h^\mu) = -\frac{1}{\beta N} \ln Z,$$

with

$$Z = Tr_S \exp\beta \left(\frac{1}{2} \sum_{i,j} J_{ij} S_i S_j + \sum_\mu h^\mu \sum_i \xi_i^\mu S_i \right),$$

where h^μ are uniform external fields. This is a simple generalization of the expressions in Section 3.4.2. The presence of the retrieval order-parameters m^μ on the left hand side expresses the anticipation that ergodicity will break in a richer way than in the ferromagnetic case of

Section 3.4.2, and that these m^μ's will be the proper parameters, *order-parameters*, which will classify the different probability distributions. What we are after is the generalization of Eq. 3.55.

Substituting the expression for J_{ij} in terms of the patterns, Eq. 4.3, and treating carefully the diagonal term, one can write:

$$Z = e^{-\beta p/2} Tr_S \exp\left[\frac{\beta}{2N} \sum_\mu \left(\sum_i \xi_i^\mu S_i \right)^2 + \beta \sum_\mu h^\mu \sum_i \xi_i^\mu S_i \right]. \quad (4.77)$$

In order to perform the Tr, one has to linearize the quadratic term in the spins in the exponent. This is a p-fold version of Eq. 3.51, namely

$$\begin{aligned} Z &= (N\beta)^{p/2} e^{-\beta p/2} \int_{-\infty}^{\infty} \prod_{\mu=1}^{p} \frac{dm^\mu}{\sqrt{2\pi}} \\ &\quad \times \mathrm{Tr}_S \exp\left(-\frac{N\beta}{2} \sum_\mu (m^\mu)^2 + \sum_i \sum_\mu (m^\mu + h^\mu) \xi_i^\mu S_i \right). \end{aligned} \quad (4.78)$$

Now the Tr can be performed, as in Eq. 3.53, to give

$$f = \tfrac{1}{2}\mathbf{m}^2 - \frac{1}{N\beta} \sum_i \ln\left(2\cosh[\beta(\mathbf{m}+\mathbf{h})\cdot\xi_i]\right). \quad (4.79)$$

On the right hand side the three vectors have p components each. Note also that as long as p is finite these are the only terms in f which are of order one. All other terms in Eq. 4.78 are at most of order $\ln N/N$.

To conclude the computation, we observe that the second term on the right hand side of Eq. 4.79 is simplified by the considerations of self-averaging, as explained in Section 4.4.2. For p finite the sum over i tends to the average over the distribution of random patterns. Thus we arrive at Eq. 4.42.

4.7.2 Free-energy and solutions – synchronous dynamics

The free-energy

Given the 'energy' $\tilde{E}$ of Eq. 3.69, the free-energy is calculated from the partition function

$$Z = Tr_S \exp(-\beta \tilde{E}).$$

The complication lies in the fact that the variables S_i are inside the $\ln\cosh$ in $\tilde{E}$. This is disentangled by the introduction of δ-functions for each μ. One writes

$$Z = Tr_S \exp\left[\sum_{i=1}^{N} \ln\left[2\cosh\left(\beta\sum_{\mu}\xi_i^\mu\left(\frac{1}{N}\sum_j \xi_j^\mu S_j\right)\right)\right]\right]$$

$$= Tr_S \int\prod_\mu dm^\mu \exp\left[\sum_{i=1}^{N}\ln\left[2\cosh\left(\beta\sum_\mu \xi_i^\mu m^\mu\right)\right]\right]\delta\left(m^\mu - \frac{1}{N}\sum_j \xi_j^\mu S_j\right)$$

The p δ-functions can be Fourier transformed according to

$$2\pi\delta(x) = \int_{-\infty}^{\infty} dt\exp(itx),$$

using p variables t^μ. After this is done, all spin variables appear linearly in the exponent and the trace can be carried out, as in Eq. 3.53. The result is

$$Z = C\int\prod dm^\mu dt^\mu$$
$$\exp\left(-N\mathbf{m}\cdot\mathbf{t} + \textstyle\sum_{i=1}^{N}\ln[2\cosh(\beta\xi_i\cdot\mathbf{m})] + \sum_{i=1}^{N}\ln[2\cosh(\beta\xi_i\cdot\mathbf{t})]\right).$$

From here, with the help of *self-averaging* and the *saddle-point* method, one arrives directly at Eq. 4.44.

Merging of the solutions

Let us rewrite Eqs. 4.46 in vector form, i.e.,

$$\begin{aligned}\mathbf{m} &= \langle\langle\xi\tanh(\beta\mathbf{t}\cdot\xi)\rangle\rangle\\ \mathbf{t} &= \langle\langle\xi\tanh(\beta\mathbf{m}\cdot\xi)\rangle\rangle.\end{aligned}$$

Subtracting the two sets of equations, one has

$$\mathbf{m}-\mathbf{t} = \langle\langle\xi\left(\tanh(\beta\mathbf{t}\cdot\xi) - \tanh(\beta\mathbf{m}\cdot\xi)\right)\rangle\rangle.$$

Multiplying both sides of this equation in a scalar product by $\mathbf{m}-\mathbf{t}$, one can write

$$(\mathbf{m}-\mathbf{t})^2 = \langle\langle(\xi\cdot\mathbf{m} - \xi\cdot\mathbf{t})[\tanh(\beta\mathbf{t}\cdot\xi) - \tanh(\beta\mathbf{m}\cdot\xi)]\rangle\rangle.$$

This equation has a left hand side which is non-negative and a right hand side which is non-positive. The equality implies then that $\mathbf{m} = \mathbf{t}$.

4.7.3 Bound on magnitude of overlaps

The vector form of the mean-field equations is

$$\mathbf{m} = \langle\langle\xi\,\mathrm{sign}(\mathbf{m}\cdot\xi)\rangle\rangle,$$

where each vector has n components, corresponding to the non-zero entries of $\mathbf{m}$ in the particular solution. Taking the scalar product of $\mathbf{m}$ with both sides of the equation and recalling that $x\,\mathrm{sign}(x) = |x|$ one has

$$\mathbf{m}^2 = \langle\langle|\mathbf{m}\cdot\xi|\rangle\rangle. \tag{4.80}$$

To estimate the right hand side, we make use of the Schwartz inequality which reads:

$$\left(\sum_k x_k y_k\right)^2 \le \sum_k (x_k)^2\sum_k (y_k)^2. \tag{4.81}$$

To employ it, we observe that Eq. 4.80 is a mean over the distribution of the ξ's, as is indicated by the double brackets. The right hand side of Eq. 4.80 is, therefore, a sum of the probabilities, p_k, for the appearance of a particular vector ξ, times the corresponding value of $|\mathbf{m}\cdot\xi|$. One can then choose

$$x_k = \sqrt{p_k}|\mathbf{m}\cdot\xi|_k \qquad y_k = \sqrt{p_k},$$

Inserting these definitions into the inequality Eq. 4.81, one has:

$$\begin{aligned}\langle\langle|\mathbf{m}\cdot\xi|\rangle\rangle &= (\sum_k x_k y_k)\\ \le \left[\sum_k (x_k)^2\sum_k (y_k)^2\right]^{1/2} &= \langle\langle(\mathbf{m}\cdot\xi)^2\rangle\rangle^{1/2}\times 1,\end{aligned}$$

where the 1 represents $\sum(y_k)^2 = \sum p_k = 1$.

Returning to Eq. 4.80, we can now write,

$$\begin{aligned}\mathbf{m}^2 &\le \langle\langle(\mathbf{m}\cdot\xi)^2\rangle\rangle^{1/2}\\ &= \left(\sum_{\mu,\nu=1}^{n} m^\mu m^\nu\langle\langle\xi^\mu\xi^\nu\rangle\rangle\right)^{1/2}\\ &= (\mathbf{m}^2)^{1/2}.\end{aligned}$$

This implies directly that

$$\mathbf{m}^2 \leq 1,$$

unless $n = 1$, which is what we have set out to prove.

4.7.4 Asymmetric spurious solution

In order to establish that a solution is stable, all one has to show is that the local field at every site cannot vanish. With the above overlaps, the local field at neuron i is, according to Eq. 4.36:

$$h_i = \tfrac{1}{4}(2z_2^i + z_3^i),$$

with z_n^i given by 4.53. The term in parentheses cannot vanish because it is a sum of an even and an odd integer. Thus, if this mixture is a solution it is a stable one.

Next, note that out of the p mean-field equations only five are not identities of the type 0=0. These five divide into equivalent equations numbers 1 and 2, and three equivalent equations for 3–5. The two groups can be condensed into two equations by averaging inside each group. One then has:

$$(m_1 + m_2) = 1 = \langle\langle z_2 \mathrm{sign}(2z_2 + z_3)\rangle\rangle$$

and

$$(m_3 + m_4 + m_5) = \tfrac{3}{4} = \langle\langle z_3 \mathrm{sign}(2z_2 + z_3)\rangle\rangle.$$

The first equation is easily verified: $z_2 = 0$ with probability $\frac{1}{2}$ and does not contribute to the right hand side. When $z_2 \neq 0$ it is either $+2$ or -2, each with probability $\frac{1}{4}$. In both cases $2z_2$ dominates the sign for all values of z_3 and the right hand side is

$$\tfrac{1}{4}[2(+1) + (-2)(-1)] = 1.$$

The second equation is slightly more complicated. If $z_2 = \pm 2$, then z_2 dominates the sign, and since it is uncorrelated with z_3 this contribution averages to zero and we have to contend with $z_2 = 0$ only. This has a probability of $\frac{1}{2}$. The second equation can, therefore, be written in the form

$$\tfrac{3}{4} = \tfrac{1}{2}\langle\langle z_3 \mathrm{sign}(z_3)\rangle\rangle = \tfrac{1}{2}\langle\langle |z_3|\rangle\rangle.$$

Now $\langle\langle |z_3|\rangle\rangle = \frac{3}{2}$, which leads to the verification of the second equation.

Bibliography

[1] J.J. Hopfield, Neural networks and physical systems with emergent computational abilities, *Proc. Natl. Acad. Sci. USA*, **79**, 2554(1982).

[2] D.J. Amit, H. Gutfreund and H. Sompolinsky, Spin-glass models of neural networks, *Phys. Rev.*, **A32**, 1007(1985).

[3] T. Kohonen and M. Rouhonen, Representation of associated data by matrix operators, *IEEE Trans. Comput.*, **22**, 701(1973).

[4] L. Personnaz, I. Guyon and G. Dreyfus, Information storage and retrieval in spin-glass like neural networks, *J. Physique Lett.*, **46**, 359(1985).

[5] C. von der Malsburg and E. Bienenstock, Statistical coding and short-term synaptic plasticity: a scheme for knowledge representation in the brain, in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman Soulié and G. Weisbuch eds. (Springer-Verlag, Berlin, 1986).

[6] D.O. Hebb, *The Organization of Behavior* (Wiley, NY, 1949).

[7] E. Goles, Positive automata networks, in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman Soulié and G. Weisbuch eds. (Springer-Verlag, Berlin, 1986) and E. Goles and G.Y. Vichniak, Lyapunov functions for parallel networks, in J.S. Denker ed. (AIP Conf. 151, NY, 1986).

[8] D.J. Amit, The properties of models of simple neural networks, in L. van Hemmen and I. Morgenstern eds. *Heidelberg colloquium on Glassy Dynamics* (Springer-Verlag, Heidelberg, 1987).

[9] M. Abeles, *Local Cortical Circuits* (Springer-Verlag, Berlin, 1982).

[10] S. Amari, Learning patterns and pattern sequences by self-organizing nets of threshold elements, *IEEE Trans. Comput.*, **21**, 1197(1972).

[11] W.A. Little and G.L. Shaw, Analytic study of the memory storage capacity of a neural network, *Math. Biosci.* **39**, 281(1978).

[12] J.P. Rauschecker and W. Singer, The effects of early visual experience on cat's visual cortex and their possible explanation by Hebb synapses, *J. Physiol. (London)*, **310**, 215(1981).

[13] L.N. Cooper, F. Lieberman and F. Oja, A theory for the acquisition and loss of neuron specificity in visual cortex, *Biol. Cybern.*, **33**, 9(1979).

[14] T. Kohonen, *Self Organization and Associative Memory* (Springer-Verlag, New York, 1984).
[15] J.C. Eccles, *The Physiology of Synapses* (Springer-Verlag, Berlin, 1964).
[16] I. Kanter and H. Sompolinsky, Associative recall of memory without errors, *Phys. Rev.*, **A35**, 380(1986).
[17] R.A. Anderson and V.B. Mountcastle, The influence of the angle of gaze upon the excitability of the light-sensitive neurons of the posterior parietal cortex, *J. Neurosci.*, **3**, 532(1983).
[18] M. Sur, J.T. Wall and J.H. Kaas, Modular distribution of neurons with slowly and rapidly adapting responses in area 3b of somatosensory cortex in monkeys, *J. Neurophysiol.*, **51**, 724(1984).
[19] M.E. Goldberg and C.J. Bruce, Cerebral cortical activity associated with the orientation of visual attention in the rhesus monkey, *Vision Res.*, **25**, 471(1985).
[20] S. Kirkpatrick and D. Sherrington, Infinite-ranged models of spin-glasses, *Phys. Rev.*, **B17**, 4384(1978).
[21] S.F. Edwards and F. Tanaka, The ground state of a spin glass, *J. Phys.*, **F10**, 2471(1980).
[22] D.C. Mattis, Solvable spin systems with random interactions, *Phys. Lett.*, **56A**, 421(1976).
[23] J.M. van Hemmen, Classical spin-glass model, *Phys. Rev. Lett.*, **49**, 409(1982); see also J.P. Provost and G. Vallee, Ergodicity of the coupling constants and the symmetric n-replicas trick for a class of mean-field spin-glass models, *Phys. Rev. Lett.*, **50**, 598(1983).
[24] M. Mezard, G. Parisi and M. Virasoro, *The Replica Method and Beyond* (World Scientific, Singapore, 1987).
[25] M. Mezard, G. Parisi, N. Sourlas, G. Toulouse and M. Virasoro, Replica symmetry breaking and the nature of the spin glass phase, *J. Physique*, **45**, 843(1984).
[26] J. Hertz, G. Grinstein and S. Solla, Neural nets with asymmetric bonds, in *Heidelberg Colloquium on Glassy Dynamics*, J.J. van Hemmen and I. Morgenstern eds., (Springer-Verlag, Berlin, 1987)
[27] G. Parisi, Asymmetric neural networks and the process of learning, *J. Phys.*, **A19**, L675(1986).
[28] J. Buhmann and K. Schulten, Influence of noise on the function of a physiological neural network, *Biol. Cybern.*, **56**, 313(1987).

5

Storage and Retrieval of Temporal Sequences

5.1 Motivations: Introspective, Biological, Philosophical

5.1.1 The introspective motivation

The type of neural network described in the previous chapter is a first prototype in the sense that:

- it stores a small number of patterns;
- it recalls single patterns only;
- once a pattern has been recalled, the system will linger on it until the coming of some unspecified dramatic event.

Such a system may provide some useful technical applications as rapid, robust and reliable pattern recognizers. Such devices are discussed in Chapter 10. It seems rather unlikely that they can satisfy one's expectations of a cognitive system.

Very rudimentary introspection gives rise to the impression that, with or without explicit instruction, a single stimulus (or a very short string of stimuli) usually gives rise to a retrieval (or recall) of a whole cascade of connected 'patterns'. Most striking are effects such as the recall of a tune, which can be provoked by a very simple stimulus, not

directly related to the tune itself. Similarly, rather simple stimuli bring about the recall of sequences of numbers, especially in children, or of the alphabet. Similarly, much of the input into the cognitive system seems to be in the form of temporal sequences, rather than single patterns. This appears to be accepted in the study of speech recognition (see e.g., ref. [1]), as well as in vision, where a strong paradigm has it that form is deciphered from motion (see e.g., ref. [2]).

To the extent that one would like to have these types of processing affected by a single network, one has to impose two requirements:

- The possibility of storing a substantial number of patterns in a single network. An issue to be taken up in the next chapter.
- The introduction of non-symmetric synaptic arrangements.

A methodological disclaimer

> We shall often discuss various questions, such as recall of tunes etc., in the parlance of cognitive psychology. This will be done in conjunction with technical exercises on dynamics of neural networks, modified one way or another. In doing this we do not presume to propose solutions for psychological problems. On the contrary, the intention is to seek inspiration from simple descriptions of complex human performances for testing the limits of the range of potential network behavior, allowing certain modifications which preserve the general nature of the interpretation of network activity. If a rich enough spectrum can be implemented on such, relatively simple, networks it will be up to the psychologist to make ontological commitments.

5.1.2 The biological motivation

Many biological systems exhibit *central pattern generators* (CPG). These are neural groups which control the muscles involved in a wide variety of rhythmic functions, such as breathing, chewing, swimming, scratching etc. Such neural systems are attractive candidates for modeling in terms of neural networks because:

1. They are anatomically well localized and identified.

2. Their rhythmic output is independent of feedback from the controlled muscles or from higher parts of the nervous system. See e.g., ref. [3].
3. CPG's function in the absence of *pacemaker cells* – neurons whose intrinsic, individual firing rate could determine the rhythm of the network. One may, therefore, expect the rhythmic behavior to be a collective property of the whole system. See e.g., the study of the swimming of the mollusk *Tritonia diomedea*[4], or of the leech[5,6].
4. The same CPG network can produce several rhythms in the same animal, to control a variety of behaviors. See e.g., ref. [7].
5. The output of a CPG *can* be modulated by afferent inputs. It can be turned on and off, for example.

The rhythmic behavior of CPG's can be envisaged as a *cyclic* transition between a variety of stored patterns. If a sufficient number of patterns are embedded in the network, then one may be able to form more than one such cycle, to account for the variety of rhythmic behavior generated by a single CPG network.

Here the disclaimer of the previous subsection can be lifted. Contact with experiment is direct. An entire network can be isolated in a preparation and the activity of its neurons can be measured individually as well as many of their mutual influences – synaptic efficacies[8] and network rhythmic output activity. See e.g., Figure 5.1, where a cyclic, *temporal sequence* of attractor states, appearing as bursts, is manifestly identifiable. This has allowed a detailed analysis in terms of model networks[9]. But, of course, not all the news is good. The price for direct accessibility to analysis in such networks is extreme simplicity, both in structure and in function. One cannot avoid some of the reservations mentioned in Section 1.2.3.

Yet, the success in modeling CPG's by ANN's, while not of profound import on the applicability of such models to cognitive processes in *homo sapiens*, may be pivotal in the debate over the allowed simplification of neural building blocks. In other words, even in situations in which the complexity of the elements entering the network is directly known, some of the main characteristics of the network as a whole can be accounted for in terms of neurons and synapses that are as simple as physicists would like them to be.

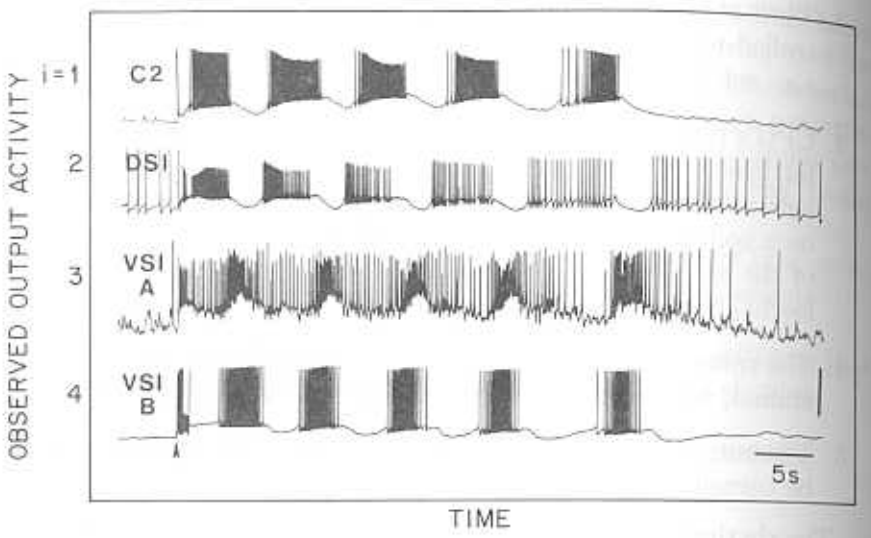


Figure 5.1: Simultaneously recorded output activity of four neurons in an isolated brain preparation from Tritonia. These neurons comprise the CPG controlling the escape swim sequence. The labels on the left are personal names given to the various neurons. The arrow indicates the initiation of the sequence. (From ref. [9] after ref. [8].)

5.1.3 Philosophical motivations

Wittgenstein, in the *Tractatus*[10], has a few stimulating observations on thought and sequences, or about how *facts* are composed of *objects*. They have been brought forward in Changeux' *Neuronal Man*[11]. In his own words,

> The facts in logical space are the world. (1.13)
> An atomic fact is a combination of objects. (2.01)
> The object is the fixed, the existent; the configuration is the changing the variable. (2.0271)
> In the atomic fact objects hang one in another, like the links in a chain. (2.03)
> In the atomic fact the objects are combined in a definite way. (2.031)

We here take the liberty of reading the early Wittgenstein as saying that *objects*, read stimuli patterns, are inexorably connected in strings

("chains"), which have an intrinsic ordering, here interpreted as a *temporal sequence*.

Next we turn to a reflection of Hebb's on a philosophical objection to brain as mind. The philosophical position is summarized by Hebb as follows:

> If mind is a brain process...we could not hear the clock strike twelve; the brain gets the same message twelve times, so, if that is all there is, what one would hear is the clock striking one over and over again...[12]

which, of course, would be the case if a neural network had single pattern attractors. Hebb goes on to propose a neurophysiological rebuttal:

> It is an example of a certain tendency to base far-reaching philosophic conclusions on misinformation about the nervous system. When the brain is exposed to a series of stimulations, the stimuli may be identical but their effects in the brain are not. The state of the brain is modified by the first stimulus, which means that the impact of the second is different from that of the first...[12]

This is clearly a problem that requires resolution on the neurophysiological level. There may be more than one possible explanation for this type of *short term memory*, namely the memory that a certain stimulus is not appearing for the first time in some sequence of stimuli. A typical Hebbian attitude would be that each appearance of the stimulus modifies synapses and hence the next appearance finds a system with modified dynamics. But the same question can be directed to an ANN. Then, in the spirit of the methodological disclaimer pronounced in the preceding subsection, one can inquire about the necessary minimal modifications which have to be introduced in an ANN, of the type described in Chapter 4, to enable it to handle such situations within its own terms of reference. One clearly has to go beyond simple attractor networks.

Finally, we will mention Fodor's objection to neural networks as cognitive *information processors*[13]. The argument is that the information that appears at the output end of a network is but the outcome of a certain computation. The structure of the computation is expressed only in the connectivity arrangement of the network. There is, therefore, no way of telling what computation the neural output

activity is the outcome of. The absence of *labels*, which indicate the structure of the computation, on the output information transmission, has, according to Fodor, a number of fundamental drawbacks for a candidate model of mental activity:

- The outcome can be taken to be that of a different computation. If some bit indicates the true or false values of an *or*, it would look the same if it were thought to be the outcome of an *and*. [p. 84]
- A network of the same structure as that computing some logical inference would also compute 'any inference generated from [the first] by substituting logical equivalents in the premise or in the conclusion'. But from the psychological point of view 'it is entirely conceivable, for example, that someone for whom the former inference is fast and obvious might find the latter inference [after substitution] obscure and slow'. [pp. 93-94]
- If 'Baby has learned to draw [the binary *or*] inference [by growing] a neural instantiation' of it, 'he has to grow a whole new network' in order to be able to draw the *or* of three propositions. [p. 94]
- 'There is no sense to the question "what is the form of the argument this network computes?" for there are no *labels* for the nodes of the machine, and anyway *labels are for us not for the machine*'.

One extreme way of summarizing this list of difficulties is to say that 'psychological processes are sensitive to *syntactic* variables (like form) but not to *semantic* variables'[13]. One does not have to accept this fundamentalist position in toto in order to be convinced that form and syntax are important as, moreover, *context* is. One possible way of transmitting the *labels* with the outcome of the computation is by manipulating the *temporal sequences*, each of which communicates the full structure of the problem. Thus at the output end (intermediate or final) one may have a temporal sequence which contains all the necessary labeling. Another possibility is that after the manipulation, which is the temporal sequence itself, the network arrives at some attractor state which is rich enough to contain the labeling and to communicate it all at the output. A third possibility is to have the incoming temporal sequence 'transduced' into a spatial pattern which is rich enough to give rise to a temporal sequence of network states.

These are all parts of an extensive research program in the present context. Some glimpses of the potential implementations will be described in Section 5.4.2, below. It may be that this approach, in which the semantic part is reduced to a selection of cognitively relevant inputs, may allay the recent criticisms of neural networks. This does not imply that they would then become valid descriptions of mind, but rather that before provoking Fodor's next attack we should have networks which can handle temporal sequences.

5.2 Storing and Retrieving Temporal Sequences

5.2.1 Functional asymmetry

Our discussion of the dynamics, synchronous or asynchronous, has led to the conclusion that if the synaptic efficacies are symmetric, then the attractors can be either single states, or if the dynamics is synchronous at most cycles of length two. See Section 3.5. It has, therefore, become clear, rather early on, that asymmetry will have to be introduced in a major and coherent way in order to provoke interesting departures from single state attractors.

Asymmetry plays a multiple role on the scene of neural networks. One is of a nuisance. Since symmetric ANN's are readily accessible to analysis, the biologists' insistence on the presence of asymmetry in real networks is a necessary evil. Fortunately, as is discussed in Section 7.3.2, it turns out that the introduction of random asymmetry into a symmetric, well functioning ANN adds some noise, but does not change the qualitative performance very drastically, even at rather high levels of asymmetry. This is good news. But, the other side of this coin is that it requires a more massive and *coherent* introduction of asymmetry, when its effect is desirable, as would be the case for *temporal sequences*.

5.2.2 Early ideas for instant temporal sequences

Ideas concerning temporal sequences, especially cyclical ones, appeared at rather early phases of the research into neural networks. (See e.g., refs. [14], [15], [16]). A more systematic study of the naive introduction

of sequences was carried out by Hopfield[17,18]. His conclusions were that

> By inserting the anti-symmetric part of J_{ij} the *direction* of where to go next is clearly indicated... It proves to be much more difficult to persuade the collection of neurons to move from one state to another... Careful adjustment of parameters ultimately permitted following a sequence of four states, but performance was not very good... More seriously, the delicate balance of parameters necessary to achieve this result suggested that an essential ingredient was lacking. [18] [p. 384].

In the present section we will clarify the source of the difficulties mentioned above. Then we shall move on to describe how these proposals have been recently extended to provide a rich and robust framework for the storage and retrieval of temporal sequences.

To realize the promise as well as the limitations of the early ideas for producing cycles in ANN's, we will resort to the type of simplified analysis that was developed in Section 4.3.1. In other words, for the sake of quick insight, random fluctuations will be neglected, which is perfectly justified as long as the ratio of the number of stored patterns to the number of neurons, p/N, is small enough.

The idea revolves around the addition of a *transition term* to the set of synaptic efficacies. The original part - the *stabilizing term* - is our familiar

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu. \tag{5.1}$$

This term, as shown in Section 4.4.1, produces a local field (PSP) on neuron i

$$h_i = \sum_{\mu=1}^{p} \xi_i^\mu m^\mu,$$

where m^μ are the p overlaps of the current network state $\{S_i\}$, Eq. 4.2. In the absence of *fluctuations*, if for some value of $\mu = \nu$, m^ν=1, the overlap with all other patterns will vanish. The absence of fluctuations implies that the random overlaps between randomly constructed patterns can be neglected.

In that case, the role of the stabilizing term is to ensure that if the network is in a state that is identical to a stored pattern it will remain

there. But perhaps more importantly, if the network is in a state in which a fraction of neurons are out of alignment with a given stored pattern, it will be quickly attracted to the nearby memory. The reason, to recapitulate some of the arguments of Section 4.3.1, is that if the network is in a state in which

$$m^1 = 1 - O(\frac{1}{\sqrt{N}}),$$

for example, then for $\mu \neq 1$

$$m^\mu = O(\frac{1}{\sqrt{N}}),$$

where $O(x)$ stands for a quantity that vanishes at least as fast as x when $x \to 0$. The field at site i will be

$$h_i = \xi_i^1 + \sum_{2}^{p} \pm O(\frac{1}{\sqrt{N}}).$$

The $\pm$ in the sum indicates that the terms are as likely to be positive as negative. As a consequence, the *signal* term, ξ_i^1, dominates the PSP at each neuron and, following one cycle, S_i will align itself with ξ_i^1, provided only that the noise, or temperature, is not too high. This will be true for either parallel or sequential updating.

Suppose next that a transition term of the form

$$J_{ij}^t = \frac{\lambda}{N}\sum_{\mu=1}^{q} \xi_i^{\mu+1}\xi_j^\mu \tag{5.2}$$

is added to the stabilizing J of Eq. 5.1. The coefficient λ being the relative strength of the two terms. At this stage, one can imagine that every two neurons are connected by two synapses, one symmetric and one non-symmetric with relative efficacy λ. This composition of two types of synapses can be implemented with one synapse between each two neurons, some symmetric and some non-symmetric. This should become clear after reading Chapter 7, where we show that the network's behavior is robust under extensive dilution of the synaptic structure. In the presence of the transition term the local fields at each neuron will have two contributions, namely

$$h_i = \sum_{\mu=1}^{p} \xi_i^\mu m^\mu + \lambda \sum_{\mu=1}^{q} \xi_i^{\mu+1} m^\mu. \tag{5.3}$$

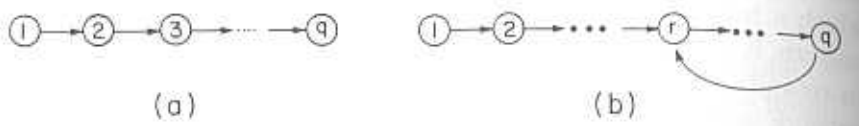


Figure 5.2: Temporal sequences. (a) A linear sequence terminating in a single state attractor. (b) A chain terminating in a cycle.

Starting in a network state $S_i = \xi_i^1$, one has

$$h_i = \xi_i^1 + \lambda \xi_i^2. \tag{5.4}$$

Consider the case with $\lambda < 1$, and a low noise level. Then, ξ^1 will determine the sign of h_i and the dynamics will attract the network state to the stored pattern number one. The asymmetry will not affect the stability. On the other hand, if $\lambda > 1$, the situation is quite different and strongly dependent on the type of dynamics. The sign of h_i will be determined by ξ^2. For synchronous, parallel, dynamics all neurons will align with pattern number two, following a **single time-cycle**, in which all S_i's are updated with the same m^μ's. Then, in the next time-cycle, the network will hop into pattern number three, if that pattern happened to follow pattern number two in the 'chain' defined by the transition term. These abrupt jumps will continue until the network has reached the last pattern, number q, in the chain. If pattern q is some isolated pattern, then on arriving there the network will settle. On the other hand, pattern q may be one of the previous patterns in the chain, e.g., number r. In that case, the network will enter a periodic cycle attractor, consisting of states r to q. This possibility is sketched in Figure 5.2, in which member states of a temporal sequence are denoted by circles and the arrows indicate their organization in time.

The conclusion is that if the dynamics is synchronous and the network is organized with synapses which are the sum of Eqs. 5.1 and 5.2, then the network will produce a temporal sequence of states. It will spend in each of these states a single neural cycle-time. This fact explains why Hopfield's difficulties had not been encountered previously. As far as the present discussion is concerned these temporal sequences are unsatisfactory on several counts.

- Even in synchronous dynamics the transitions will be extremely sensitive to noise. If random correlations between patterns misalign some neurons there will be nothing to correct the errors. This implies very low storage or, alternatively, very high values of λ.

- Full synchrony is not a plausible approximation to the random continuous updating taking place in a realistic network – biological or artificial.

- Even if synchrony were the rule, since the network spends a single neural cycle-time in each of the states of the sequence, none of these states can be cognitively recognized in the sense of Section 1.4.4. Each state in the sequence is no different from any transient network state. At best, the entire sequence, or perhaps the repeated cycle of states, can be considered as a single cognitive event. But that would satisfy none of the motivations raised in Section 5.1.

As soon as the updating departs from full synchrony, the simple picture derived from Eq. 5.3 is severely modified. The reason is that as one moves from one neuron to the next, in the updating sequence, the overlaps composing the field h_i, in Eq. 5.3 vary. They vary because the updating of a certain neuron is performed taking into account the new states of the neurons formerly updated. Hence, if the network starts in state $S_i = \xi_i^1$, or m^1=1, then the first neurons to be updated will feel a field as in Eq. 5.4. For $\lambda > 1$ they will flip into alignment with pattern number 2, as before. But as the number of flips increases, m^1 decreases and m^2 becomes larger. Suppose that these changes have brought about a situation in which $m^1 = \frac{1}{2} - z$ and $m^2 = \frac{1}{2} + z$, which must be an intermediate state if the system is to make a transition from state number 1 to state number 2. Then, the field of the next neuron to be updated will be, according to Eq. 5.3,

$$h_i = (\tfrac{1}{2} - z)\xi_i^1 + [\tfrac{1}{2} + \lambda + z(1-\lambda)]\xi_i^2 + \lambda(\tfrac{1}{2} + z)\xi_i^3. \tag{5.5}$$

This is a new kind of game. It can be exercised on one's PC, but the resulting incoherence can be appreciated from the following simple considerations. The next neuron to be updated is still in neural state $S_i = \xi_i^1$. Consider then the two possibilities:

1. If $\xi_i^2 = \xi_i^1 = -\xi_i^3$, then

$$h_i = \left(1 + \frac{\lambda}{2} - 2z\lambda\right)\xi_i^1$$

and, for $z > (2+\lambda)/4\lambda$, $\mathrm{sign}(h_i) = -\xi_i^1$.

2. If $\xi_i^2 = -\xi_i^1 = -\xi_i^3$, then

$$h_i = 2z(\lambda - 1)\xi_i^1$$

and, since $\lambda > 1$, $\mathrm{sign}(h_i) = \xi_i^1$.

These are two examples in which the process of updating will start deviating from the transition from 1 to 2. In the first case, which can occur only if $\lambda > 2$, because $z < \frac{1}{2}$, the i-th neuron will take on its value in state 3, and **not** that in state 2 or 1. In the second case the i-th neuron will choose to remain in its state in the initial pattern 1 and not flip into its state in pattern 2.

In Figure 5.3 we show simulation results of a network of 300 neurons storing 11 patterns. The three curves present overlaps of the network state with the pattern in the sequence to which it is expected to have moved as a function of time. In (a) the iterations are synchronous and λ=1.7, which should have sufficed to drive the sequence. Random overlaps destroy the operation. When in (b) λ is raised to 2.5, the sequence works properly with synchronous dynamics, but with this very high λ it fails when dynamics becomes asynchronous in (c).

5.3 Temporal Sequences by Delayed Synapses

5.3.1 A simple generalization and its motivation

The difficulties associated with the early types of proposals for storage and recall of temporal sequences have led to several alternatives. We will restrict ourselves here to the attempts which do not involve dynamic modification of synapses (see e.g., ref. [19]), a topic we shall evade as long as we possibly can. One way of dealing with the cognitive difficulty mentioned in Section 5.2.2 is to assume that the asymmetric (transition) synapses are activated from outside the network at time intervals which allow a sufficiently long presence in each attractor to allow for a cognitive interpretation[20,21]. One might imagine that

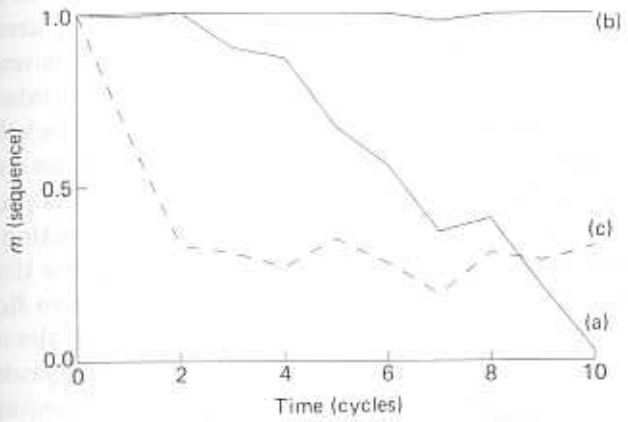


Figure 5.3: Simulation of instant sequence dynamics (N=300, p=11) – the overlaps of the network state with expected sequence state at consecutive time steps. (a) synchronous (λ=1.7) – destroyed by fluctuations. (b) synchronous dynamics (λ=2.5) – good operation. (c) asynchronous (λ=2.5) – disaster.

such control could be exercised by some brain component via *heterosynaptic regulation* of the efficacy of *allosteric receptors* by the activity of neighboring synapses[22]. For such a proposal[20] to work, one would still require parallel (synchronous) updating, which is not robust. The *synaptic triad* mentioned above has been employed in modeling another network which can learn and retrieve temporal sequences — bird songs[21].

What one would like to have is a network that has quasi-attractors with the following three features:

- They attract in a robust way – insensitive to the updating procedure and to noise.
- They are stable for a period of time which is significantly longer than the neural cycle-time, so that the presence in the attractor can be recognized as a cognitive event.
- Prolonged presence in one of the attractors destabilizes it and the network moves on.

The reason the sequence retrieval mechanisms of the previous section did not work was that the transition term J^t, Eq. 5.2, became operative

as soon as the system started entering a new attractor. This left no time for the dynamics to either allow the network to arrange itself comfortably inside the attractor – retrieve it in a robust way, nor did it allow for a long enough stay inside the attractor to attain *meaning*.

A simple remedy comes immediately to mind. If the effect of the network's presence in an attractor on its transfer dynamics could be delayed, then both requirements could be met. This, it was proposed[24, 25], could be done in a rather simple fashion. If the transition synapses, of Eq. 5.2, were to act with a time delay which would make the influence of a pre-synaptic neuron transmitted through them arrive significantly later than the influence coming through a synapse of the stabilizing type, Eq. 5.1, then the network should deal satisfactorily with temporal sequences, as we shall see following a few comments on synaptic delays.

Comments on slow synapses

Very significant synaptic delays are quite familiar in invertebrates. See e.g., ref. [9]. Certain kinds of delays are known in vertebrate synapses, such as the internal dynamics of the allosteric receptor under the influence of the activity in its neighboring synapses[26,21]. These delays will not do for our present purposes. What we need here is a 'synaptic mechanism' which will slow down, by a number of neuronal cycle-times, the arrival of the chemical signal from neuron j to neuron i. In the triad synapse, a receptor can be *desensitized* for some time and will be resensitized only by the intermediary of the arrival of an additional signal at the synapse. The delays to be employed here should be a permanent feature of the synapse, in the absence of learning, which should slow down the communication of every signal trying to come through it, independently of recent signal history. Some indication that such synapses are present in the brain may be inferred from the significant delay with which visual stimuli reach the visual cortex. This delay can be as long as 80–100ms, although there are no more than three synapses on the route from the eye to the cortex.[1]

If such delays do not exist in a simple variety, they will have to be invented by the nervous system in a more roundabout, but perhaps a more versatile, way. It is almost inconceivable that the wide range of temporally related phenomena confronting a living system could be

[1] I owe this remark to Professor Valentino Braitenberg.

successfully manipulated by the central nervous system without such delays. That is the reason for the introduction of the term *synaptic mechanism*. After all, synapses are but contacts between neurons. Such contacts could conceivably be expanded to include neurons or even entire networks in their midst. If they are so constructed then the arrangement will have to include an anatomy of divergent-convergent connections, similar to that invoked by Abeles[23] for the synfire mechanism.

Finally, it should be recalled that the neat arrangement of pairs of one fast and one slow synapse on a neuron is convenient for analysis but is not essential for the effective functioning of the network. This is one more aspect of the robustness of distributed dynamics. We return to this point when the wider issue of robustness is discussed in Chapter 7. Moreover, the simple, sharp time delay in the synapses may not be quite rich and representative enough. It may very well be that whatever *synaptic mechanism* is invoked, its resulting outcome will be an average of the PSP's transmitted over a preceding time interval, with some characteristic weight function[24]. For a wide variety of weight functions, the network remains an effective memory of temporal sequences, much like the one with a sharp time delay described below. The salient technical features required for the introduction of such modifications are described in Appendix 5.6.1.

5.3.2 Dynamics with fast and slow synapses

Irrespective of the origin of the delays, we can now go back and analyze the dynamical behavior of the system. As usual, the system is described by its set of connections. Those are given by Eqs. 5.1 and 5.2. The fact that the transition synapses act with a delay is expressed in the computation of the PSP's. Instead of Eq. 5.3 we would have

$$h_i(t+\delta t) = \sum_{\mu=1}^{p} \xi_i^\mu m^\mu(t) + \lambda \sum_{\mu=1}^{q} \xi_i^{\mu+1} m^\mu(t-\tau) \qquad (5.6)$$

where, as usual, m^μ is the overlap of the state of the network $\{S_i\}$ with the memorized pattern $\{\xi_i^\mu\}$:

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

The first term in Eq. 5.6, the stabilizer, depends on the activities of the neurons one cycle-time earlier, via the $m^\mu(t)$, while the transition term depends on the activities of the neurons τ cycle-times earlier, via $m^\mu(t-\tau)$.

In order to make the dynamics more transparent, we assume again that fluctuations do not play a role. This is justified if p/N is very small. We will comment on the effect of fluctuations later on. Just like in the discussion in Section 5.2.2, the absence of fluctuations implies that when the network is in a given memory attractor its overlap with that attractor is unity, and all other overlaps with stored patterns are negligible. Suppose that at time t=0 the system is in a state perfectly aligned with pattern number 1, while previously it had been in a random state, with negligible overlaps with **all** patterns in the sequence. This is expressed by the overlaps:

$$m^1(t=0)=1, \qquad m^\nu(t=0)=0 \quad \text{for } \nu>1$$
$$m^\mu(t<0)=0 \quad \text{for all } \mu.$$

In the first time interval, when $0<t<\tau$, the PSP on neuron i, according to Eq. 5.6, will be

$$h_i=\xi_i^1 \tag{5.7}$$

and the network will remain in network state 1 until $t=\tau$. At time τ the information that the network has settled in pattern 1 will have reached the neurons via the slow transition synapses. In the next one or two cycle times the situation will be somewhat complicated. At the beginning of the first cycle time the PSP's are

$$h_i=\xi_i^1+\lambda\xi_i^2. \tag{5.8}$$

If $\lambda<\lambda_c$ ($\lambda_c=1$ in the absence of fluctuations), the network will remain indefinitely in the state corresponding to pattern 1. But, for $\lambda>\lambda_c$ the second term will dominate and neurons will start aligning themselves with their appropriate states in pattern 2. At those sites where patterns 1 and 2 agree, no change takes place, but neurons which would have different activity states in the two patterns will flip from their state in pattern 1 to their state in pattern 2.

This will go on, as long as the number of neurons that have gone over is small compared to N. Since the dynamics is asynchronous, once the number of flipped neurons has increased, we have to modify

our assumptions about the various overlaps, which in turn determine the PSP's. Suppose that a fraction y of the neurons has flipped from pattern 1 to 2. Then the current overlaps with pattern 1 and 2 will be

$$m^1=1-2y \qquad m^2=2y.$$

The PSP of the next neuron in line will be

$$h_i=(1-2y)\xi_i^1+(2y+\lambda)\xi_i^2. \tag{5.9}$$

Compare Eq. 5.8. Note that this expression for the PSP differs also from Eq. 5.5 by the absence of a term pulling toward pattern 3. This term is absent here because the news that the network has entered a state with a finite overlap with pattern 2 will take τ time steps to arrive through the transition synapses. The result is that the asynchronous drifting into 2 becomes a robust avalanche, without any confusing side-effects from the next elements in the chain. Note that for a duration τ a quasi-attractor is more stable than the corresponding attractor in the absence of a sequence. Following a short transient time, the network settles in the quasi-attractor, y=0.5 and the PSP is

$$h_i=(1+\lambda)\xi_i^2,$$

which is a significantly greater signal term than in the usual attractors discussed in Chapter 4.

This situation persists for τ time steps and then the next transition begins. The sequence of transitions continues as inscribed in the slow synapses. Each pattern is visited for an interval τ. If the present interpretation of the role of sequences is to be plausible, then τ should be of the order of magnitude of the *cognition time*. Whether this is or is not the case is first and foremost a question about the available *logical* delaying synapses.

5.3.3 Simulation examples of sequence recall

In Figure 5.4 we present a simulation of the dynamics of a network of 200 neurons in which 10 patterns are stored as a sequence, which at state number 6, runs into a cycle. The synaptic time delay is τ=5, which shows up as the visiting time of each of the quasi-attractors. The updating is asynchronous. We plot the overlaps with the memorized patterns. In part (a), the initial state, the stimulus has a large overlap

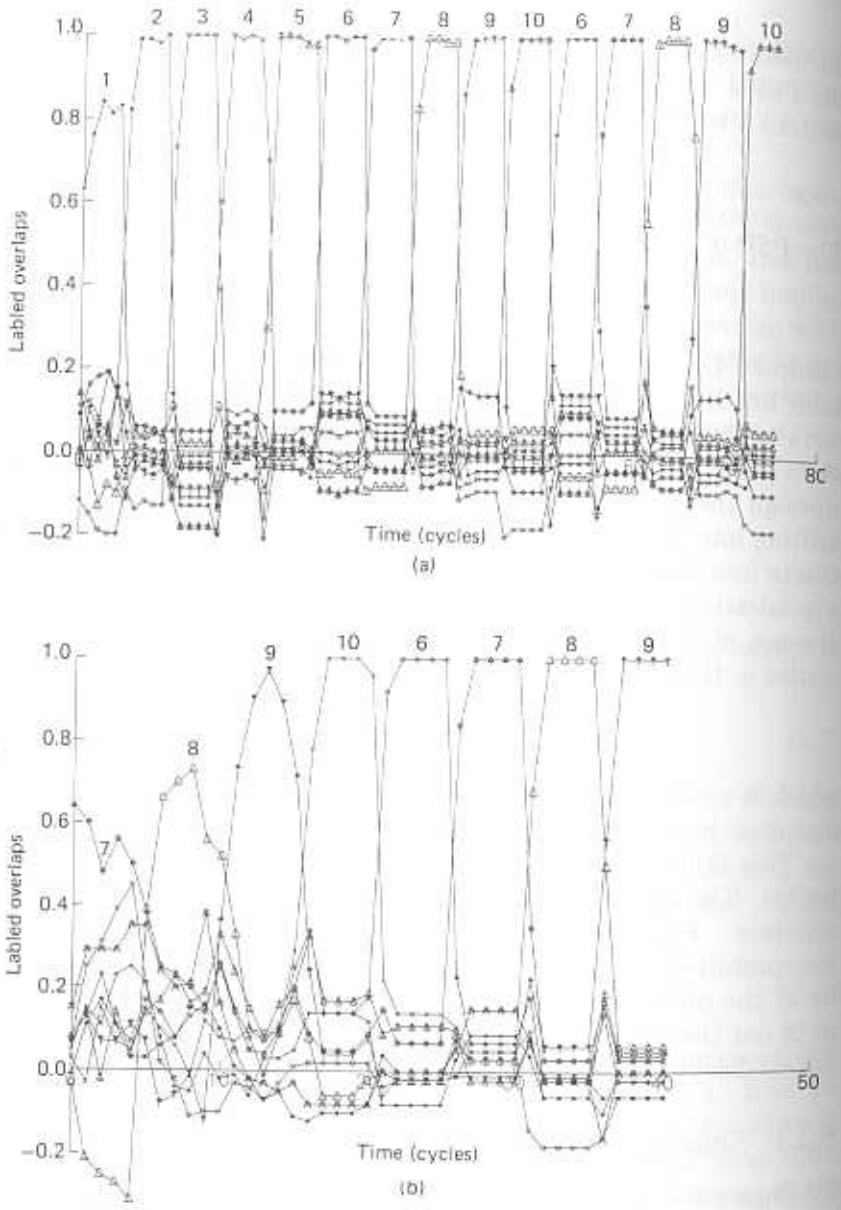


Figure 5.4: Simulation of the asynchronous dynamics of a network with N=200, p=10, λ=1.5, $\tau = 5$, T=0.6. The sequence is 1–2–3–4–5–6–7–8–9–10–6. Curves are marked by the number of the pattern to which the particular overlap corresponds. Stimuli: (a) $m^1(0)$=0.6, (b) m^7=0.6. Following a short transient the network follows the sequence, identifying clearly each pattern for the duration of the synaptic delay.

with pattern number 1. In (b), the stimulus has its largest overlap with pattern 7.

The same phenomenon can be exhibited in a more visually pleasing way. The 200 neurons can be conceived as a 20×10, two-dimensional square of pixels, on which a retina image of a digit or a letter is mapped. The retrieval of a pattern or of a sequence is then described as the reconstitution of a faithful two-dimensional image, or sequence of images, from a noisy stimulus. See e.g., refs. [27,28]. As far as artificial devices are concerned, this direct mapping is under the designer's control. When it comes to the recall of images in the nervous system, there is a rhetorical element in such presentations, despite the fact that ultimately they may correspond to reality. At this point there is essentially no knowledge as to the kind of pre-processing that may be affecting sensory stimuli before they are memorized or used as cues for retrieval. Yet, such displays are irresistible.

Figure 5.5 presents a recall of a sequence of three out of ten visual digits drawn on a square of 300 pixels. The stimulus is a noisy digit 3 and in the transient stages each consecutive digit is noisy, until the network settles into the attractor. If the digits are stored as direct geometrical mappings of pixels, there would be large overlaps between the patterns, and the synaptic matrix Eq. 5.6 will not do. The large overlaps have two origins. The first is the large ratio of background to foreground. If white pixels are mapped onto inactive neurons, then in each pattern there would be many more inactive neurons than active ones. The second origin is the correlated use of foreground pixels by different digits. Compare e.g., the digits 3 and 5. One would tend to think that this fact may be a strong incentive against direct storage.

In Section 4.2.3 it was shown that a most complete way of dealing with general correlations between patterns is the construction of a synaptic connection matrix which is an orthogonal projection matrix. This technique was generalized to the recall of sequences[9] (see e.g., Appendix 5.6.2) and was used to generate Figure 5.5.

There are two other approaches to the internal representation of visual recall. The first retains the direct visual mapping of pixels on neurons, but tries to preserve the local relation between the synaptic efficacies and the activities of the neurons in the memorized patterns. This can be achieved by an elimination of the mean correlations between the stored patterns from the synaptic efficacies. See e.g., Section 8.2. The second approach would be to consider that the

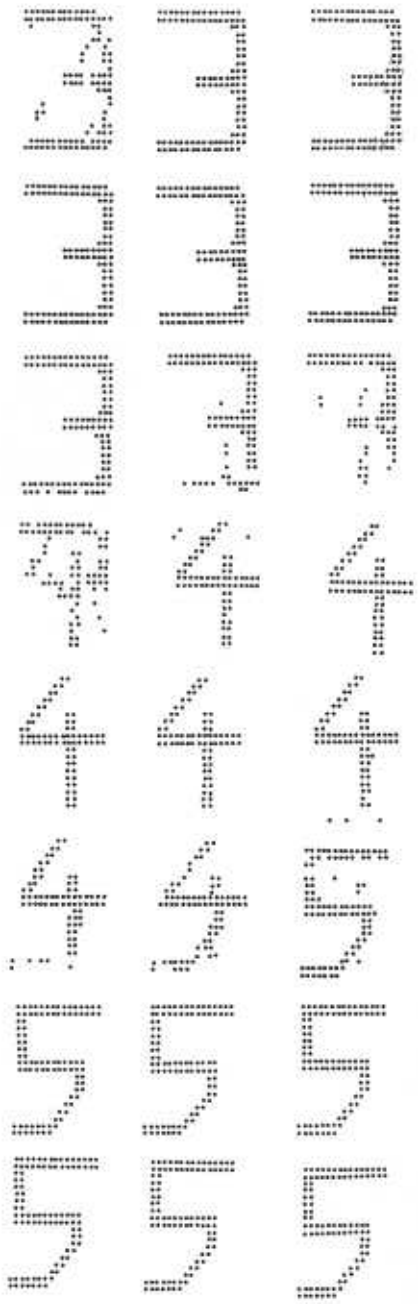


Figure 5.5: Recall of a sequence of visual images of digits. Synapses are constructed according to the orthogonal projection matrix.

nervous system pre-processes the images into uncorrelated neural activity patterns which are based on the characteristics of each image. Such preprocessing would allow a simple synaptic matrix for storage and retrieval. These patterns could still be translated into visual images for the purpose of literary presentation, with the fraction of unaligned neurons in transient states sprinkled around the clean images as random noise. To the best of our knowledge, currently available empirical data does not exclude any of the three alternatives.

5.3.4 Adiabatically varying energy landscapes

The existence of an energy function or of a Lyapunov function plays a central role in the analysis of the properties of neural networks with symmetric synaptic connections. This has been amply discussed in Chapters 3 and 4 and will be taken up again in Chapter 6. The synaptic matrix composed of the sum of the J_{ij}'s of Eqs. 5.1 and 5.2 is clearly not symmetric, because J^t_{ij} is not, while the fast synapses are.[2] Fortunately, due to the slow action of the slow synapses, some vestige of this useful construct can be salvaged[9].

As we have seen in the previous section, the system does not have stable states, but at best quasi-attractors. Therefore, a real energy function cannot possibly exist. Yet, for periods of duration τ, the system behaves as if an attractor exists, implying the existence of a corresponding 'temporary' energy function. The first term in that 'energy' would be the former expression, Eq. 3.59, which describes the relaxation of the network into the minima determined by the fast synapses. This relaxation takes place on a time scale of the neuronal cycle-time of about one milli-second. It has the form

$$E^1 = -\tfrac{1}{2} \sum_{i,j\ i \neq j} J_{ij} S_i S_j. \tag{5.10}$$

On the other hand, the slow synapses act like an external field (PSP) $h^t_i(t)$ which is determined by the state of the network at a time

[2] On interchanging i and j in J^t_{ij} of Eq. 5.2 one replaces the i-th element of pattern $\mu + 1$ with the corresponding element of pattern μ, and those two will usually not be the same.

τ earlier and acts on the neurons at the present time. This field, the second term in Eq. 5.6, is given by

$$h_i^t(t) = \lambda \sum_{\mu=1}^{q} \xi_i^{\mu+1} m^\mu (t - \tau).$$

Since it varies slowly within a period τ, it can be considered fixed and then its contribution to the energy will be a term just like Eq. 3.14, namely

$$E^2 = -\sum_{i=1}^{N} h_i^t S_i. \tag{5.11}$$

The sum of E^1 and E^2 acts like an energy, with its landscapes, for a time of order τ. After the network settles into a quasi-attractor corresponding to one of the patterns stored in the sequence, ξ^μ, the field h_i^t acts to deepen that well, relative to its depth in the landscape of E^1. See e.g., Figure 5.6(a). Following a time τ, the field starts to evolve due to the effect of the slow synapses. The effect being that the valley at ξ^μ begins to be filled up and the one at $\xi^{\mu+1}$ begins to deepen – Figure 5.6(b). This is the transition period. If $\lambda > \lambda_c$, eventually the valley at ξ^μ disappears, the network spills into $\xi^{\mu+1}$, Figure 5.6(c), and the show repeats.

The scenario is depicted graphically in Figure 5.6, in a restricted part of the network state space. Only two dimensions are displayed, one (horizontal in the plane of the page) along the transition line between states μ and $\mu + 1$ and one (drawn into the page) in another direction. The cross-points on the surfaces are network states and adjacent points differ by a single neuronal (spin) state. The distance between the two valleys represents the number of neurons that must change their state for the network to go between the two states.

The existence of a slowly (*adiabatically*) varying energy surface is one way of establishing and presenting the fact that the dynamical behavior of the network, in the vicinity of each member of the sequence and for a time of order τ, is robust to various types of noise and disruption, much as would be the situation in a symmetric network. See e.g., Chapter 7. Moreover, if the noise is stochastic, as in the case of the Glauber dynamics, Eq. 3.23, representing the noisy synapses of Section 2.1.3, then the above discussion can be extended to show that there would be an effective 'temperature' and a *free-energy* which is

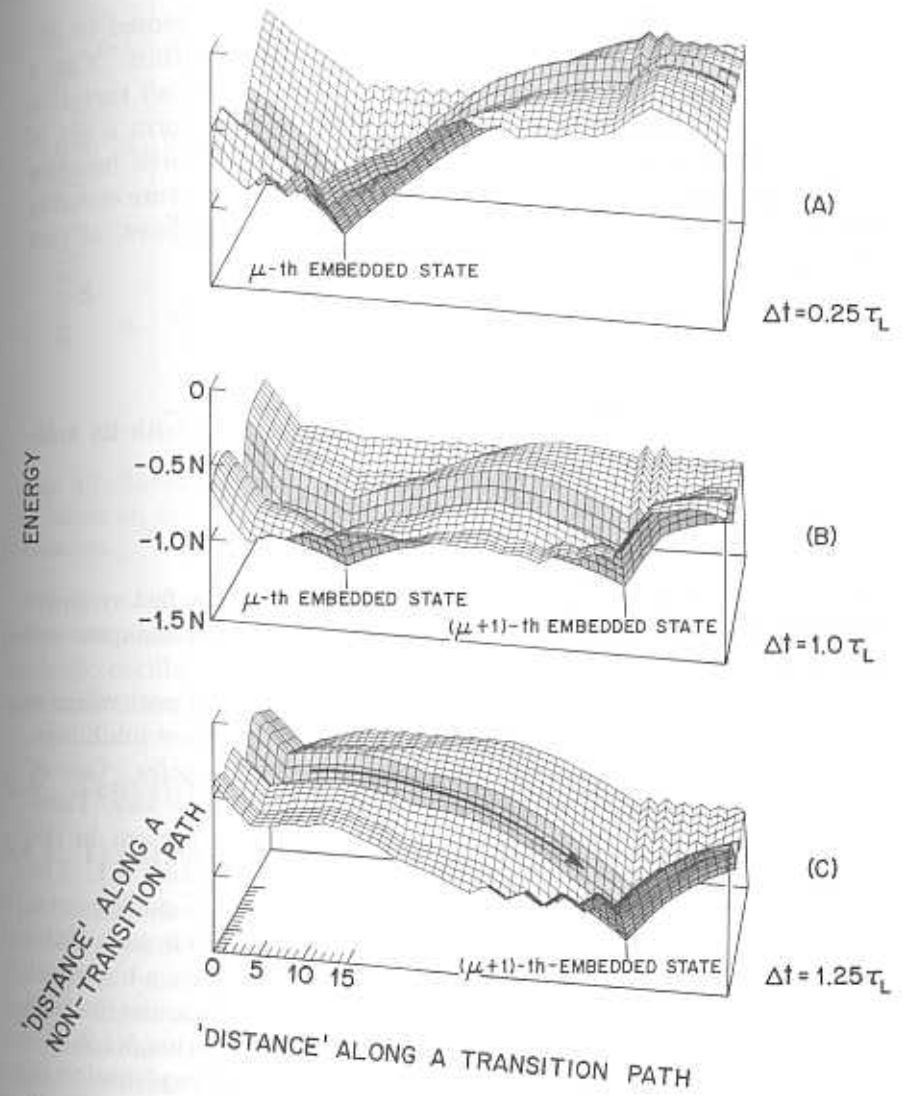


Figure 5.6: Adiabatically varying energy surface describing the dynamics of a network storing temporal sequence. (From ref. [9], by permission.)

minimized by a relaxation process, for periods of length τ. Within these durations the landscape picture is valid and useful.

5.3.5 Bi-phasic oscillations and CPG's

In Chapter 4 we observed that when a pattern $\{\xi_i\}$ is stored in the network, its *anti-phase* $\{-\xi_i\}$ is automatically stored with it. This a result of the symmetry of the dynamical process. One can therefore combine pairs of anti-phasic patterns in a network to form a set of oscillations with a cycle of two, each going back and forth between one of the patterns and its anti-phase. The synaptic structure creating k 2-cycle oscillations, out of the p stored patterns will have, as fast synapses, the usual (Eq. 5.1):

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$$

and the slow ones will couple the first k patterns, each with its anti-phase, according to

$$J_{ij}^t = \lambda \frac{1}{N}\sum_{\mu=1}^{k} \xi_i^\mu [-\xi_j^\mu],$$

which is just the negative of the corresponding part of the fast synaptic efficacies. Note that in this special case the transition synapses are symmetric.

In this network each of the synaptic connections has a part which is excitatory in the short term and following a delay becomes inhibitory, or vice versa. There are k oscillating sequences – *2-cycles*. One of them is selected by the incoming stimulus in an associative way. These oscillations should be distinguished from the 2-cycles present in the synchronous network, which have been described in Section 4.1. The latter are cycles of states in which the network spends a single neural cycle-time. In the present oscillations, the network spends in each of the two anti-phasic states of the cycle a time τ, which can have a biological *meaning*. It can, for example, control the rhythmic muscular motion of an organ, as does a CPG, which was described in Section 5.1.2.

Such oscillations have in fact been proposed as an explanation of the CPG in the escape swimming of *tritonia*[9]. A second glance at Figure 5.1 may now reveal that the four neuron groups oscillate between two anti-phasic states – in one neurons 1 and 2 have bursting activity and neurons 3 and 4 are essentially quiescent, in the other, 3 and 4 are bursting and 1 and 2 are relatively tranquil. The theoretical network has most of its parameters – synaptic strengths, thresholds, synaptic

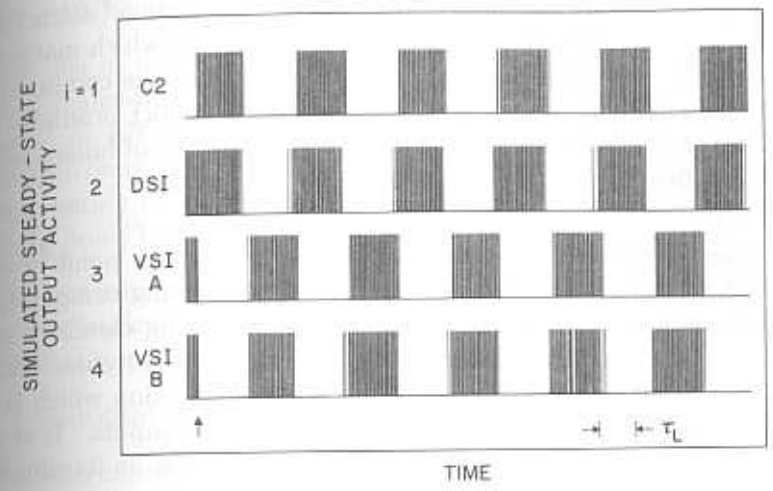


Figure 5.7: Simulated output activity of model ANN of the CPG in *tritonia*. The labels on the left correspond to those in Figure 5.1. (From ref. [9], by permission.)

delays etc. – taken directly from the biological system. It produces the bi-phasic oscillations of Figure 5.7 in impressive agreements with the measured biological activity of Figure 5.1.

5.4 Tentative Steps into Abstract Computation

5.4.1 The attempt to reintroduce structured operations

The general view taken in our discussion so far has been that ANN's identify *cognitive events* when they arrive rapidly enough to their attractors or quasi-attractors. The relation of the binary string composing the input stimulus to the content of the attractor has not been interpreted as a *computation* in any structured sense. See e.g., the discussion in Section 1.5.2. To recapitulate, a computing relationship has to be sensitive to the variation of single bits in the input data. Compare, for example, the *computing* network described in Section 1.5.2. This is, of course, just the antithesis of the whole concept of ANN's, whose pride is robustness and extensive associative basins. The error-correcting property of neural networks is synonymous to the insensitivity of the output to variations in the input.

- To propose that the binary word which represents an attractor is the outcome of a computation on a binary word which makes the network flow to that attractor, is to imply that the computation (represented by the organization of the network) produces the same outcome for every one of the vast multitude of binary words which flow into the same attractor.

In a sense pattern recognition is, from the vantage point of conventional computers, an elaborate and often frustrating computation. ANN's treat it as an elementary 'reflex', compressing clouds of possible input stimuli in almost no time into single prototypes. When considered as such, pattern recognition is a computation which is insensitive to a wide distribution of **closely spaced** inputs. It is out of such structureless building blocks that computation, in its simplest, abstract, algorithmic form, must be reconstituted.

One major problem facing this construction is the choice of a meaningful yet manageable abstract operation to be implemented. Both requirements depend strongly on which operation is considered primitive and which is supposed to be structured. In our case the primitives are the recall and classification of single patterns and of temporal sequences. In other contexts, this may be a very structured set of logical operations, as in a digital computer, or a set of parsing rules, as in a compiler. Fodor[29] considers that mind operates essentially like a compiler, with a more elaborate syntax. He is inspired to write his *Modularity of Mind* by a comment of Garrett's who suggested that "what you have to remember about parsing is that basically it's a reflex."

If parsing is a reflex, as it is for a compiler or for Fodor's man, then computational constructs would be linguistic compositions of a rather high level. The type of computational constructs that one contemplates starting from a set of logical operations is of a much lower level, and a compiler is considered a major achievement. It appears to us that in the universe of ANN's the task is different again. If one considers arguments as those of Shanon, discussed in Section 1.5.3, it is not quite obvious that any of the strict computational faculties mentioned above should be attempted. In fact, it may be that a much more diffuse and not fully recursive operational structure is a better approximation to observations on mind. Yet, some structured abstract computations must be shown to coexist, even within an extreme structureless interpretation of ANN's.

5.4.2 ANN counting chimes

With this in mind, we describe now what may be the simplest non-trivial realization of a reconstruction of an abstract operation in the context of ANN's[30]. The particular choice was inspired by Hebb's reflection on counting chimes, which was quoted in Section 5.1.3, above. The question mentioned by Hebb, the counting of identical chimes to tell the hour, has several dimensions

- Recognition of the arrival of several identical stimuli requires some mechanism of short term memory that does not transform into long term memory.
- Identification of a generic chime in order to provoke counting.
- Discrimination between different temporal sequences of chimes, according the *abstract* property which is their cardinal number.
- Robustness to variations in the sound, period or duration of each of the chimes. Yet, this robustness should be limited, in accord with observed cognitive instances.

It seems appropriate at this point to recapitulate the methodological disclaimer of Section 5.1.

One may, of course, invoke *synaptic plasticity* to account for chime counting. Synaptic plasticity has been employed by neurobiologists as well as by theoreticians to explain short term memory[31] and the storage and recall of temporal sequences[19]. To the best of our knowledge, there is no empirical evidence or theoretical argument against frivolous application of synaptic changes. Yet, it seems more prudent to reserve them for the more complex tasks of learning. It is gratifying that a rather delicate task, which represents all four desiderata listed above can be described by a network with a fixed set of synapses. In fact, recent theoretical progress indicates that one may even do without delayed synapses[32], whose role is taken up by noise. All of which goes in a direction of greater parsimony in neuro-biological elements, leaving more of them for the overbearing task of learning.

5.4.3 Counting network – an exercise in connectionist programming

It would be a rather elementary assignment to program a conventional computer to count the number of chimes arriving at a microphone

which is connected to the computer through some interface. Such a system may have some difficulty identifying the wide range of acoustical data that would naturally be considered as legitimate chimes. An ANN is particularly well suited for this task. It will be very sensitive to physical disruption. Above all it will operate along principles that do not even remotely resemble a plausible functioning of a cortical culture. In other words, in what follows we do not necessarily suggest that the ANN proposed is the most efficient device for telling the hour. Rather, we submit the following two points:

- A rather randomly connected network can be organized to perform the structured task of counting identical stimuli.
- The organization of the network (i.e., its connections) is a special type of programming.

Still conditioned by conventional programming, we proceed by sketching the requirements of the task (its flow chart) in terms of familiar modules (subroutines). The building blocks we have are:

1. Pattern recognizer.
2. Sequence generator.

One clearly needs the first to identify a chime, as such. The second will be employed with a slight modification. The sequence of *numbers* should be available, but it should not be spontaneously generated. Counting, the transition from number to number should proceed only under the impact of arriving chimes. But this is a minor change in the available temporal sequence retriever, namely the ratio of the transition synapses to the stabilizing synapses should be below its critical value λ_c of Section 5.3.2. Then, when input stimuli stop, the network will remain for a significantly longer time in the last quasi-attractor it has reached, which may be identified as the tally – the hour. What this ensures is only that transitions will **not** take place without chimes. To show that transitions will take place under the impact of chimes is the main aim of this section.

Having constructed an ANN that will make transitions down an ordered sequence of quasi-attractors, which can be identified as *numbers*, one must still ensure that it will recognize the first chime as special so that equal numbers of chimes will end up in the same attractor.

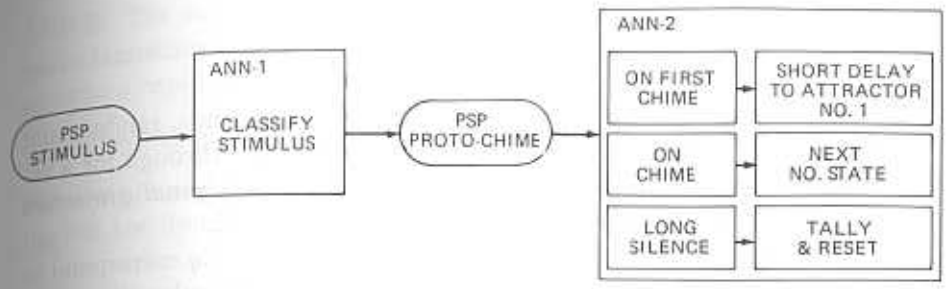


Figure 5.8: The logical operation of the two-tiered ANN counting chimes.

All these requirements and components are sketched in the Figure 5.8, which is a hybrid flow-chart and ANN flow.

5.4.4 The network

As is indicated in Figure 5.8, the construction is based on one simple element of architecture. There is an associating network at the lower level, ANN-1, whose task it is to associate the class of chime-like stimuli with counting. This can be a standard ANN. It should produce the **same** output for every chime, and communicate it to the counting ANN-2. The communication is by means of PSP's from the lower ANN to the counting network. The standard set of PSP's arriving at the counting network must:

1. Initiate the counting.
2. Identify the beginning.
3. Prompt transitions.

In a sense ANN-2, Figure 5.8, is a *universal* counting network. Different sets of stimuli to be counted can be identified by different lower level associators, of type ANN-1 and communicated for counting to ANN-2.

The connection matrix is constructed as in Section 5.3.2, out of fast and slow synapses whose form was given in Section 5.2.2. They are reproduced below.

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu$$

$$J_{ij}^{t} = \frac{\lambda}{N}\sum_{\mu=1}^{q}\xi_i^{\mu+1}\xi_j^{\mu}$$

where $\lambda < \lambda_c$. The patterns stored are, for convenience, random and uncorrelated. The ensuing discussion can be carried through for a set of correlated patterns, using, for example, the *orthogonal projection* synaptic matrix of Kohonen[33], as in Appendix 5.6.2.

The network states in the sequence $\{\xi_i^{\mu}\}$ for $1 \leq \mu \leq q$ correspond to numbers in a sense on which we elaborate in the next subsection. The prolonged presence of the network in any of these states allows for a cognitive event related to the number which equals the position of the quasi-attractor in the sequence.

To deal with the identification and special treatment of the first chime, we proceed as follows: Suppose that ANN-1, on identifying a chime, communicates PSP's η_i. We add an additional quasi-attractor to the counting network by adding to its synaptic connections the term

$$J_{ij}^{0} = \rho\xi_i^{0}\xi_j^{0}, \tag{5.12}$$

where the attractor ξ^0 is closely correlated with the incoming distribution of PSP's, namely

$$\xi_i^{0} = \mathrm{sign}(\eta_i).$$

It is random and uncorrelated with the q patterns in the sequence of 'numbers'. The synapses of Eq. 5.12 are instant, and thus stabilizing. They stand out in the special amplitude of their efficacy, ρ, which will be chosen to be especially weak, for reasons which will be explained below.

The pattern ξ^0 is not to be an attractor, but rather a quasi-attractor. It is appended to the front end of the sequence of 'number states' by transition synapses of the usual form

$$J_{ij}^{t0} = \lambda\xi_i^{1}\xi_j^{0}, \tag{5.13}$$

but with a relatively short time delay, τ_0. We impose

$$\tau_0 \ll \tau.$$

The role of this delay is to allow for the correction of errors on the initial input of the chime stimulus, as communicated by ANN-1 to

ANN-2. The value of λ is chosen to ensure that there are no spontaneous transitions in the sequence $1, \ldots, q$. The value of ρ is chosen to provoke a rapid, unaided transition $0 \rightarrow 1$.

- The four equations for J_{ij} are the program.

We now proceed to study the operation of the program, which is nothing but the dynamics of the ANN with this set of synapses.

5.4.5 Its dynamics

Given the synaptic efficacies and time delays, one can immediately write down the PSP induced on every neuron when the network is in any network state $\{S_i\}$. It is a simple extension of Eq. 5.6, which reads:

$$h_i(t+1) = \rho m^0(t)\xi_i^0 + \sum_{\mu=1}^{p} m^{\mu}(t)\xi_i^{\mu} + \lambda\left(m^0(t-\tau_0)\xi_i^1 + \sum_{\mu=1}^{q} m^{\mu}(t-\tau)\xi_i^{\mu+1}\right) \tag{5.14}$$

If the system is in state $\{S_i = \xi_i^{\nu}\}$ for $1 < \nu < q$ and if $\lambda < \lambda_c$ (=1 when overlaps between the random patterns can be neglected), then according to the discussion in Section 5.3.2 the system will remain indefinitely in that state, as required.

The central point is that the arrival of an identical input ξ_i^0 will produce a transition from any attractor to the next one, down the sequence. This can be seen as follows: Suppose that the network has been in state $\nu > 1$ for a time duration $> \tau$ and a chime arrives. In that case (neglecting fluctuations) the only non-zero overlaps are $m^{\nu}(t)$ and $m^{\nu}(t-\tau)$. Both of these are equal to unity. The PSP is therefore

$$h_i(t+1) = \xi_i^{\nu} + \lambda\xi_i^{\nu+1} + h\xi_i^{0}. \tag{5.15}$$

The first two terms are familiar from Eq. 5.8. The last one is the *standard* potential input due to the arrival of the chime. Recall that we have not only assumed that the q patterns in the chain are uncorrelated, but also that they are all uncorrelated with the state of the chime attractor.

Since all three states in Eq. 5.15 are random and uncorrelated, one concludes that

1. When the network is in state ν, i.e., $S_i = \xi_i^\nu$, one half of the neurons are already in neural state $\nu+1$, because for 50% of the sites $\xi_i^\nu = \xi_i^{\nu+1}$.

2. For the remaining $\frac{1}{2}N$ sites

$$S_i = \xi_i^\nu = -\xi_i^{\nu+1}.$$

For those sites

$$h_i(t+1) = (\lambda - 1)\xi_i^{\nu+1} + h\xi_i^0.$$

3. For one half of these remaining $\frac{1}{2}N$ sites the two patterns 0 and $\nu+1$ are identical, i.e.,

$$\xi_i^0 = \xi_i^{\nu+1}.$$

At these $\frac{1}{4}N$ sites the PSP is

$$h_i(t+1) = (h + \lambda - 1)\xi_i^{\nu+1}.$$

Note first that one must choose

$$h < 1 + \lambda,$$

or else Eq. 5.15 will conduct the network into pattern 0 whenever it tries to make a transition under the impact of a chime. On the other hand, the choice

$$h > 1 - \lambda,$$

will ensure a safe transition into state $\nu+1$. Thus we choose

$$1 - \lambda < h < 1 + \lambda. \tag{5.16}$$

Let us follow the transition: If the updating were synchronous and the chime sounded for a single neural cycle-time, then after one updating round, 75% of the neurons would be aligned with pattern $\nu+1$, and 75% of them would be aligned with pattern 0. This, in turn, implies that

$$m^0(t+1) = m^\nu(t+1) = m^{\nu+1}(t+1) = 0.5.$$

Hence, at the next updating, the PSP's will be, according to Eq. 5.14

$$h_i(t+2) = \tfrac{1}{2}\rho\xi_i^0 + \tfrac{1}{2}\xi_i^\nu + (\tfrac{1}{2} + \lambda)\xi_i^{\nu+1}.$$

Finally, choosing

$$\lambda > \rho, \tag{5.17}$$

which we will find quite handy for another task below, all h_i's will be aligned with pattern $\nu+1$ after the second updating sweep, and consequently all neurons will follow suit. The discussion at the end of Section 5.3.2 should be sufficient as an argument that what has been shown above to hold for parallel dynamics holds *a fortiori* for sequential dynamics. For a systematic study of this dynamics, see e.g., ref. [34].

It is interesting that the only requirement for the proper operation of the enumeration is that the impacting stimulus be uncorrelated with the 'number states'.

To complete the story, we must check that things start right. We assume that a proper initiation takes place with the network in states which are uncorrelated with any of the 'number states'. This is a rather plausible requirement from the psychological point of view. After all, it should be rather difficult to start counting a sequence of stimuli if one is in the midst of counting another. Before the arrival of the first chime we must have, for $\mu = 1, \ldots, q$,

$$m^\mu = 0.$$

The contribution to the PSP from Eq. 5.14 will be negligible and with the arrival of the chime

$$h_i(t=0) = h\xi_i^0$$

and the network will settle into the state number 0. Following a short delay, τ_0, the network will make a spontaneous transition to the first 'counting state', provided

$$\rho < \lambda.$$

From here on the network enters its normal counting course.

This concludes the demonstration that the *program*, expressed in the present context as a set of connections, should work. That it actually does work will be shown below by simulations.

5.4.6 Simulations

The above analysis of the dynamics of the counting network has been restricted to the case in which fluctuations could be neglected. Fluctuations can have several origins,

1. Noise in the elements of the network – temperature.
2. Finite N, promoting exceptional correlations.
3. High loading (p/N) – accumulation of a large number of small random overlaps between the patterns.

It turns out that fluctuations have a number of common effects: They facilitate transitions between consecutive quasi-attractors. This leads to a lowering of the critical value of λ, above which the transitions become spontaneous. Eventually, when fluctuations become large enough, they destabilize the attractors and destroy the system's associative recall. The universality of the phenomenon that noise stimulates transitions can be understood rather naturally in the present context. In fact, the operation of the counting network itself depends on noise, since the transitions are prompted by a stimulus which relative to the content of the quasi-attractor states is random noise. A similar effect takes place in the recently proposed model[32] of a network which recalls temporal sequences, without invoking delaying synapses.

Simulations are performed by choosing a set of p random patterns and generating the synaptic couplings as described in Section 5.4.4, with q of the patterns connected in a linear temporal sequence. An additional random pattern is designated as number 0 and is connected to the sequence as described in Section 5.4.4. The parameters λ, ρ, τ, τ_0 and h are set inside their appropriate windows. The network is then prepared in a random state, uncorrelated with any of the stored patterns, $0, \ldots, p$.

At time t=0 a chime stimulus is introduced, represented as a set of N PSP's which are proportional to the activities in state 0, i.e.,

$$h_i^0(t=0) = h\xi_i^0.$$

The network then develops under its own dynamics, Eq. 5.14, in an asynchronous, sequential updating procedure. When the next chime is sounded, the same external PSP, h^0, is added to the PSP's generated within the network as in Eq. 5.15. The stimulus PSP is added for a single cycle-time. The overlaps of the current state of the network with the patterns representing the 'number states' are represented in Figures 5.9 and 5.10 as a function of time.

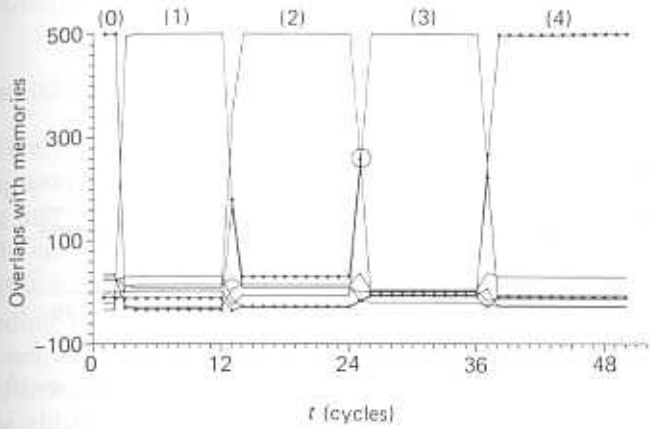


Figure 5.9: Simulation of a counting network (N=500, p=7, λ=0.7, ρ=0.5, h=0.4, τ=7). Chimes strike every 12 cycles. Plateaux at maximal overlaps are consecutive recognized numbers. Marked curves are overlaps with state 0 and with the state of final count. The circle indicates a moment in which the network enters a mixture of three states, during a transition from 1 to 2.

Figure 5.9 presents a counting run with relatively low fluctuations. Chimes strike, for a duration of a single cycle, once every 12 cycles and the delays are τ=7 and τ_0=3 cycles. The remaining fluctuations impose a value of λ=0.7, rather lower than 1. Note that h is in the window defined by Eq. 5.16 and ρ satisfies 5.17. Following the time dependence of the seven overlaps, one observes the network moves rapidly into state 0 and from there, spontaneously, into 'number state' 1, which is retrieved perfectly. It remains in state 1 until a chime arrives, at which point the network moves into a state which is a mixture of three states – 1, 2 and 0, surrounded by a circle. With each of them it has an overlap of about 0.5. See e.g., the discussion following Eq. 5.16. Subsequently, the network relaxes into the next number state, where it remains pending the next chime. Number state 1 as well as the next three states are perfectly retrieved, for a duration which should

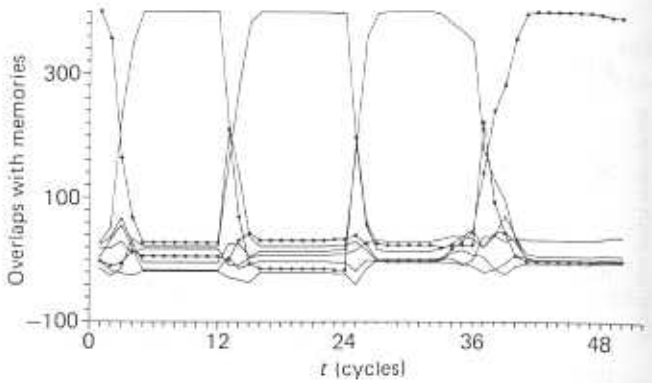


Figure 5.10: Counting at high storage levels. N=400, p=40, q=10. λ=0.35, ρ=0.5, h=0.9, τ=7 and chime period is 12.

suffice for their recognition. When the network arrives in state number 4, corresponding to the last chime, it remains there for an indefinitely long time, because the mechanism resetting the counting network has not been simulated.

Figure 5.10 presents the operation of a counting network at much higher storage - $\alpha = p/N$ =0.1. Ten of the patterns are strung in a temporal sequence. The fluctuations lower the value of λ_c, and hence λ=0.35!. See e.g., the discussion at the beginning of the present subsection. As a consequence, the amplitude of the chime must be increased, to satisfy the constraint Eq. 5.16. It is h=0.9. The constraint Eq. 5.17 would have implied a very low value of ρ, yet we chose ρ=0.5, as in Figure 5.9. That it can still work is due to the fluctuations, which assist in the transition from state 0 to state 1. This is a reflection of a real limitation of this model network, related to the mechanism of the initiation of counting. Higher storage increases fluctuations and lowers λ. This entails a lowering of ρ, but if ρ becomes too small, then the same fluctuations quickly destabilize the 0-state, and counting does not start properly. A precursor of this can be seen in Figure 5.10. Unlike the situation in Figure 5.9, the state 0 begins to disintegrate before the short time delay τ_0=3 has gone by. This implies that the 0-state is not stabilized by the fast synapses, and a further decrease of ρ would result in a total loss of coherence. The end result is a decrease in the *storage capacity* of the network, namely the maximal number of patterns that

can be stored without severely impairing its performance. The general topic of *storage capacity* is the subject of the next chapter.

5.4.7 Reflections on associated cognitive psychology

The relatively simple network described above satisfies the requirements imposed on it. But it does more. It emulates a number of features that one would have expected of a cognitive system; it reproduces some of the limitations of such systems and shares a few acceptable developmental suppositions.

According to our cognitive interpretation, the presence of the network in each of the 'number' quasi-attractors is long enough for the output neurons to be able to identify the particular cognitive event. When the chimes stop, the network will spend a significantly longer time in the last attractor. This again is an identifiable event, which may be biologically distinguished from the numbers encountered on the way. After a while, significantly longer than is required for identification of the final count, the network should shift from the last 'number state' into a state which is uncorrelated with any of the states in the sequence. There it is ready again to start counting properly.

Given the mechanism and the interpretation of the dynamics of the counting network, it is clear that if two stimuli arrive within a short time interval only one will be counted. Any stimulus which arrives before the delayed synapses become effective will not suffice to affect a transition. The minimal separation must, therefore, be the synaptic delay, The separation between stimuli cannot be too long either, because, within the present interpretation, if the network remains in the same attractor beyond a certain duration, that number state is identified as the final count. The time intervals between chimes are flexible, as long as they remain between the synaptic delay, on the short side, and the final count time on the long side. These are features which are rather reminiscent of common experience.

The next issue is the meaning assigned in the present model to the 'number states'. The fact that one usually does not expect to be able to count a very large number of chimes, or any other set of stimuli which are not especially organized for counting, lends some plausibility to an account in which there is one state for one number. Yet, it might be objected that a limited set of numbers needs few bits, and hence few neurons, for its representation. In fact, while we do not intend

such networks to describe counting of more than a few dozen, we do expect that the counting network should comprise hundreds, if not thousands, of neurons. Each 'number state', then, is represented by this vast number of bits.

This feature is related to the wider issue of *mental representations* raised in Section 1.5.3. The attitude informing us here is that the cognitive event associated with an attractor is rich and replete with context. One may speculate, for example, that the conscious awareness of a number may be accompanied by a visual image of its shape, by elements which may lead to the production of its sound, as well as many other relics of contextual data which might have been imprinted in the process of learning. In point of fact, there is no reason why the 'representation' of a number should be any poorer than that of any other concept.

Clearly, the above description cannot possibly be applicable to a different type of familiar counting, generating and reciting numbers, which can be carried on almost indefinitely. In order to perform such long counting operations, one seems to be using a very different type of activity from that of counting chimes. The impression is that of an organization of an abacus. How one organizes a neural network which operates as a dynamic abacus is something we are only beginning to understand.[3]

The issue of the content of the number attractor states connects rather naturally, though very tentatively, with a developmental subject. The construction of the 'program' somehow presupposes the presence of a temporal sequence of number states. This rather weighty requirement appears quite natural in the context of a naive observation of the cognitive development of children. Infants can and do learn to recite relatively long sequences of numbers, to identify their visual forms and to associate them with the corresponding sounds. The process of enumeration – the assignment of cardinal numbers to sets – is a much later development[35].

Once again, one should recall the methodological disclaimer. We do not, by any stretch of the imagination, purport to describe children by the little model counting chimes. Yet it is of interest to observe that a number of features, which one usually associates with human cognitive

[3]Preliminary abacus networks have been simulated in collaboration with M. Usher.

psychology, do find a simple and well defined echo in this version of a neural network.

5.5 Sequences Without Synaptic Delays

5.5.1 Basic oscillator - origin of cognitive time scale

The network to be described below[32] marks important new dimensions into which ANN's can grow. It has a number of attractive features:

- It does not require delaying synapses, and consequently reduces the amount of necessary special purpose elements.
- It allows for a continuous range of *cognitive times*, namely of possible periods of presence in each quasi-attractor.
- It allows for an interesting interpretation of some of the anatomical features in the neural connectivity in various brain areas.
- It provides an additional instance of the instrumental role of synaptic noise (temperature) in biological systems, in addition to the elimination of spurious states.

The main disadvantages we can perceive in such model networks are:

- the very fine tuning of the noise parameter required to achieve long time periods;
- the elaborate synaptic structure required to eliminate spurious attractors;
- the rather low storage capacity.

Yet, it may be that oscillators of this type can be used sparingly.

The oscillator is a network storing a sequence of two patterns only. It consists of two sub-networks, each comprising $\frac{1}{2}N$ neurons. The two sub-networks, Figure 5.15 below, are internally fully connected[4] by *ferromagnetic* symmetric interactions, namely the neurons in each sub-network are connected to all other neurons in the same sub-network by

[4]In fact, the mechanism will operate even if only one sub-network is internally connected.

excitatory synapses. Our discussion of the fully connected Ising model, Section 3.2.4, has shown that each of the two sub-networks has two attractors. In the absence of noise (T=0), either all spins are up or all are down. Let us denote the fraction of up-spins in sub-network 1 by x^1 and in sub-network 2 by x^2. Both variables lie between 0 and 1. In terms of these variables, the two attractors of each sub-network a (=1, 2) are

$$x^a = 1 \quad \text{or} \quad x^a = 0.$$

Considered as a combined network there are four attractors, since any of the two attractors of one sub-network can be combined with any of the two attractors of the other. The four states will have the first $\frac{1}{2}N$ bits all either +1 or −1 and similarly for the last $\frac{1}{2}N$ bits.

Next the two sub-networks are coupled in the following fashion: Every neuron in network 1 has an excitatory connection of efficacy α with every neuron in 2. Every neuron in 2 connects in an inhibitory way to every neuron in 1. If sub-networks 1 and 2 are in a states with a given x^1 and x^2, respectively, then the local fields, or PSP's, on every neuron in the two networks will be

$$H_i^1 = x^1 - \beta x^2 - h, \quad H_i^2 = \alpha x^1 + x^2 - h. \tag{5.18}$$

The dynamics of the neurons in each of the sub-networks is just the Glauber (heat-bath) dynamics, Eq. 3.23. It is given in terms of the probability that the neuron will be in state S at time $t + \delta t$, if the combined network was in a state with (x^1, x^2) at time t, namely

$$\Pr[S_i^a(t+\delta t)] = \frac{\exp[H_i^a(t+\delta t)S_i^a(t+\delta t)/T]}{\exp[H_i^a S_i^a/T] + \exp[-H^a S_i^a/T]}. \tag{5.19}$$

This determines the time development of the network.

There are a few special features that should be observed.

- The synapses are not symmetric. The contribution of x^1 to H^2 is different from the contribution of x^2 to H^1, unless $\alpha = -\beta$. There is no detailed balance and one may expect a behavior richer than simple fixed points.
- Sub-network 2 affects the neurons in 1 in an inhibitory fashion, while sub-network 1 affects the neurons of 2 excitatorily.
- The subtraction h in each of Eqs. 5.18 plays the role of a threshold.

5.5.2 Behavior in the absence of noise

As $T \to 0$, the probabilities in Eq. 5.19 are unity if the argument is positive and zero if it is negative. Thus, S_i^a=1 if $H^a > 0$ (a = 1, 2), otherwise it will be −1. Correspondingly, σ_i^a, where (see e.g., Section 1.4.2)

$$\sigma = \tfrac{1}{2}(S+1)$$

will be either 1 or 0. Suppose now that $h \approx 0.5$ and that the parameters α and β are in the respective intervals $0 < \alpha < h$ and $h < \beta \leq 1$. It is then easy to see that three of the four ferromagnetic attractors mentioned above are perfectly stable. They are $V^2 = (x^1 = 1,\ x^2 = 0)$, $V^3 = (0, 1)$ and $V^4 = (0, 0)$. In other words, if the combined network is started in any one of these states, it will remain in it indefinitely. If it is started near one of them, it will drift to it rapidly. For example, if the network is in state V^4, then $H^1 = H^2 = -h$, and no neuron will assume the state +1 in the next time-cycle. Clearly, if in the initial state of either network a few neurons were in state +1, the two x's would be small, of order $1/N$. Hence, the two H's would still be negative and the next network state would be V^4, again.

When the network is in state V^2 – (x^1=1, x^2=0), one has $H^1 = 1 - h > 0$ and all S_i^1 will be +1 again; $H^2 = \alpha - h < 0$ and all S_i^2=−1. Similarly, in state V^3, where x^1=0 and x^2=1, $H^1 = -\beta - h < 0$ and $H^2 = 1 - h > 0$. In the absence of noise, this state is reproduced. One can easily extend these arguments to show that the noiseless dynamics will correct errors – recall associatively – in the neighborhood of these three states.

- The three states are real attractors at T=0. The time the network will spend in any one of them will be infinitely long.

The remaining ferromagnetic state – $V^1 = (1, 1)$ – is unstable, even in the absence of noise. In fact, if $x^1 = x^2 = 1$, then

$$H^1 = 1 - \beta - h < 0$$

$$H^2 = \alpha + 1 - h > 0$$

which implies that while sub-network 2 remains in its ferromagnetic state with x^1=1, sub-network 1 begins flipping into the opposite ferromagnetic state in which all neurons are quiescent (all spins are down). This new state is stable, as we have seen above.

5.5.3 The role of noise

In the presence of noise ($T > 0$) none of the attractors remains a real fixed point. There is always a finite probability that a spin will flip against its local field – that a neuron will fire below threshold. As we have seen in Section 3.2.4, the ferromagnet preserves such parameters as its mean magnetization despite the fluctuations. In the present context, this implies that when $\alpha = \beta = 0$, the mean values of x^1 and x^2 will drift independently, in the presence of noise, to attractor distributions with fixed-point values, which are the solutions represented in Figure 3.5. These attractors are stable to small deviations.

When the two sub-networks are coupled, as in Eqs. 5.18, the situation changes drastically. To see this, suppose that the number of neurons in each sub-network is very large. In that case, due to the theorem of large numbers, the probabilities in Eq. 5.19 will be equal to the fractions of spins in the corresponding states. In other words, the probability that at time $t+1$ spin i of sub-network 1 is in state S_i^1=1 equals the fraction of spins in that sub-network with S=1, namely x^1 at time $t+1$. Consequently, the Eqs. 5.18 and 5.19 can be rewritten in the form

$$x^1(t+1) = f\left(\frac{x^1(t) - \beta x^2(t) - h}{T}\right)$$

$$x^2(t+1) = f\left(\frac{\alpha x^1(t) + x^2(t) - h}{T}\right), \qquad (5.20)$$

where

$$f(x) = \frac{\exp(x)}{\exp(x) + \exp(-x)}. \qquad (5.21)$$

This type of relation is called a *map*. It would trace just the type of flow trajectories of Figure 5.11, since for any point in that plane it gives the point at which the network will be at the next instant in time. The particular form of the equations implies parallel (synchronous) updating, because all neurons choose their new state on the basis of the same previous values of x^1 and x^2. One can write the corresponding equations for asynchronous Glauber dynamics[32]. They would read:

$$\Gamma \frac{dx^1(t)}{dt} = -x^1(t) + f\left(\frac{x^1(t) - \beta x^2(t) - h}{T}\right)$$

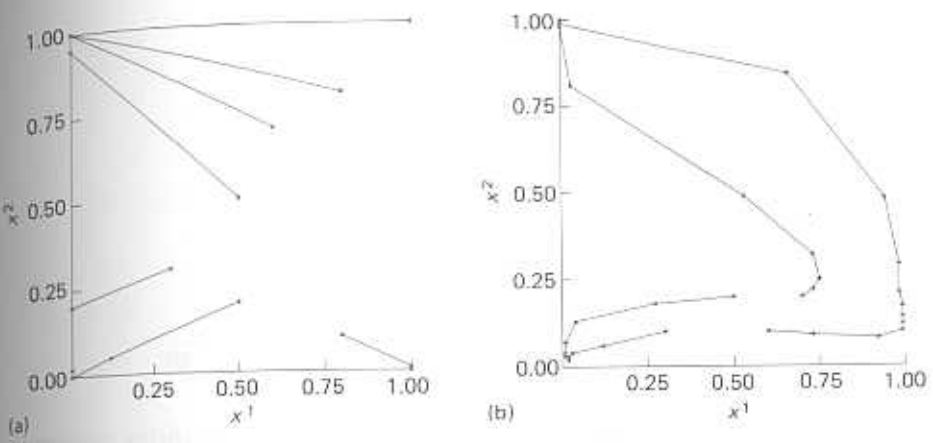


Figure 5.11: Flow maps in the x^1-x^2 plane. h=0.4; α=0.1; β=1.0. (a) T=0.1, (b) T=0.21. Different trajectories correspond to different initial conditions. Marks on the curves indicate successive temporal positions.

$$\Gamma \frac{dx^2(t)}{dt} = -x^2(t) + f\left(\frac{\alpha x^1(t) + x^2(t) - h}{T}\right), \qquad (5.22)$$

where Γ is the single neuron updating cycle-time. Compare Eq. 3.35.

In Figure 5.11, we draw the variation of x^1 and x^2 as a function of time, according to Eqs. 5.20, for three values of the temperature. In the presence of noise, the solutions of Eqs. 5.22 will be very similar. For each temperature, we plot the development of both variables starting from x^1=1 and x^2=0. At very low temperature, Figure 5.11(a), this state is stable for a very long time, which tends to infinity when $T < T^*$, or as $\alpha, \beta \to 0$. As the temperature becomes higher, T=0.21 in Figure 5.11(b), the initial state V^2 is destabilized and the network flows to state V^3 within some 7 cycle-times. The states V^2, V^3, V^4 are stable, as can be seen from the fact that the flows are into them. Then, in Figure 5.12 it is shown that the time which the network spends near the first state can be controlled by the temperature. As the temperature increases, the time decreases significantly.

The instability of the network state (x^1=1, x^2=0) does not necessarily imply that this state is not retrieved. In Figure 5.13 we plot the time development of the combined network for the parameters of Figure 5.11. For two sets of initial conditions (a) and (b) there is no retrieval. But, for (c) (x^1=0.7 and x^2=0.1), one observes that first the

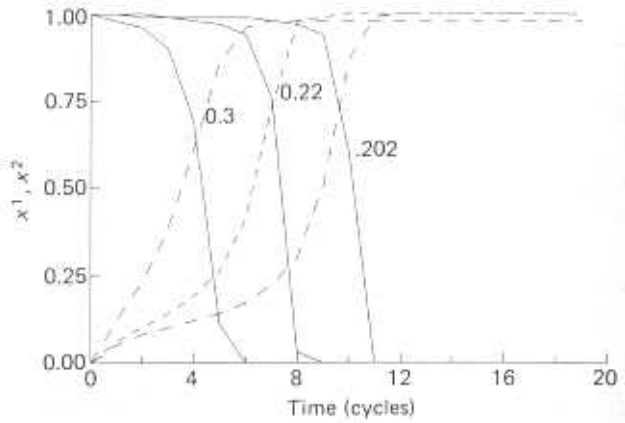


Figure 5.12: Time development of the fraction of up spins (firing neurons) in the two sub-networks, starting from (1,0). α=0.1 and β=1.0, h=0.4, for three levels of noise, T=0.3, 0.22, 0.202.

state with x^1=1 is clearly retrieved and only then begins the transition to the state $(0,1)$.

What about the state $(0,0)$, which was stable at T=0? It turns out that in the presence of noise there still is an attractor right near that state. This can be seen in Figure 5.13(b), where the same network is presented with the pattern (0.7,0.2). The network then runs to the attractor near the origin in the x^1–x^2 plane.

The state (1,0) was stable at T=0. The duration time $\tau(T)$ is finite when the noise crosses a certain threshold and is decreasing with increasing noise. In other words, it is noise which drives the transitions from one attractor to the next. Moreover, noise allows for control of the duration time in each quasi-attractor. This can be seen in Figure 5.14, where the function $\tau(T)$ [32], is presented. Here we observe that at relatively low levels of noise one has a sensitive control over the oscillator's period. In fact, the curve is fairly well approximated by the relation

$$\tau = \frac{K}{T}$$

with $K \approx 0.6$. This cannot be an exact relation because the instability of the attractor $(1,0)$ does not start before T crosses some threshold T^*, which can be deduced from a more elaborate stability analysis of the equations 5.20.

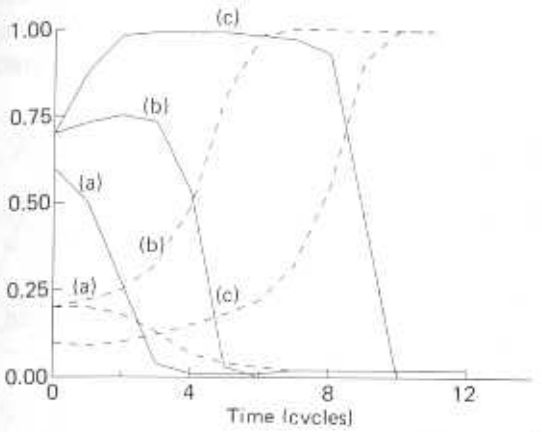


Figure 5.13: Retrieval of (1,0) from a pattern with errors. x^1 full curves, x^2 dashed. (a) Stimulus (0.6,0.2) – 40% errors. (b) Stimulus (0.7,0.2) – 30% errors. No retrieval. (c) Stimulus (0.7,0.1) – retrieval. (T=0.2).

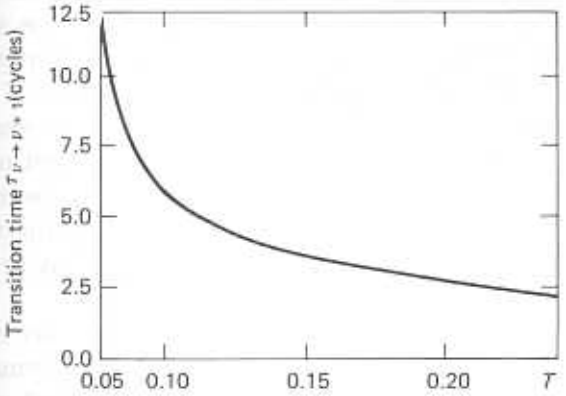


Figure 5.14: The duration time in a quasi-attractor in the two-state network vs temperature. The values of the parameters are α=0.1, β=1.0, h=0.35. (From ref. [32], by permission.)

5.5.4 Synaptic structure and underlying dynamics

We will now proceed to represent the dynamics of the oscillator in a form more reminiscent of temporal sequences in ANN's. In the present case, it is particularly convenient to use a $(0,1)$ for the neuronal states,

rather than $(1,-1)$. The two representations are, of course, equivalent, as we saw in Section 1.4.1. We must, therefore, display a synaptic structure in terms of stored *network states*. We choose the two patterns as:

$$\eta_i^1 = \begin{pmatrix} 1 & \text{if} & i \leq \frac{1}{2}N \\ \\ 0 & \text{if} & i \geq \frac{1}{2}N \end{pmatrix} \qquad \eta_i^2 = \begin{pmatrix} 0 & \text{if} & i \leq \frac{1}{2}N \\ \\ 1 & \text{if} & i \geq \frac{1}{2}N \end{pmatrix}. \tag{5.23}$$

They correspond to our previous V^2 and V^3. Out of these two patterns the synapses are formed as follows:

$$J_{ij} = \frac{2}{N}\left(\sum_{\mu=1}^{2} \eta_i^\mu \eta_j^\mu + \alpha \eta_i^2 \eta_j^1 - \beta \eta_i^1 \eta_j^2\right). \tag{5.24}$$

This is a very interesting synaptic structure. Note first that it includes no delays. All synapses act on the same time scale. The first term is of the stabilizing, symmetric type. These synapses act only within one of the two sub-networks and are purely excitatory (ferromagnetic). The second term represents the effect of the activity in the first network on that of the second, and is also excitatory. Its effect is to try to imprint the activity pattern in the first network onto the second. The last term is the back influence of the second network on the first. It is inhibitory (anti-ferromagnetic), trying to induce in the first network the opposite activity pattern to that present in the second.

The construction is sketched in Figure 5.15. There are two neuronal blobs. Each internally fully connected by excitatory synapses, which are drawn as open forks. The axons from sub-network 1 afferent on 2 are also excitatory, while the axons from 2 afferent on 1 are inhibitory, which is denoted by full circles. One can envision such an arrangement, which can become recursive to allow for richer networks, as consisting of two types of neurons in each module – excitatory and inhibitory. All excitatory neurons within a module synapse on each other and on the inhibitory ones. The excitatory neurons also make *efferent* connections on modules down the chain. The inhibitory ones do not connect inside the modules but only *efferently* onto modules from which their sub-network receives excitatory inputs.

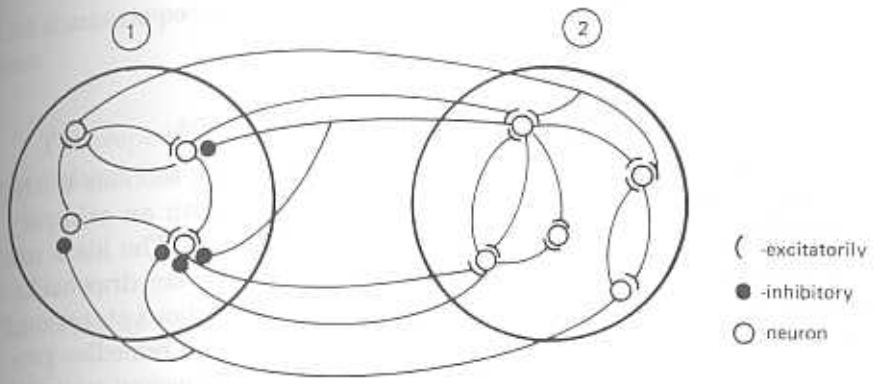


Figure 5.15: Two neural modules with a temporal sequence of two states. Each is internally fully connected by excitatory connections. Sub-network 1 affects 2 excitatorily while 2 affects 1 inhibitorily. Open forks are excitatory synapses, closed circles are inhibitory.

The $N \times N$ matrix J_{ij} for the combined network has the form

$$J_{ij} = \frac{2}{N}\begin{pmatrix} 1 & 1 & \dots & 1 & -\beta & -\beta & \dots & -\beta \\ 1 & 1 & \dots & 1 & -\beta & -\beta & \dots & -\beta \\ \cdot & & & \cdot & \cdot & & & \cdot \\ \cdot & & & \cdot & \cdot & & & \cdot \\ \cdot & & & \cdot & \cdot & & & \cdot \\ 1 & 1 & \dots & 1 & -\beta & -\beta & \dots & -\beta \\ \alpha & \alpha & \dots & \alpha & 1 & 1 & \dots & 1 \\ \alpha & \alpha & \dots & \alpha & 1 & 1 & \dots & 1 \\ \cdot & & & \cdot & \cdot & & & \cdot \\ \cdot & & & \cdot & \cdot & & & \cdot \\ \cdot & & & \cdot & \cdot & & & \cdot \\ \alpha & \alpha & \dots & \alpha & 1 & 1 & \dots & 1 \end{pmatrix}$$

If the network states are also represented by N-bit words of ones and zeroes, then the PSP arriving at any of the N neurons in the combined network is given by the usual summation

$$h_i = \sum_{i=1}^{N} J_{ij} S_j$$

and subtracting the threshold h leads directly to the equations 5.18. From here on the dynamics is standard.

5.5.5 Network storing sequence with several patterns

A generalization of the considerations of the preceding sections to the construction of a network which stores a sequence with an arbitrary number of memorized patterns has been suggested[32]. The ideas are quite stimulating yet the actual proposal has two severe drawbacks. The first is related to the appearance of pestering spurious states and the second is the extremely low storage capacity. The remedies proposed for the first difficulty lead to a rather subtly engineered synaptic matrix which seems to preclude robustness. In order not to prejudge the issue we shall refer the reader at this stage to the original study, ref. [32]. These comments are not intended as a depreciation of the connectionist mechanism for storage and retrieval of sequences without synaptic delays. The concept is almost surely of direct relevance to brain functioning. The modulated oscillator itself finds a very natural physiological counterpart in the breathing control system which has been very carefully studied in cats[36].

5.6 Appendix: Elaborate Temporal Sequences

5.6.1 Temporal sequences by time averaged synaptic inputs

The main issue is the generalization of Eq. 5.6 for the dependence of the afferent synaptic input into neuron i, ref. [24]. There is the fast, stabilizing part, which remains unchanged. But the slow part, which leads to transitions could, in general, have the form

$$h_i^t(t) = \sum_{i=1}^{N} J_{ij} \bar{S}_j(t) \tag{5.25}$$

where

$$\bar{S}_i(t) \equiv \int_0^\infty S_i(t-t')w(t')dt'.$$

These equations express the fact that neuron i receives its input potential as a weighted time average over the preceding activities of other neurons and the synapses can remember the past activities over a time interval much longer than the neural cycle-time. The particular case of

the sharp time delay discussed in the text corresponds to the extreme case

$$w(t) = \delta(t-\tau). \tag{5.26}$$

The response function of the slow synapses will be non-negative and normalized to unity, i.e.,

$$\int_0^\infty w(t)dt = 1.$$

Its structure defines a mean time interval τ, given by:

$$\tau = \int_0^\infty tw(t)dt,$$

which is the typical averaging time of the past neuronal activity contributing to the PSP's.

The arguments employed in Section 5.3.2 to describe the dynamics can be extended to the more general situation. Writing the expression for $h_i(t)$, which results from a transition synaptic matrix of the form given by Eq. 5.2 together with the contribution of fast synapses, in terms of state overlaps, one has:

$$h_i(t+\Delta t) = \sum_{\mu=1}^{p} \xi_i^\mu m^\mu(t) + \lambda \sum_{\mu=1}^{q} \xi_i^{\mu+1} \bar{m}^\mu(t) \tag{5.27}$$

where

$$\bar{m}^\mu(t) = \int_0^\infty m^\mu(t-t')w(t')dt'.$$

If the network, prior to time t_1, has been in states uncorrelated with the patterns and from t_1 to t_2 has been in state number ρ, then the contribution of the transition part to h_i in Eq. 5.27 will be

$$h_i^t = \lambda \xi_i^{\rho+1} W(t_1, t_2) \tag{5.28}$$

with

$$W(t_1, t_2) \equiv \int_{t_1}^{t_2} w(t)dt. \tag{5.29}$$

Suppose that the network is in state ρ $(1 < \rho < q)$, after transients have subsided. What we would like to know is how long will it remain in that particular quasi-attractor. The length of this time interval, τ,

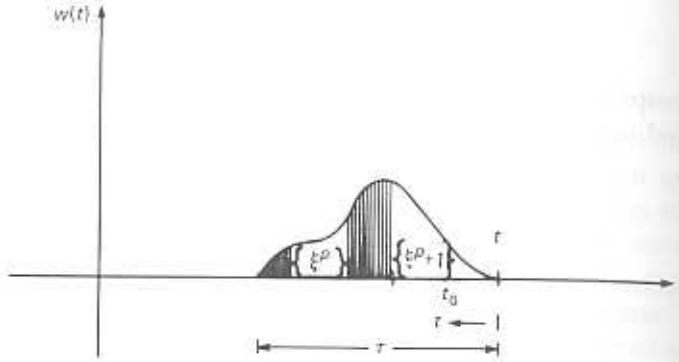


Figure 5.16: Contributions to the PSP at time t due to successive presence in **two** quasi-attractors with a weight function $w(t)$ whose variable runs from right to left starting at t.

will depend on two factors – the relative strength, λ, of the transition synapses to the stabilizing ones and the shape of the weight function $w(t)$. Shortly after its relaxation into pattern ρ, all PSP's will be consistent with the neural states in this network state. The longer the network remains in this state the larger becomes the buildup for a transition, via contributions such as Eq. 5.28.

The typical situation to be considered is depicted in Figure 5.16, where time is denoted by s and the present moment is indicated as $s = t$. The time dependence of $w(t)$ should be considered from right to left, since it samples past states, and $s = t$ is the zero point for the variable of w. The network has made a transition into state $\rho + 1$ a time t_0 ago, having arrived from an attractor corresponding to state ρ. We will assume that $w(t)$ does not overlap more than two patterns. In other words, the size, τ, of the support of $w(t)$, the interval in t where $w \neq 0$, must be less than twice the resulting period, t_0, which the network spends in any one quasi-attractor. How long can t_0 be before the network moves on?

At time t there are two terms which stabilize the presence of the network in state ρ.

1. The input from the fast synapses, contributing $\xi_i^{\rho+1}$ to the PSP, via the first term in Eq. 5.27.

2. The memory of the prior presence in state ρ, coming through the slow synapses, weighted by the hatched part of the curve $w(t)$ in Figure 5.16.

Its contribution to the PSP is

$$\lambda \bar{m}^{\rho}(t)\xi_i^{\rho+1} = \lambda \int_{t_0}^{\infty} w(t)dt\xi_i^{\rho+1},$$

according to the second term in Eq. 5.27, together with Eq. 5.29. On the other hand there is only one term working for a transition from $\rho + 1$ to $\rho + 2$. It comes through the slow synapses and represents the recent memory of the network's presence, for a time t_0, in state $\rho + 1$, under the unhatched part of the function w in Figure 5.16. The corresponding contribution to the PSP is:

$$\lambda \bar{m}^{\rho+1}(t)\xi_i^{\rho+2} = \lambda \int_{0}^{t_0} w(t)dt\xi_i^{\rho+2}.$$

The resulting PSP at time t is

$$h_i(t) = [1 + \lambda \bar{m}^{\rho}(t)]\xi_i^{\rho+1} + \lambda \bar{m}^{\rho+1}(t)\xi_i^{\rho+2}, \tag{5.30}$$

which will provoke a transition after a time t_0 at which the coefficient of the second term becomes as big as the first. One can easily convince oneself, much as in the case of the sharp time delay described in the text, that once the transition starts it becomes an avalanche. The time t_0, neglecting the short time needed for the transient, is the time the network will spend in each of the quasi-attractors. It is determined from the equation

$$1 + \lambda \int_{t_0}^{\infty} w(t)dt = \lambda \int_{0}^{t_0} w(t)dt,$$

which, due to the normalization of $w(t)$, is equivalent to

$$1 + \lambda = 2\lambda \int_{0}^{t_0} w(t)dt. \tag{5.31}$$

This is our final result. An immediate consequence is that

$$\lambda > \lambda_c = 1$$

because the integral on the right hand side cannot be greater than unity. Next, it can be applied to special cases. For example, if w is

given by Eq. 5.26, then the right hand side of Eq. 5.31 will be non-zero only if $t_0 > \tau$ and a transition can take place only if $\lambda > 1$, as we have seen in the text. The result $t_0 = \tau$ is, of course, consistent with our restriction $\tau < 2t_0$.

Another example is the step-function for w, namely

$$w(t) = \frac{1}{\tau} \quad \text{for } 0 < t < \tau$$

and w=0 otherwise. In that case Eq. 5.31 becomes:

$$1 + \lambda = 2\lambda \frac{t_0}{\tau}$$

or

$$t_0 = \tau \left(1 + \frac{1}{\lambda}\right).$$

It implies that $\tau < t_0 < 2\tau$, as required.

5.6.2 Temporal sequences without errors

The synaptic connection matrix, for both fast (stabilizing) synapses as well as for the slow (transition) synapses, which was used to generate the sequence in Figure 5.5 was a generalization of the orthogonal projection matrix[9] introduced in Section 4.2.3.

Suppose that there are p patterns stored in the network. Let them form, for simplicity, a single temporal sequence, so $p = q + 1$. The patterns in the sequence will be denoted, as usual, by N-bit words - $\{\xi_i^\mu\}$, with $\mu = 1, \ldots, p$ and $i = 1, \ldots, N$. As in Section 4.2.3, one first forms the correlation matrix of the p patterns,

$$C_{\mu\nu} \equiv \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu, \tag{5.32}$$

with both $\mu, \nu = 1, \ldots, p$. This is a $p \times p$ matrix.

The inverse of the matrix C, when operating on the original states, projects out a set of p states, $\{\eta_i^\mu\}$, orthogonal to the patterns. These states are given by:

$$\eta_i^\mu = \sum_{\nu=1}^{p} C^{-1}_{\mu\nu} \xi_i^\nu. \tag{5.33}$$

Clearly, they satisfy

$$\frac{1}{N} \sum_{i=1}^{N} \xi_i^\nu \eta_i^\mu = \delta^{\nu\mu}.$$

Using these orthogonal states, one defines the two sets of synaptic efficacies. The fast ones are defined just as in Section 4.2.3, namely

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \eta_j^\mu,$$

for $i \neq j$, and zero otherwise. The slow ones are defined in a similar way, by rewriting Eq. 5.2 in the form

$$J_{ij}^t = \lambda \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^{\mu+1} \eta_j^\mu \tag{5.34}$$

for $i \neq j$. These last synapses act with a time delay τ.

One can now proceed to obtain equations like 5.7, 5.8 or 5.9 without worrying about possible overlaps between the stored patterns.

Bibliography

[1] D.W. Tank and J.J. Hopfield, Neural computation by concentrating information in time, *Proc. Natl. Acad. Sci. USA*, **84**, 1896(1987).

[2] S. Ullman, *Interpretation of Visual Motion* (MIT Press, Cambridge Mass., 1979).

[3] S. Grilner, Locomotion in vertebrates. Central mechanisms and reflex interaction, *Physiol. Rev.* **55**, 247(1975).

[4] P.A. Getting, Mechanism of pattern generation underlying swimming in *Tritonia*. I. Neuronal network formed by monosynaptic connections, *J. Neurophys.* **46**, 65(1981).

[5] G.S. Stent, W.B. Kristan Jr., W.O. Friesen, C.A. Ort, M. Poon and R.L. Calabrese, Neuronal generation of the leech swimming movement, *Science* **200**, 1348(1978).

[6] J.C. Weeks, Neuronal basis of leech swimming: Separation of swim initiation, pattern generation and intersegmental coordination by selective lesions, *J. Neurophys.* **45**, 698(1981).

[7] P.S.G. Stein, A.W. Camp, G.A. Robertson and L.I. Mortin, Blends of rostral and caudal scratch reflex motor patterns elicited by simultaneous stimulation of two sites in the spinal turtle, *J. Neurosci.* **6**, 2259(1986).

[8] P.A. Getting, Mechanism of pattern generation underlying swimming in *Tritonia*. III. Intrinsic and synaptic mechanisms for delayed excitation, *J. Neurophys.* **49**, 1036(1983).

[9] D. Kleinfeld and H. Sompolinsky, Associative Neural Network Models for the generation of temporal patterns: Theory and Application to central pattern generators, *J. Neurosci.* (Submitted 1987).

[10] L. Wittgenstein, *Tractatus Logico-philosophicus* (Routledge & Kegan Paul Ltd, London, 1922).

[11] J.-P. Changeux, *Neuronal Man* (Oxford University Press, NY,1986).

[12] D.O. Hebb, *Essay on Mind* (Lawrence-Erlbaum Assc., Hillsdale NJ, 1980).

[13] J.A. Fodor, Information and association, in M. Brand and R.M. Harnish eds. *The Representation of Knowledge and Belief* (The University of Arizona Press, Tucson, 1986).

[14] E.R. Caianiello and L.M. Ricciardi, Reverberations and control of neural networks, *Kybernetik*, **4**, 10(1967); E.R. Caianiello, A. de Luca and L.M. Ricciardi, Neural networks: Reverberations, constants of motion, general behavior, in E.R. Caianiello ed., *Proceedings of the Ravallo School on Neural Networks* (Springer-Verlag, Berlin, 1967).

[15] S. Amari, Learning patterns and pattern sequences by self-organizing nets of threshold elements, *IEEE Trans. Comput.*, **21**, 1197(1972).

[16] W.A. Little and G.L. Shaw, A statistical theory of short and long term memory, *Behavioral Biology* **14**, 115(1975).

[17] J.J. Hopfield, Neural networks and physical systems with emergent selective computational abilities, *Proc. Natl. Acad. Sci. USA*, **79**, 2554(1982).

[18] J.J. Hopfield, Collective processing and neural states, in C. Nicollini ed., *Modeling and Analysis in Biomedicine* (World Scientific, NY,1984).

[19] P. Peretto and J.J. Niez, Collective properties of neural networks, in E. Bienenstock, F. Fogelman-Souliè and G. Wiesbuch eds. *Disordered Systems and Biological Organization* (Springer-Verlag, Berlin, 1986).

[20] I. Nebenzahl, Recall of associated memories, *J. Math. Biol.*, **25**, 511(1987).

[21] S. Dehaene, J.-P. Changeux and J.-P. Nadal, Neural networks that learn temporal sequences by selection, *Proc. Natl. Acad. Sci. USA* **84**, 2727(1987).

[22] T. Heidmann and J.-P. Changeux, Un modèle moléculaire de régulation d'efficacité d'un synapse chimique au niveau postsynaptic, *C.R. Acad. Sci. Ser.2*, **295**, 665(1982).

[23] M. Abeles, *Local Cortical Circuits* (Springer-Verlag, Berlin, 1982).

[24] H. Sompolinsky and I. Kanter, Temporal association in asymmetric neural networks, *Phys. Rev. Lett.* **57**, 2861(1986).

[25] D. Kleinfeld, Sequential state generation by model neural networks, *Proc. Natl. Acad. Sci. USA* **83**, 9469(1986).

[26] J.-P. Changeux, A. Klarsfeld and T. Heidmann, The acetylcholine receptor and molecular models for short- and long-term learning, in J.-P. Changeux and M. Konishi eds. *The Neural and Molecular Bases for Learning* (John Wiley and Sons, NY, 1987).

[27] T. Kohonen, E. Reuhkala, K. Mäkisara and L. Vainio, Associative recall of images, *Biol. Cybernetics*, **22**, 159(1976).

[28] W. Kinzel, Learning and pattern recognition in spin glass models, *Z. Phys.*, **B 60**, 295(1985).

[29] J.A. Fodor, *The Modularity of Mind* (MIT Press, Cambridge, Mass., 1983).

[30] D.J. Amit, Neural networks counting chimes, *Proc. Natl. Acad. Sci. USA*, **85**, 2141(1988).

[31] E.T. Rolls, Information representation, processing and storage in the brain: Analysis at the single neuron level, in J.P. Changeux and M. Konishi eds., *Learning* (Springer-Verlag, Berlin, 1986).

[32] J. Buhmann and K. Schulten, Noise-driven association in neural networks, Techniche Universität München, preprint (1987).

[33] T. Kohonen and M. Rouhonen, Representation of associated data by matrix operators, *IEEE Trans. Comput.*, **22**, 701(1973).

[34] H. Gutfreund and M. Mezard, Processing temporal sequences in neural networks, *Phys. Rev. Lett.*, **61**, 235(1988).

[35] J. Piaget, *Child's Conception of Number* (Humanities Press, Atlantic Highlands NJ, 1964).

[36] F. Bertrand, A. Hugelin and J.F. Vibert, A steleologic model of

pneumotaxic oscillator based on spatial and temporal distribution of neuronal bursts, *J. of Neurophysiology*, **37**, 1(1974) and J.F. Vibert, F. Bertrand, M. Denavit-Saubié and A. Hugelin, Three dimensional representation of bulbo-pontine respiratory networks architecture from unit density maps, *Brain Research*, **114**, 227(1976).

6

Storage Capacity of ANN's

6.1 Motivation and general considerations

6.1.1 Different measures of storage capacity

The properties of ANN's described in the preceding chapters should make them interesting candidates both for models of some brain functions as well as for technical applications in certain areas of computer development or artificial intelligence. In either case, one of the first questions that comes to mind is the storage capacity of such systems, namely the quantity of information that can be stored **and** effectively retrieved from the network. It is of primary interest to know, for example, how the number of possible memories, in single patterns or in sequences, varies with the number of elements, neurons and synapses, of the network[1].

The storage capacity of a network can be quantified in a number of possible ways. It must be expressed per unit network element. Here we mention a few such possible measures:

1. The number of stored bits per neuron.

2. The number of stored bits per synapse.

[1] It is one of the weaknesses of the PDP program that so far no technique has been proposed for estimating the scaling of the network capabilities with the number of its elements. It is partly a result of the fact that the corresponding capabilities of PDP networks are much more complex.

3. The number of stored patterns per neuron.

4. The number of stored patterns per synapse.

5. The number of stored bits per coded synaptic bit.

Any one of the items in this list must be supplemented by informational qualifications. It should be realized that the usefulness of any of these three quantifications is strongly dependent on the level of correlation between the stored bits. In a network that stores a number of highly correlated patterns, the retrieval of any one pattern would predict, with high probability, the structure of a pattern retrieved by any other stimulus. On the other hand, if the stored patterns are completely uncorrelated, then the retrieval from memory by one stimulus leaves the result of retrieval by another completely unknown. The informational store in the second case is obviously greater.

There is yet another dimension to the quandary. The informational content of an associative memory storage should depend also on the size of the basins of attraction of each of the stored memories. If retrieval requires that the *stimulus* differ from the *memory* in a relatively very small number of bits, then the process of retrieval adds almost no information to what has arrived in the stimulus. This again would be a rather impoverished case.

Not all these questions can be answered. Some remain open because of technical difficulties. Others await a clearer specification. Much of the progress that has been made in the study of ANN's has concentrated on questions of this kind and many results have been obtained. A fair amount of these results are analytic (see e.g., refs. [1,2,3,4,5,6, 7,8]). Further results are obtained by numerical simulation (see e.g., refs. [9,10]).

One should keep in mind that the choice of the measure for the storage capacity is strongly connected to the available resources, as well as to the task one assigns the network. If one is trying to answer very general questions about the brain, such as posed by Von Neumann[11] (see e.g., Section 6.1.2), without specifying the computational structure, in order to put a lower bound on the total amount of resources required, then the pattern structure becomes irrelevant. The appropriate measures would be either the first or the second in the list above. On the other hand, in designing an electronic device for pattern recognition which is to be fully connected, the main constraint is the difficulty in

implementing, in a reasonable volume, N^2 connections between N neurons . Hence, the crucial variables are the number of different patterns (which is to be maximized) and the number of neurons (which should be minimized). Consequently, one would like to optimize the third criterion. If it is easier to supply neurons than synapses in a sparsely connected network[6], then one would try to optimize the fourth criterion. Finally, if the *analog depth* of the synapses – the number of bits they can faithfully represent – is a valuable resource, then the relevant measure of storage would be the fifth one.

6.1.2 Storage capacity of human brains

Von Neumann[11] (pp. 63-64) has estimated the order of magnitude of the total number of information bits that may penetrate the central nervous system through the sensory organs in a human lifetime. The estimate is based on the neurobiologically determined mean rate of neural spike activity of some 14 per second, on the number of incoming channels, which he estimated as equal to the total number (10^{10}) of neurons in the brain, and on the mean length of a human life, which is about 10^9 seconds. He arrives at the figure of 10^{20} bits. This number would become 10^{21} if a cortex of 10^{11} neurons is considered.

As tentative as this estimate may be it does produce, via its order of magnitude, a pressure which models of brain must relate to. As far as ANN's are concerned, the general tenor of the results on storage capacity is that

- The number of uncorrelated **patterns** that can be stored in a network is proportional to the mean number of **synapses** per neuron in the network.

The proportionality constant α ranges between 0.14 (of ref. [1]) and 2 (of ref. [5]). A fully connected cortex with 10^{11} neurons, and 10^{11} synapses per neuron can store $\alpha 10^{11}$ uncorrelated patterns, each consisting of 10^{11} bits. Hence, a total of $\alpha 10^{22}$ bits, which seems plenty, even if the figure 10^{20} is taken literally.

But cortex is not fully connected. There are only about 10^4 synapses per neuron. We have been viewing it as a system of 10^7 elementary ANN's, each comprising 10^4 fully connected neurons. Such a system

can store and retrieve $\alpha 10^4$ patterns, of 10^4 bits, in each elementary network. Hence a total of

$$\alpha \times 10^7 \times (10^4)^2 = \alpha 10^{15}$$

bits. A similar order of magnitude results also if one considers the network of 10^{11} neurons as a uniform network with a mean of 10^4 synapses per neuron [6]. That system will stock $\alpha 10^4$ patterns of 10^{11} bits each – a total of $\alpha 10^{15}$ bits again.

Clearly, it is not the factor of 14 between the high and the low values of α which will close the gap between Von Neumann's 10^{20} and our 10^{15}. The explanation would lie in the consideration of the correlations between possible incoming information clusters. Once the network is endowed with some functional structure, such as content addressability, there is no conceivable reason for inputs which lie in the basin of attraction of some stored memory to be stored independently. Since *basins of attraction* (see e.g., Section 6.5.2) of uncorrelated patterns themselves increase linearly with the number of neurons in the network[12] there should be ample redundancy in the Von Neumann figure to allow for a comfortable closing of the gap.

A more empirical measure of storage capacity comes out of the experiments of Standing[13], for example. In these experiments, subjects are shown very many different pictures, as many as 10,000, and two days later are asked to *recognize* previously seen pictures out of pairs of pictures, one of which had been seen and one randomly chosen. Subjects seem to retain about 6,600 of them and the number of stored patterns seems to be still almost linearly increasing with the size of the *learning set*, showing no sign of saturation. Such numbers are very impressive indeed. Yet, if all these pictures are to be stored in a single network, as their rapid recognition indicates, then the estimates to be developed below lead to the following conclusion:

- The storage of 10,000 pictures can be affected in a network of between 5,000 to 100,000 neurons, depending on the learning procedure.

This estimate looks much less forbidding, especially if one keeps in mind that any significant correlations between the pictures reduce the number of required neurons.

6.1.3 Intrinsic interest in high storage

The discussion of the previous chapter emphasized the central role attributed to temporal sequences in the reconstruction of abstract computational capabilities in ANN's. The potential complexity, as well as the richness of the syntax of the resulting networks, depend to a large extent on the ability of the network to store and retrieve faithfully substantial numbers of patterns. One could, instead, resort to a system in which each pattern in the temporal sequence is stored in a separate sub-network[14]. But then the overall capacity of the cortex, or of a conceivable device, becomes unacceptably low.

But even prior to the consideration of temporal sequences, the attitude toward ANN's advocated here calls for networks which can store and retrieve substantial numbers of individual patterns. If ANN's are to be viewed as cognitive classifiers of stimuli, then one would naturally desire that each network, on recognizing a stimulus, would thereby discriminate it from the largest possible set of other classes of associatively grouped stimuli in the same category of input. Since reactions to a stimulus and higher level operations on a stimulus should be relatively specific, the lower the number of classes discriminated by a given network the higher the number of serial discrimination tasks that have to be performed. The price is time, which is where biology appears to be rather efficient. A similar motivation for large storage capacity comes from model networks which store and retrieve hierarchically organized data[15]. See e.g., Section 8.4, below. In this approach the stored patterns are correlated in a tree structure. Eventually, the richness of the tree is constrained by the storage capacity of the network storing the most detailed memories – the leafs of the tree – which are the most numerous.

6.1.4 List of results

In the present chapter, we will discuss the capacity for storing uncorrelated random patterns. In the absence of a motivated specific data structure for the memorized patterns, random patterns are a useful generic case to study. In some sense it is both generic and unique. The results have general implications and yet can be quite specific. A representative list of results would be composed of two general classes:

1. Estimation of the upper limit on the number of random patterns which can be stored and retrieved effectively, given a specific prescription for the synaptic efficacies in terms of the stored patterns.

2. Estimation of the upper limit on the number of stored random patterns which can be stored and recalled associatively with the most general set of synaptic efficacies J_{ij}.

In the first class, we describe results for coupling matrices which are symmetric and are a variation on the simple storage prescription:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu. \tag{6.1}$$

In order of increasing complexity, one has the following results:

- If the storage level p satisfies

$$p < \frac{N}{2 \ln N}, \tag{6.2}$$

then, in the limit $N \to \infty$, the probability that there will be an unstable bit in a network state corresponding to any one of the memorized patterns will vanish.

- If the storage level satisfies the stricter constraint:

$$p < \frac{N}{4 \ln N}, \tag{6.3}$$

then the probability that there will be an unstable bit in an entire set of p memorized patterns will vanish[2].

- If, with the same J_{ij}'s, one allows that the network have stable states not necessarily at the exact memorized patterns, but very near them, then the network can store and retrieve

$$p < p_c \equiv \alpha_c N \tag{6.4}$$

patterns, with $\alpha_c = 0.145$. No more than 1% of the neurons deviate from the memorized patterns [1,16].

- As α becomes greater than α_c no pattern can be retrieved.

- With finite noise (finite T) one has

$$p < \alpha_c(T) N, \tag{6.5}$$

with $\alpha_c(T) < \alpha_c$. The retrieval quality at $\alpha_c(T)$, measured by the overlap, is $m_c(T)$ (the error fraction is $\frac{1}{2}(1 - m_c)$). For $\alpha > \alpha_c(T)$, retrieval is impossible and for $\alpha < \alpha_c(T)$ retrieval is possible with an overlap better than $m_c(T)$[1]. See e.g Figures 6.1 (a) and (b).

- A storage prescription which allows forgetting upon learning[7], rather than the total blackout at α_c implied by the couplings Eq. 6.1, is given by

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \Lambda(\frac{\mu}{N}) \xi_i^\mu \xi_j^\mu. \tag{6.6}$$

with

$$\Lambda(x) = \epsilon \exp\left(-\frac{x\epsilon^2}{2}\right).$$

The most recently acquired pattern is μ=1 and the oldest is μ=p. See also Sections 6.6.1 and 6.7.3. In this network, memorization and storage capacity depend on the value of ϵ. Its optimal value is $\bar{\epsilon} = 4.108$. For this value of ϵ, the storage capacity is α_c=0.049, which implies that the $0.049N$ most recent patterns can be retrieved with less than 1.5% error.

The second class of capacity computations has a very different point of departure. The question asked is: Given a set of p random patterns, can a set of J_{ij}'s be found such that, at $T = 0$, these patterns are stable? This is a search in the $N \times (N-1)$-dimensional space of J_{ij}. Note that symmetry is not imposed. The results are:

- There can be no solution to the stability equations if $\alpha > 2$[8].

- For $\alpha < 2$, there is a finite volume in the space of normalized J_{ij} for which a solution exists. The finite volume ensures some level of robustness[5].

- If one requires a finite basin of attraction (measured by a parameter K) to exist around each pattern, one obtains $\alpha_c(K)$, a decreasing function of K.

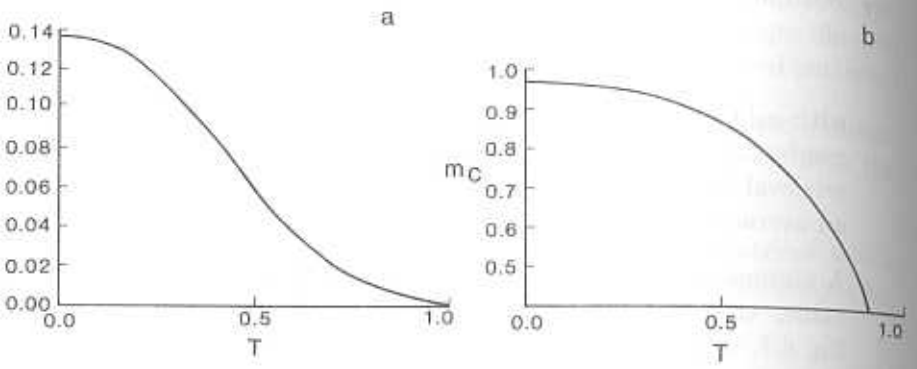


Figure 6.1: The dependence on the noise (temperature) (a) of the fractional storage capacity, α_c, (b) of the overlap – retrieval quality – at maximum storage.

- If one allows for correlations among the memorized pattern, the storage capacity increases, diverging as the patterns become fully correlated.
- The retrieval error $f(\alpha, K)$, which is the minimal fraction of unstable bits per pattern[17], is represented in Figure 6.2.

6.2 Statistical Estimates of Storage

6.2.1 Statistical signal to noise analysis

The first hint that loading the memory of an ANN may affect its functioning in retrieval has appeared in the signal-to-noise arguments of Section 4.3.1, following the original paper of Hopfield[18]. The point there was that if the patterns are chosen at random and the synaptic efficacies have the special form Eq. 6.1, then every pattern contributes to a noise term in the PSP of any other. This is a result of the fact that the patterns are not mutually orthogonal. Yet this noise term, which was described as a simple random walk of steps $+1$ and -1, was estimated to be of relative magnitude $\sqrt{p/N}$ compared to the signal term, which stabilizes the pattern. As long as p was kept finite as N became very large, the noise was negligible and the zero temperature

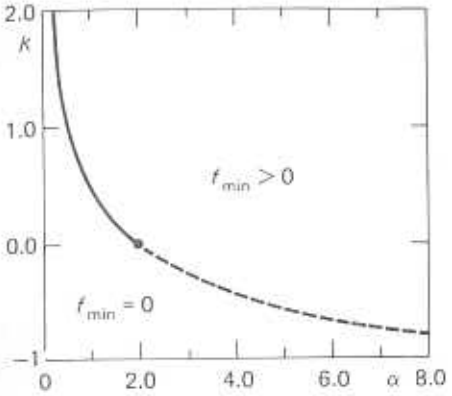


Figure 6.2: The minimal fraction of errors in an optimal network trying to stabilize a set of N random patterns. The dot on the curve is the highest value of α (=2) for which one may have stable points ($K > 0$) with no errors. (From ref. [17], by permission.)

stability of all the memorized patterns was assured. This would lead to the naive estimate that storage is safe as long as $p \ll N$ and that storage capacity $p_c = \alpha_c N$.

The first extension of the finite p estimate of the noise is to assume that there is a range in which p becomes large but the noise is still a sum of pN uncorrelated $+1$'s and -1's, so that for large N it is a random variable with a gaussian distribution[2]. It turns out that the storage capacity resulting from this assumption, that is the maximum p for which the exact patterns are stable at T=0, is consistent with the assumption of a gaussian distribution. But it requires a major effort to prove this statement[19]. Assuming a normal distribution of the noise one finds that for the probability that any bit in a pattern be unstable to be vanishingly small, as $N \to \infty$, one must have

$$p < \frac{N}{2 \ln N}.$$

While if one makes the stronger requirement of vanishingly small probability for an unstable bit in a set of p patterns, than it must be that

$$p < \frac{N}{4 \ln N}.$$

The argument proceeds as follows: With the choice Eq. 6.1 for the synaptic efficacies, the condition that a given firing pattern $\{S_i\}$ be dynamically stable is that at every site the field be in the direction of the spin, i.e.,

$$S_i h_i > 0 \qquad (i = 1, 2, \ldots, N), \tag{6.7}$$

with the local fields h_i given by

$$h_i = \frac{1}{N} \sum_{j,j\neq i} \sum_{\mu} \xi_i^\mu \xi_j^\mu S_j. \tag{6.8}$$

We will start by computing the probability that one of the bits, $i = 1$, of one of the stored memories, say ξ^1, satisfies the corresponding inequality 6.7. We therefore substitute $\{\xi_i^1\}$ for $\{S_i\}$ in Eqs. 6.8 and 6.7 to find:

$$\xi_1^1 h_1 = \frac{N-1}{N} + \frac{1}{N} \sum_{j,j\neq i}^{N} \sum_{\mu=2}^{p} \xi_1^1 \xi_1^\mu \xi_j^\mu \xi_j^1, \tag{6.9}$$

where the sum over μ has been separated into a *signal* term, induced by the pattern $\mu = 1$, and the remaining *noise* term, just as was done in Section 4.3.1.

In a large system (as $N \to \infty$), the signal term tends to unity. The noise term is a sum of $(N-1)(p-1) \approx Np$ random bits of $+1$ and -1 divided by N. Recall that now $p \to \infty$ when N does. By the central limit theorem the noise must tend to a gaussian random variable with zero mean and a mean square deviation

$$\sigma^2 = p/N.$$

The probability for $S_1 = \xi_1^1$ to be stable is equal to the probability that the right hand side of Eq. 6.9 is positive, namely, that the noise term is larger than -1 (see Figure 6.3). This probability is given by:

$$\Pr(\xi_1^1 h_1 > 0) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-1}^{\infty} dx \exp\left(-\frac{x^2}{2\sigma^2}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\sqrt{\frac{1}{2\sigma^2}}\right)\right]. \tag{6.10}$$

The *error function* appearing in the above equation has been defined in Eq. 2.16.

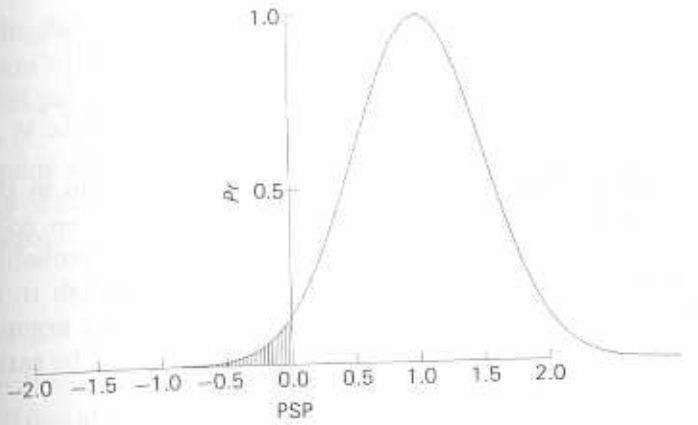


Figure 6.3: Graphical representation of the probability of the stability of a specific bit in one of the stored patterns. The shaded area is the probability that the noise term overrides the signal term.

In order to estimate the loading level, p_c, at which the signal-to-noise argument breaks down and unstable bits begin to appear, even in networks with very large N, we investigate Eq. 6.10 for $\sigma = \sqrt{p/N}$ small. This implies that the argument of the error function is large. The asymptotic behavior of the error function, for large values of its argument, is[20](Eq. 8.254)

$$\operatorname{erf}(x) \approx 1 - \frac{1}{\sqrt{\pi}x} e^{-x^2}. \tag{6.11}$$

Consequently, the probability that neuron no. 1 in pattern no. 1 is stable, Eq. 6.10, can be approximated by

$$\Pr(\xi_1^1 h_i > 0) \approx 1 - \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{1}{2\alpha}\right), \tag{6.12}$$

where we have defined

$$\alpha \equiv \frac{p}{N}.$$

But there are N bits in our pattern and the probability that they are all stable is the product of N such terms. Thus, the probability that the entire pattern no. 1 is stable is:

$$\Pr(\text{stable pattern}) \approx \left[1 - \sqrt{\alpha/2\pi}\exp(-1/2\alpha)\right]^N \quad (6.13)$$

$$\approx 1 - N\sqrt{\alpha/2\pi}\exp(-1/2\alpha). \quad (6.14)$$

and it must be close to unity. Hence, the second term in the above equation must be negligible compared to unity as $N \to \infty$. Put in other words, since the second term in Eq. 6.12 is the probability that the bit will be unstable, one finds that it is not enough to have the probability for one unstable bit be small, as the naive argument had it, but it must be small compared with $1/N$. This will be satisfied if

$$\alpha = \frac{1}{2\ln N}. \quad (6.15)$$

In fact, if one writes

$$\alpha = (2x\ln N)^{-1} \quad (6.16)$$

one finds, for this second term in Eq. 6.14 (to be denoted by N_{er} for reasons which will become clear below), then

$$N_{er} \approx \frac{N^{1-x}}{\ln N}. \quad (6.17)$$

For this term to vanish as N tends to infinity we must have $x \geq 1$.

Looking at the argument from the perspective of the unstable bits, one observes that the mean number of errors in a pattern is the probability for a bit to be unstable times the total number of bits N. But this is just the second term in Eq. 6.14. Hence, the mean number of errors per pattern, N_{er} is

$$N_{er} \approx N\sqrt{\frac{\alpha}{2\pi}}\exp\left(-\frac{1}{2\alpha}\right), \quad (6.18)$$

which should explain the notation in Eq. 6.17. If the number of errors per neuron in the pattern has to vanish, then we are led again to Eq. 6.15.

The maximal number of patterns that can be stored under those conditions is

$$p_c = \frac{N}{2\ln N}. \quad (6.19)$$

This loading is high enough to ensure that the assumptions about the distribution of the noise term hold[19]. If one demands further that all p stored patterns be reproduced without error, then pN_{er}/N has to vanish as $N \to \infty$. Since p is itself of order N, the requirement that pN bits be stable implies an even narrower probability distribution for the instability of an individual neuron. This is tantamount to an even lower storage capacity, since now the condition on x of Eq. 6.16 is $x \geq 2$, and therefore

$$p_c = \frac{N}{4\ln N}. \quad (6.20)$$

These severe limitations on the number of stored patterns can be removed if one allows a small finite *fraction* of misaligned bits. Actually one should not be concerned with the exact stability of the memorized patterns. For purposes of storage and retrieval it is sufficient that there are stationary points of the dynamical process which are near enough to the patterns so that the biological readout mechanism would draw the correct conclusions. This will be the subject of Section 6.3.2. Moreover, the demand of stationarity can also be relaxed. The network would work effectivly provided the dynamics draws the network to the neighborhood of a memory, even if a small fraction of neurons keep changing their states. If the time averaged overlap is high enough, the memory will be properly recalled.

6.2.2 Absolute informational bounds on storage capacity

It is the expressed view of this book that ANN's do not create information. This has already been stated in Section 4.5.4 and it has dictated our attitude toward spurious states – a nuisance to be disposed of in order to facilitate the retrieval of the patterns intentionally stored in the network. The function of the network is to retrieve, via its dynamics, as best it can the memorized patterns. Given this attitude, it is clear that the best the network could possibly do is to invert its synaptic connections and to reproduce the patterns that served to form them. But the total number of bits of **random** patterns that can be thus reconstituted cannot exceed the total number of bits that are in the synapses. After all, one would need as many bits as there are in the patterns, merely for the specification of the arrangement of the Np random bits in the p patterns.

Suppose that each synapse can store D bits. Then the synaptic matrix of a large, fully connected symmetric network can store no more than $\frac{1}{2}DN^2$ bits. A set of p random patterns contains pN bits and hence the absolute limit on p is[21]

$$p < \tfrac{1}{2}DN.$$

This bound would imply, for example, that if a synapse can hold a single bit, i.e., it can be either $+1$ or -1, then no more than $N/2$ patterns can be retrieved. If the constraint of symmetry is lifted, this number increases by a factor of 2 and $p_c < N$.

It is quite interesting that this constraint, which has not yet found a proper formalization[2], and is usually considered to be a very weak constraint, has recently found a very significant application. In estimating p_c by scanning the space of coupling constants, described in the second part of Section 6.1.4, one finds[17] that for a distribution of one-bit synapses $p_c = (4/\pi)N$. This is about 30% higher than the absolute bound discussed above. It has turned out that this excess is due to an approximation (*replica symmetry*) which is usually extremely accurate, but fails in this special case. In fact, that same approximation is excellent when one repeats the computation, provided the synapses are not discretized to a single bit[17] each. Moreover, recent numerical studies[22] indicate that the optimal storage obtained with binary synapses is around p_c=0.8N.

For a model with standard synapses, Eq. 6.1, which stores p patterns the synapses must be able to contain at least $\log_2(2p+1)$ bits, as the values stored in them range from $-p$ to p. The absolute constraint will therefore be:

$$Np \leq \tfrac{1}{2}N^2 \log_2(2p+1). \tag{6.21}$$

This implies[21] that, for large N,

$$p_c \leq \tfrac{1}{2}N \log_2 N,$$

which is indeed a weak upper limit on the actual storage capacity which is a fraction of N. On the other hand, it does predict the correct N-dependence of the storage capacity in a neural network with

[2]Perhaps because it so fundamental.

multi-neural (multi-spin) synapses. If the network is fully connected and each synapse connects x neurons, according to,

$$J_{i_1,\ldots,i_x} = \frac{1}{N^{x-1}} \sum_{\mu=1}^{p} \xi_{i_1}^{\mu} \cdots \xi_{i_x}^{\mu}, \tag{6.22}$$

then there would be $N^x/x!$ synapses. Each synapse will have $\log_2(2p+1)$ bits, as before. The absolute constraint, corresponding to Eq. 6.21 will read:

$$Np \leq \frac{1}{x!} N^x \times \log_2(2p+1) \tag{6.23}$$

from which one derives[21]

$$p_c \leq \frac{1}{x!} N^{x-1} \times \log_2 N.$$

Apart from the logarithmic term, this gives the correct power dependence of p_c on N[23,24,8,25].

6.2.3 Coupling (synaptic efficacies) for optimal storage

A quite different approach to the question of storage capacity is to formulate it backwards. Rather than ask about how many patterns can be stored and retrieved with a given prescription for the J_{ij}'s, one leaves the couplings free. The question becomes one about the availability of any set of J_{ij}'s which will ensure the stability conditions for the given set of patterns. Since the J's are continuous variables, this is translated into a question about the volume in the space of J_{ij}'s for which the stability conditions are satisfied[8,26]. This type of computation is based on a few presuppositions:

- The memorized patterns are random and *quenched*. (see e.g., Section 4.1.2).
- Only solutions with a finite volume in coupling space are of interest, since other solutions are neither generic nor robust. This volume decreases as p increases and it vanishes when the storage capacity of the network has been reached.
- As usual in statistical physics, we look for solutions in the typical case (not the worst case, for example).

- Solutions of the stability conditions for the patterns should produce reasonable basins of attraction.

What one is looking for is the volume in J-space in which the set of inequalities

$$R_i^\mu = \xi_i^\mu h_i\{\xi^\mu\} > K \tag{6.24}$$

is satisfied, with

$$h_i\{S\} = \frac{1}{\sqrt{N}} \sum_{j,j\neq i} J_{ij} S_j \tag{6.25}$$

for all $\mu = 1, \ldots, p$. The positive constant K is a measure of the basin of attraction of each of the patterns[27,28].

The overall scale of the couplings has to be chosen so that PSP's be finite as $N \to \infty$. This is ensured by the prefactor in Eq. 6.25, which keeps the values of the h_i's of order unity. Moreover, if a certain set of J_{ij}'s is a solution of the inequalities 6.24, then any positive multiple of this set of couplings will solve them. The J_{ij}'s must therefore be constrained, to obtain a finite volume in J-space when the volume is non-zero. This is done by imposing, for every i,

$$\sum_{j,j\neq i} J_{ij}^2 = N. \tag{6.26}$$

The most relevant (order) parameter for this problem is the similarity of different possible solutions for the stability equations. If the volume of the solutions shrinks, the available solutions become increasingly similar. If it expands, the variety increases and the overlap between two **typical** solutions is very small. It is therefore a natural choice to measure this parameter by

$$q = \overline{\frac{1}{N} \sum_j J_{ij}^\alpha J_{ij}^\beta} \tag{6.27}$$

where the bar indicates an average and the superscripts on the J's denote particular solutions in the volume. One expects that as $q \to 0$ the volume is large, and that it shrinks to zero as $q \to 1$, when all solutions become essentially identical. This will be identified as the signal for the saturation of the network. For a larger number of random patterns there is no solution.

These intuitions are substantiated by detailed, forceful analysis[26]. Here, only the flavor of the technical aspects will be conveyed. A very similar technique is described in Appendix 6.7.1, in the context of the thermodynamic calculation of the storage capacity. The mean fractional volume is computed as the ratio between the volume in J-space in which the inequalities 6.24 hold to the total volume of J's obeying the constraint 6.26. This ratio can be written down for a given choice of the memorized patterns. To obtain the typical value of this fractional volume one must average over the *quenched* randomness of the stabilized patterns and hope that the average represents well the typical situation.

The average is performed not on the volume itself, which is of order $\exp N$, but on the logarithm of the volume, which is of order N. As is common practice in dealing with quenched randomness, see e.g., the discussion in Section 6.7.1. Again, as is typical in such cases, the final result is obtained by a computation of a complicated integral of an integrand which behaves like $\exp N[g(q)]$,[3] by the method of Laplace, as in Section 3.4.2. It consists of finding the maxima of $g(q)$, and replacing the integral by the value of the integrand at this maximum. In this maximization procedure one finds that the maximum of g is at a value of q which depends on $\alpha(= p/N)$ as well as on the basin parameter K, which allows for an evaluation of $\alpha_c(K)$.

To cut a long story short, one arrives at the following equation for q, at which g is maximal[26]:

$$q = (1-q)\frac{\alpha}{2\pi} \int_{-\infty}^{\infty} \frac{dt}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) \exp(-Z^2) \left(\frac{1}{2}[1 - \mathrm{erf}(Z/\sqrt{2})]\right)^{-2}$$

with the error function defined in Eq. 2.16, and

$$Z \equiv \frac{q^{1/2}t + K}{\sqrt{1-q}}.$$

As was mentioned above, saturation is reached if the value of q approaches one. When this happens, Z tends to infinity, and using the asymptotic expansion of the error function, Eq. 6.11, one arrives at the simple equation:

$$\frac{1}{\alpha_c} = \int_{-K}^{\infty} \frac{dt}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})(t+K)^2.$$

[3] After assuming that replica symmetry is intact.

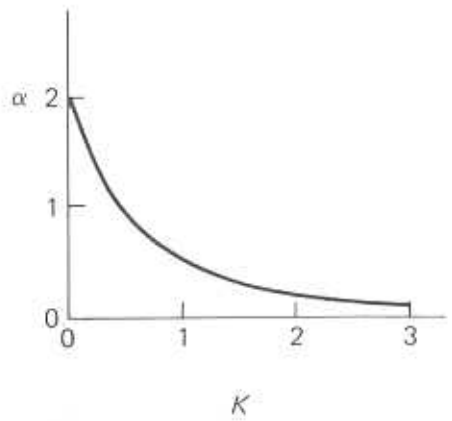


Figure 6.4: The storage capacity as function of the basin parameter for the optimally chosen synaptic efficacies. (After ref.[26]).

From this equation it follows immediately that when K=0, α_c=2. As K increases α_c decreases as is depicted in Figure 6.4.

We follow this brief account of the storage calculation by a few brief comments:

- The main approximation has been that of *replica symmetry*. This symmetry is, technically speaking, stable, which implies that it is self-consistent.
- There is a version of the perceptron learning algorithm, to be discussed in Section 9.2, which ensures that when a solution matrix J exists, stabilizing a given set of patterns, the algorithm will arrive at some solution matrix J **in a finite number of steps**. The above discussion implies that it is actually possible to teach a network as many as $\alpha_c(K)N$ patterns.
- There is an extension of this computation which relaxes the constraint of perfect stability of the stored patterns, allowing a certain fraction of errors f[17]. One then calculates the lowest possible number of errors, f, for a given α and K, which is presented in Figure 6.2.

Anticipating the next section, where we compute thermodynamically the storage capacity of a network with standard couplings, we

emphasize the significance of the different measures of storage capacity. The standard network has $\alpha_c \approx 0.14$. On the face of it, the optimized network has a storage capacity 14 times higher. If one is concerned with storage efficiency per synaptic bit, one will have to obtain an estimate of the range of synaptic values required for the optimization of the J_{ij}'s.

6.3 Theory Near Memory Saturation

6.3.1 Mean-field equations with replica symmetry

We now return to the standard model to extend the considerations of Chapter 4 to the case where the number of memorized patterns can be of the same order as the number of neurons. The main difference is that now the small random overlaps of a pattern being retrieved with the rest of the stored patterns, each of order $1/\sqrt{N}$, add up to have a finite effect on the stability of the attractor. These overlaps act like a second source of noise, *slow noise*, in addition to the *fast noise* associated with the temperature. Their effect on the structure of the landscape, which determines the flows in the space of network states, is quite pronounced[1].

- They deform completely the distribution of spurious state, merging them into a spin-glass state, which is largely shapeless.
- They shift the position of the stable states of the dynamics away from the memorized patterns slightly.
- As the number increases further, the overlaps abruptly destroy the retrieval states – attractors in the neighborhood of the stored patterns – and all attractors become essentially uncorrelated with the stored patterns. This is sometimes referred to as a *blackout*.

First, we must establish the set of relevant parameters for the description of the retrieval dynamics. In the presence of fast noise there will, of course, be no fixed network states but only attractive distributions, breaking the system's ergodicity. It is the averages parametrizing these distributions which we are after, as is explained in Section 3.3.4. An obvious set of *order-parameters* is the same as that introduced in

Section 4.4. These are the mean overlaps of the states of the network visited by the dynamics and the memorized patterns. They are:

$$m^\mu = \left\langle\left\langle \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \langle S_i \rangle \right\rangle\right\rangle, \tag{6.28}$$

where $\langle ... \rangle$ is an average over that part of the space of network states which the dynamics allows. It can be conceived either as a time average or as an ensemble average. The latter is just the standard thermal average of statistical mechanics. Retrieval, we recall, is identified by a large time-averaged overlap with consecutive single states, which is essentially the average overlap defined above. But, one has to keep in mind that in order to obtain a typical result, rather than a consequence of some particular choice of the random patterns, we must perform the (quenched) average over the distribution of the patterns. This is the significance of the double angular brackets.

This set of order-parameters will not suffice near saturation. As the number of patterns increases, so does the randomness of the synapses. This raises the possibility of a *spin-glass* order.

Brief digression on spin-glass ordering

In the ferromagnetic system we have identified two basic types of dynamical development. One was ergodic. The other, at low temperature, was constrained to a small part of the space of states. The size of that part shrinks to zero as the temperature vanishes. In that case the system has a large magnetization. Spins at distant parts of the system tend to point in the same direction. Thus when ergodicity breaks, two things occur simultaneously:

- A dynamical (temporal) freezing of the individual spins.
- Long range correlations between the orientations of the spins.

Since the two go hand in hand, the *magnetization* is a sufficient parameter to signal both. When the magnetization becomes macroscopic, of order N (of order one per spin), the system freezes dynamically.

As the mutual influence among the spins becomes random, with mean zero, a new phenomenon takes place[29,30].

At low temperatures, the dynamics of the spins slows to a halt. Yet the contradictory (*frustrating*) interactions prevent long range correlations between the orientations of the different spins. Instead freezing takes place with the spins oriented at random with respect to each other. The magnetization upon freezing is therefore zero, just like in the paramagnetic state. The magnetization cannot be the order-parameter to describe this state of affairs. Rather, it has been proposed[29] that an additional order-parameter should be introduced, whose average can discriminate between spin-glass freezing and paramagnetism. Specifically, it is

$$q \equiv \left\langle\left\langle \frac{1}{N}\sum_{i=1}^{N} \langle S_i \rangle^2 \right\rangle\right\rangle. \tag{6.29}$$

As usual, the angular brackets have the dual meaning – an ensemble average and a temporal average. The double brackets express an average over the (*quenched*) randomness in the couplings. In a paramagnetic phase, the spins keep flipping freely and $\langle S_i \rangle$ would vanish, for every i. Consequently, q will vanish. If the system is in a pure ferromagnetic phase, then q will equal the average magnetization per spin squared. However, if the system enters a spin-glass phase, with dynamical freezing and random orientations, then q will not vanish, and will serve as a measure of the extent of the freezing.

The spin-glass counterpart of the thermal jumping around in the paramagnetic phase is a great abundance of stable states. If the system is in a pure spin-glass phase – one in which there is no ferromagnetic ordering – then random frozen configurations appear all over the space of states. It has been shown[31] that the number of such random, frozen states is exponentially large. More specifically, in a spin-glass of N spins the number of such metastable states N_s behaves as

$$N_s \approx \exp(0.1992N) \tag{6.30}$$

which diverges strongly as N becomes large. The system will, typically, flow into one of these attractors, but the

choice will depend sensitively on the initial state and on the particular realization of the randomness in the couplings.

For a system which is both ferromagnetic and glassy, which is very relevant for the discussion of retrieval at high storage levels, the order-parameter q will differ from m^2 – the square of the magnetization – and will usually be higher. This is a consequence of the fact that most of the spins can organize in a state with long-range correlations, but a fraction can freeze in random orientations, which are uncorrelated with the magnetization. The corresponding thermal situation is the presence of fast noise in the ferromagnetic phase, which allows a fraction of the spins to keep flipping while the majority of the spins is coherently organized. Around every state with a finite magnetization there will be a large number of metastable states in which a small fraction of spins are randomly oriented, even at $T = 0$. Nonergodicity takes place in a double sense. The initial state determines into which of the globally magnetized states the system will flow as well as the particular spin-glass stable state in its vicinity.

In describing the ANN at high storage levels, we will also resort to the order-parameter q of Eq. 6.29, to anticipate possible freezing into disordered stable states. The external double brackets stand, as above, for the averaging over the (quenched) random patterns, which underlie the randomness in the synaptic couplings. In fact, we shall find[1], that coexistence between the ferromagnetic (retrieval) and the spin-glass state is possible. In the ANN ferromagnetism is tantamount to large, macroscopic, retrieval overlaps. The fact that these overlaps are less than unity may be the result of fast fluctuations of some of the neurons under the influence of temperature. Alternatively, some of the neurons (or spins) may freeze in random individual states under the influence of conflicting synaptic inputs.

Finally we need an auxiliary variable, to describe the noise due to the *uncondensed* patterns. In other words, if the network is in a state with large (macroscopic) overlaps with a few of the memorized patterns, the accumulation of the random overlaps with all the other patterns creates a significant amount of noise. If s is the number of

condensed patterns (with finite overlaps as $N \to \infty$), it is described by:

$$r \equiv \left\langle\left\langle \frac{N}{p} \sum_{\mu=s+1}^{p=\alpha N} \langle m^\mu \rangle^2 \right\rangle\right\rangle = \frac{1}{\alpha} \sum_{\mu>s} \left\langle\left\langle \left[\frac{1}{N} \sum_i \xi_i^\mu \langle S_i \rangle \right]^2 \right\rangle\right\rangle . \quad (6.31)$$

Note the special normalization of this parameter. Since each of the m^μ's for $\mu > s$ is of order $1/\sqrt{N}$, the sum is of order p/N and the coefficient makes it of order unity, even if p increases linearly with N.

With these three sets of parameters, the equations for the attractors of the network can be written in the limit when N becomes very large. These are the mean-field equations, which for a fully connected network are exact[1].[4] The derivation of these equations is left to Appendix 6.7.1. Here we will only discuss their structure and their consequences. The equations read:

$$m^\nu = \langle\langle \xi^\nu \tanh \beta[\sqrt{\alpha r}\, z + (\mathbf{m}+\mathbf{h})\cdot\xi] \rangle\rangle \quad (6.32)$$

$$q = \langle\langle \tanh^2 \beta[\sqrt{\alpha r}\, z + (\mathbf{m}+\mathbf{h})\cdot\xi] \rangle\rangle \quad (6.33)$$

$$r = \frac{q}{(1-\beta+\beta q)^2}. \quad (6.34)$$

The vectors $\mathbf{m}, \mathbf{h}$ and ξ have s components each, corresponding to the patterns that have condensed. The fields h^μ are externally imposed local fields (PSP's). The field h_i^ν produces at each neuron a PSP proportional to ξ_i^ν. The double angular brackets, the average over the random patterns, is composed of two parts. The first is an average over the distribution of discrete ξ^ν's, just like in the case of finite p. The second is an average over a continuous gaussian distribution of the noise generated by the 'other' patterns. The combined average can be written as:

$$\langle\langle O \rangle\rangle \equiv 2^{-s} \sum_{\mu=1}^{s} \sum_{\xi^\mu=\pm 1} \int_{-\infty}^{\infty} dz \frac{\exp(-z^2/2)}{\sqrt{2\pi}} O\{\xi\}. \quad (6.35)$$

In Appendix 6.7.1 it is shown that the three equations for m^μ, q and r are the equations for the extrema of a free-energy in the space

[4] Except for small effects of replica symmetry breaking.

of these variables, which extends the landscape picture to the present situation. This free-energy reads:

$$f = \frac{\alpha}{2} + \frac{1}{2}\sum_{\nu=1}^{s}(m^\nu)^2 + \frac{\alpha}{2\beta}\left(\ln(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}\right)$$
$$+\frac{1}{2}\alpha\beta r(1-q) - \frac{1}{\beta}\langle\langle \ln 2\cosh\beta\left[\sqrt{\alpha r}\, z + (\mathbf{m}+\mathbf{h})\cdot\xi\right]\rangle\rangle. \tag{6.36}$$

What these equations bring forth is the fact that when the number of stored patterns becomes extensive (of order N) the local field at each neuron is composed of two parts:

1. A ferromagnetic part, $\mathbf{m}\cdot\xi$, which is the signal induced by the s condensed patterns.

2. A spin-glass noise part, $\sqrt{\alpha r}\, z$, generated by the random overlaps.

This noise, in the present (*replica symmetric*) context is still gaussian, but of a rather special kind. Its width is no longer simply the root-mean square of pN independent, random bits ± 1's, as was the case in Section 6.2.1. Instead it is connected via Eq. 6.34 to the spin-glass order-parameter q. As we shall see below, this parameter does not simplify to the expression of Section 6.2.1 even when $T \to 0$, unless p/N is small enough. This is a consequence of the fact that when the randomness is large enough the situation at the observed site reacts back on the other sites, because of the requirement that the system be at the bottom of a valley, and this *reaction* contributes back to the noise.

We will not elaborate here on this point, which is an inalienable part of spin-glass culture[32,33], except to reemphasize that while the distribution of the noise remains gaussian it is not a consequence of the independence of the terms composing it, even when $N \to \infty$. When the problem is treated more carefully, the deviations become so large that the distribution is no longer gaussian.

6.3.2 Retrieval in the absence of fast noise

The properties of the attractors of the network can be derived by analyzing the solutions of Eqs. 6.32–6.34, as a function of the memory

loading (α), the fast noise (T), and the external PSP's (h^ν). To begin with, we set the external fields to zero and take the limit $T \to 0$. The solutions become functions of the single parameter α.

The analysis of the equations, to be sketched in the next section, brings out the following picture. There are two main regimes:

1. $\alpha < 0.138$ – magnetic regime – effective retrieval.

2. $\alpha > 0.138$ – the spin-glass regime – no retrieval.

Retrieval regime

Retrieval is restricted to attractors with a single macroscopic overlap with the memorized patterns. When they appear, all $2p$ of them appear together, just as in the finite-p case. Our system has full symmetry between the different memories as well as between these memories and their complementary states, in which the state of each neuron is reversed. Later on more discriminatory networks will be discussed. See e.g., Sections 8.2 and 6.6.1.

In the retrieval regime the network behaves very much as if it were at a finite temperature, with $T = \sqrt{2\alpha r}$. This effective temperature is due to the randomness in the synapses. It implies, by analogy with the finite-p case (see e.g., Section 4.1.4) that the overlaps in retrieval will not be unity. Retrieval will contain errors. The number of errors in retrieval decreases with α, as it would with decreasing temperature at finite p. One major difference from the noisy finite-p case is that here retrieval is represented by a real attractor, rather than by small fluctuations around a large mean. This implies that errors are not a result of dynamical fluctuations about a stored pattern, but are due to the fact that the fixed attractors are not at the memories. They are slightly displaced. Another important contrast is that the quality of retrieval does not erode continuously to zero, as it did when $T \to 1$, in Section 4.1.4. Instead, the overlap of the attractors with the memorized patterns decreases from unity, for vanishingly small α, to 0.97 as α approaches the value 0.138. Then the overlap drops abruptly to zero.

The level of errors in retrieval is represented in Figure 6.5. It should be appreciated that before the total collapse of retrieval at a storage level of $0.138N$ patterns the fractional error is never higher than 1.5%.

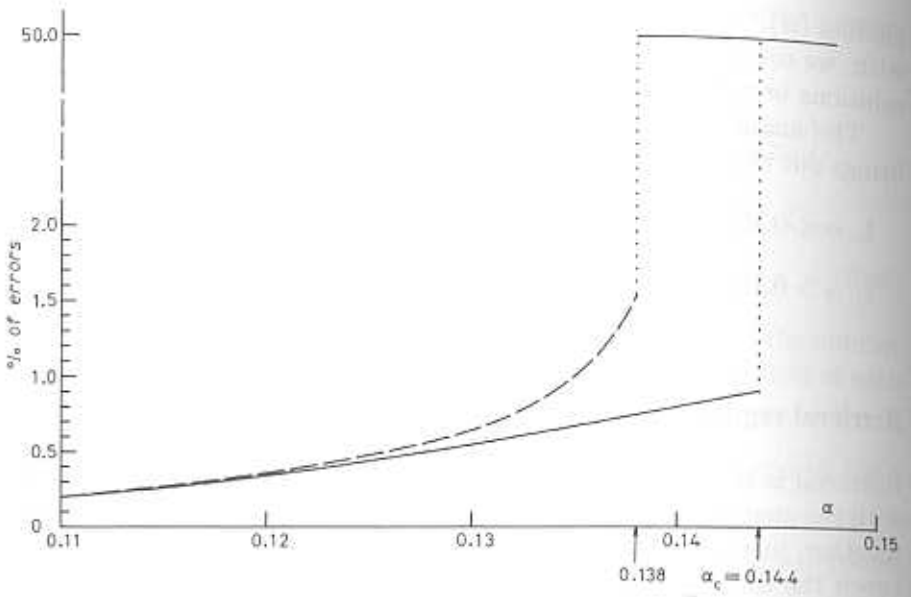


Figure 6.5: The fraction of errors in retrieval vs memory loading. At α=0.138 (dashed curve) there is a jump from 1.5% to 50% errors, which is complete lack of correlation. Full curve includes improvement due to replica symmetry breaking[16].

This follows from the relation between the overlap and the Hamming distance, Eq. 1.9, which implies that the error fraction is

$$\frac{N_{err}}{N} = \frac{1}{2}(1 - m)$$

where m is the overlap.

The fraction of errors in retrieval decreases exponentially as $\alpha \to 0$. In the next section it is shown that the fraction of errors behaves as:

$$\frac{N_{err}}{N} \approx \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{1}{2\alpha}\right). \tag{6.37}$$

which is the asymptotic form of the function represented in Figure 6.5.

The small deviations of the attractors from the memorized patterns is the price paid for the expansion in storage from $\alpha_c \approx 1/(2 \ln N)$, of Section 6.2.1 to $\alpha_c \approx 0.138$. The two results must, of course, merge for very small α. This can be verified on two levels. First, the noise

distribution obtained in the thermodynamic calculation becomes normal, i.e., its width tends to $\sqrt{\alpha}$, as is the case in Eq. 6.10. In fact, it is shown in the technical digression below that as $\alpha \to 0$, in a situation of retrieval, $r \to 1$ and the width, $\sqrt{\alpha r}$, is just $\sqrt{\alpha}$. Secondly, Eq. 6.37 can be used to derive the behavior of α which will ensure a vanishing number of errors per pattern, or per set of p patterns. Clearly, as long as α remains finite, the errors remain a finite fraction of the number of neurons in the network, Eq. 6.37. For the total number of errors not to increase with N as $N \to \infty$, α must vanish in that limit. Setting

$$\alpha = \frac{1}{2x \ln N}$$

one finds

$$N_{err} \approx N^{(1-x)}.$$

Hence, we must have $x \geq 1$, as in Eq. 6.19. Moreover, if the total number of errors in p patterns is to remain finite for large N, one arrives at the result $\alpha_c = 1/(4 \ln N)$, Eq. 6.20.

At $T = 0$, for $\alpha < \alpha_c$ one finds that $q = 1$, while $m \neq 1$, because of the randomness in the connections, as was explained in the digression above. This is a clear indication that there is dynamical freezing which is not associated with the alignment of retrieval. This additional freezing is random and is of the spin-glass type.

In the retrieval regime, one finds the following structure for the attractors:

- Just below α=0.138 there are two types of attractors, those associated closely with one of the memorized patterns (having a single m^μ near unity), and a spin-glass states (with all m's zero). If α is in the interval (0.051–0.138), the spin-glass state is lower in the energy landscape than the retrieval state.

This should be contrasted with the situation at finite p and finite temperature. There, as we approached the disappearance of retrieval (at T=1), the retrieval states were absolute minima to the bitter end. The contrast is related to the fact that the abrupt disappearance of the retrieval state at α_c is a physicist's thermodynamic transition of first order, while the finite-p transition at T_c is of second order. We return to this point below.

- At $\alpha = 0.051$ the $2p$ retrieval states become absolute minima of the landscape.

- As α decreases further spurious mixture attractors appear.

First to materialize are the symmetric mixtures of three memorized patterns, much like in the finite-p situation upon the reduction of temperature. This takes place at $\alpha \approx 0.03$, where the amplitude of the three equal overlaps is 0.496, compared with $m = 0.5$ for the three-mixture of finite-p at $T = 0$. We will not elaborate further on the rich structure of spurious states as part of our general attitude that they are truly spurious.

Spin-glass phase

As far as associative memory is concerned, the main interest of the spin-glass phase is the absence of retrieval. When the network is in a spin-glass attractor, the overlaps with all memorized patterns vanish. This is strictly true only in the framework of the approximations made to derive the mean-field equations, namely $N = \infty$ and replica symmetry. The first approximation is under our control. Replica symmetry is very well satisfied in the retrieval states. The most dramatic correction is that the value of α_c increases from 0.138 to 0.145, when the breaking of replica symmetry is taken into account[16]. This is the full curve in Figure 6.5. Hardly an issue to write home about.

The situation is not as well described in the spin-glass state, where deviations from replica symmetry are expected to be significant. For example, one finds in extensive simulations, which will be described below, that the attractors in the spin-glass phase do not really have zero overlaps with the patterns. These overlaps appear to be reaching a finite limit as N increases. Yet, those overlaps are always below 0.4, an impressive jump from 0.97 in the worst retrieval situation. This may be sufficient for detecting retrieval.

In the spin-glass state as $T \to 0$, $q \to 1$ as well, while all $m^\mu = 0$. This is a typical spin-glass phenomenon (see e.g., digression on spin-glasses). It is quite interesting to observe that this very complicated spin-glass state can be visualized as the merging of very many finite-p spurious attractors as $p \to \infty$. To see this recall that in Section 4.5.2 it has been shown that if n, the number of memories symmetrically admixed into the spurious state, is large, then each overlap tends to zero as

$$m_n \approx \sqrt{\frac{2}{n\pi}}.$$

Moreover, the value of q in each of these states is unity if the number of admixed patterns is odd and is given by

$$q = 1 - \Pr(z_n = 0)$$

if n is even, where z_n is the the sum of n randomly chosen ± 1's. The probability that $z_n = 0$ tends to zero as $1/\sqrt{n}$. Hence, for both odd and even n, $q \to 1$ as $n \to \infty$. In other words, if p is large, there are very many symmetric spurious states each with $n \approx p$, with vanishing overlaps and with $q = 1$. This gradual construction of the spin-glass state can be further corroborated by a demonstration that the values of the noise parameter r, as well as of the energy, in this state are continuous limits of the corresponding quantities in the spurious states at finite p, as $p \to \infty$.

6.3.3 Analysis of the $T = 0$ equations

The technical details involved in the analysis of the solutions of the mean-field equations are relatively simple, so rather than relegate them to an appendix we discuss them in this subsection.

First we deal with the retrieval states. These are the solutions with a single finite overlap m. To study this regime analytically, one observes the fact that in the limit $T \to 0$, or $\beta \to \infty$

$$\begin{aligned}\int_{-\infty}^{\infty} dz \frac{\exp(-\frac{1}{2})z^2}{\sqrt{2\pi}} \tanh \beta(Az + x) &\to \sqrt{\frac{2}{\pi}} \int_0^{x/A} dz \exp\left(-\frac{1}{2}z^2\right) \\ &= \operatorname{erf}(x/\sqrt{2}A). \end{aligned} \tag{6.38}$$

When this is applied to Eq. 6.32, with $\mathbf{h} = 0$, one arrives at:

$$\mathbf{m} = \left\langle\left\langle \xi \operatorname{erf}\left(\frac{\mathbf{m}\cdot\xi}{\sqrt{2\alpha r}}\right)\right\rangle\right\rangle, \tag{6.39}$$

where here the double brackets represent an average over the s condensed discrete ξ's.

This zero temperature equation should be compared with the corresponding equation of the finite-p network, Eq. 4.40,

$$\mathbf{m} = \left\langle\left\langle \xi \tanh \frac{\mathbf{m}\cdot\xi}{T}\right\rangle\right\rangle.$$

The error function in Eq. 6.39 is rather similar in behavior to the hyperbolic tangent in the last equation. If r remains finite for finite m, the zero temperature finite-α overlaps will behave very much like the finite temperature finite-p ones. Moreover, as $\alpha \to 0$, Eq. 6.39 reduces to

$$\mathbf{m} = \langle\langle \xi \, \mathrm{sign}(\mathbf{m} \cdot \xi) \rangle\rangle,$$

provided $\alpha r \to 0$ in this limit. But this is just the zero noise limit of Eq. 4.40.

Now we turn to Eq. 6.33. The right hand side has the limit unity as $T \to 0$. But, as $\beta \to \infty$, the appearance of the term $\beta(1-q)$ in the denominator in Eq. 6.34 requires the computation of the $T = 0$ limit of this expression, which involves the term of $0(T)$ in q. This is found to be:

$$C \equiv \lim_{T \to 0} \beta(1-q) \sqrt{\frac{2}{\pi \alpha r}} \left\langle \left\langle \exp - \left[\frac{(\mathbf{m} \cdot \xi)^2}{2\alpha r} \right] \right\rangle \right\rangle . \tag{6.40}$$

The two limiting expressions Eqs. 6.39 and 6.40 lead to the basic equation determining both *storage capacity* and *retrieval quality*.

To see this, we take $\mathbf{m}$ to have a single non-vanishing component m. In this case, Eq. 6.39 reduces to

$$m = \mathrm{erf}\left(\frac{m}{\sqrt{2\alpha r}} \right) \tag{6.41}$$

and Eq. 6.33 reads

$$r = (1 - C)^{-2} \tag{6.42}$$

with

$$C = \sqrt{\frac{2}{\pi \alpha r}} \exp\left(-\frac{m^2}{2\alpha r} \right) . \tag{6.43}$$

These equations reduce to a single equation for the variable $y \equiv m/\sqrt{2\alpha r}$, namely

$$y = \frac{\mathrm{erf}(y)}{\sqrt{2\alpha} + (2/\sqrt{\pi}) \exp(-y^2)} . \tag{6.44}$$

This is a relation between the retrieval quality m and the storage level α. Storage capacity is simply the value of α above which the equation has no solution, except $y = 0$. The graphical solution of 6.44 is shown in Figure 6.6. The straight line is the left hand side. The

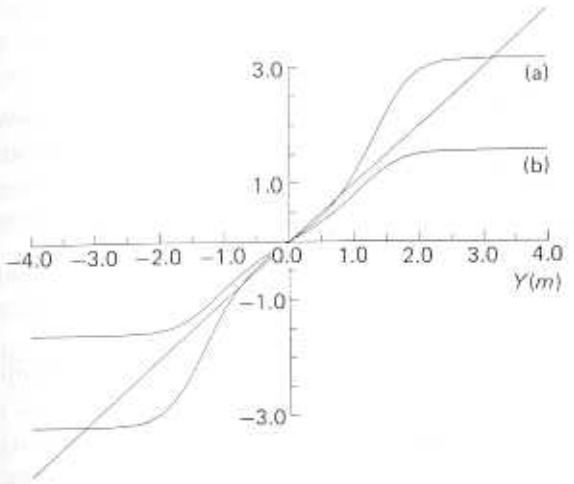


Figure 6.6: Graphical representation of the solutions of the retrieval equation. (a)α=0.05; (b)α=0.15.

curve is the right hand side, plotted for two values of α, one below and one above $\alpha_c = 0.138$. For low values of α there are three intersections. The two extreme solutions, the spin-glass at $m = 0$ and the retrieval state with high m, are dynamically stable, in that they are minima of the energy. Beyond α_c, only the $m = 0$ intersection persists. The figure should make it intuitively clear that the disappearance of the retrieval solution takes place abruptly, a result that is obtained by a numerical solution of Eq. 6.44[1].

Next, inspecting Eqs. 6.43 and 6.42 one concludes that, as $\alpha \to 0$, $m \to 1$ and $\alpha r \to 0$. Consequently, in this limit Eq. 6.41, or Eq. 6.44, lead directly to

$$m \approx \mathrm{erf}\left(\frac{1}{\sqrt{2\alpha r}} \right) .$$

Recalling that the fraction of errors in retrieval is expressed in terms of the overlap as $\frac{1}{2}(1-m)$ and that for large argument the error function has the asymptotic behavior Eq. 6.11, one arrives directly at the asymptotic behavior of the fraction of errors for small α, Eq.6.18.

The energy of a solution can be obtained either as the zero temperature limit of the free-energy, Eq. 6.36 or alternatively, directly from the basic equation:

$$E = \left\langle\left\langle -\frac{1}{2N^2}\sum_{mu=1}^{P}\sum_{ij}\xi_i^\mu S_i \xi_j^\mu S_j \right\rangle\right\rangle + \frac{\alpha}{2}, \tag{6.45}$$

where the last term is the self-interaction which has been included in the first term. Separating the single retrieved pattern and expressing the rest in terms of the definition of the noise parameter r, Eq. 6.31, one finds, in the replica symmetric solution, that

$$E = -\frac{m^2}{2} + \frac{\alpha}{2}(1-r). \tag{6.46}$$

This is the energy at the solutions, but not away from them. One cannot, for example, derive the mean-field equations as derivatives of this expression for the energy, because the zero temperature limit cannot be exchanged with those variations. It can, however, be used for comparing the energies of different solutions.

One finds that for very small α, $r \to 1$ and $m \to 1$. Hence, the energy of the retrieval state tends to -0.5. At $\alpha = \alpha_c$, m=0.967 and the energy is computed to be -0.5014. This is a rather amusing result. If the alignment were perfect at α_c with r=1, the energy would have been -0.5 again. The system finds it slightly more advantageous in energy (a reduction of 0.28%) to relax about 1.5% of the spins from perfect alignment with the pattern.

In the spin-glass state, m=0 and Eq. 6.43 gives

$$C = \sqrt{\frac{2}{\pi\alpha r}}.$$

Substituting C in the equation for r one finds

$$r = \left(1 + \sqrt{\frac{2}{\pi\alpha}}\right)^2.$$

When this expression for r is substituted into the the energy, Eq. 6.45, and m is set to zero, the result is

$$E_{SG} = \frac{\alpha}{2}(1-r) = -\frac{1}{\pi} - \sqrt{\frac{2\alpha}{\pi}}.$$

At $\alpha = \alpha_c$ the value of the spin-glass energy is E_{SG}=$-$ 0.615. This energy is lower than the energy of the retrieval state, E=$-$0.5014, at

the same α, a fact we have mentioned in discussing the retrieval states in the previous section. As $\alpha \to 0$ the energy tends to $-1/\pi$, which is just the value of the energy for highly mixed symmetric spurious states in the finite p system, supporting again the identification of the spin-glass state as the collective remnant of the finite-p spurious states.

Dynamical interpretation of the abrupt transition

We have emphasized above the conjunction of two phenomena as α crosses its critical value,

1. The transition is abrupt. At $\alpha = 0.138$ the overlaps drop suddenly from 0.97 to 0.
2. The energy of the retrieval state is higher than the energy of the spin-glass state, which is also stable at and below α_c.

This situation deserves a few comments.

First, in Section 4.5.4 we have discussed the difference in attitude to meta-stable states between thermodynamics and ANN dynamics. Here we are dealing with a very specific dynamical process, which is the schematized neural dynamics, alias Glauber heat-bath dynamics, sequential or parallel. In such a process, any state in which all spins are aligned with their local fields is perfectly stable, even if it lies higher in energy. Since retrieval is our business, the scope of thermodynamics must be transgressed.

A second issue is the discontinuous change in the retrieval overlap as α crosses α_c continuously. This apparent discrepancy disappears if one contemplates for an instant Figure 6.7. In this figure we draw the qualitative shape of $E(m)$ for three values of α. The variable r has been eliminated in favor of m. The rule of the game of associative memory is that a ball is rolling on a rough surface whose shape is the energy function. It necessarily stops at minima, local or global. If the energy surface is like curve (a), then initial conditions to the right of the maximum lead to the absolute minimum, at $m = m_1$. This would be the situation for $\alpha < 0.051$. Despite the fact that the $m = 0$ state is a local minimum which is higher in energy, it will attract balls starting to the left of the hump. As α increases, the energy surface becomes like (b).[5] As far as we are concerned, the situation in (b) is

[5] The thermodynamicist should note that on the way there was a surface at which the two minima were at equal height. This would have been a point of a

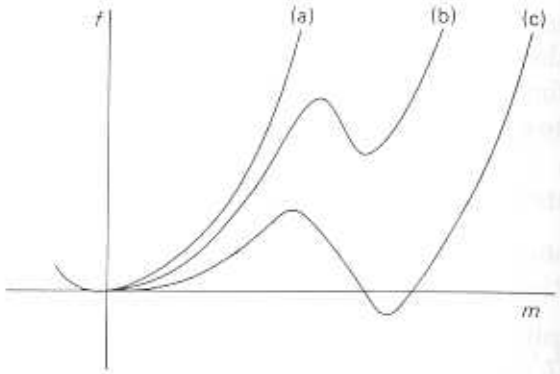


Figure 6.7: Energy as a function of overlap (magnetization) near a 'first-order transition': (a) $\alpha < 0.051$; (b) $\alpha < 0.138$; (c) $\alpha > \alpha_c$.

not much different from (a), except that the basin of attraction might have shrunk. On a surface like (b) there is good retrieval with quality m_2. As α crosses α_c the hump in the surface disappears continuously. But, the ball which might have stopped near the bottom of the hump, must now roll all the way down to the bottom, at m=0.

6.4 Memory Saturation with Noise and Fields

6.4.1 A tour in the T-α phase diagram

We now return to the full mean-field equations, 6.32, 6.33, 6.34, keeping T, or β finite, but setting $\mathbf{h} = 0$. The first part of the discussion will consist of a guided tour through the T–α *phase diagram*, which is depicted in Figure 6.8.

- It represents the lines in the T–α plane across which the network changes its qualitative behavior.

In contrast to thermodynamic phase diagrams, here the boundaries are not restricted to changes between different thermodynamic equilibrium phases, but include the appearance of metastable states, since those are relevant to the performance of the network as an associative memory. Once the tour is over we shall turn to some technical details.

thermodynamic transition of first order. As α crosses that value, the **equilibrium** value of m jumps from a non-zero value to zero.

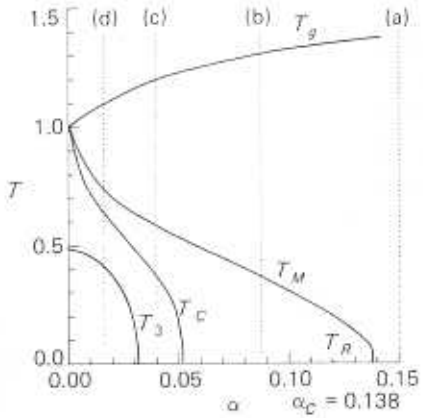


Figure 6.8: T–α phase diagram of a network near saturation. T_g is the transition temperature into the spin-glass phase vs α; at T_M retrieval appears; at T_c the retrieval phase becomes global minimum of the free-energy; below T_3 mixtures of three patterns appear; below T_R replica symmetry breaks[1,34].

In the plane of α and T we have presented the five most important lines. The line T_g separates a high noise phase, which is ergodic (paramagnetic), from a spin-glass phase. The line $T_g(\alpha)$ gives the dependence of the transition temperature on the loading of the network's memory. Varying the noise level in a network with a fixed number of stored patterns, fixed α, amounts to moving along straight vertical lines like the lines marked (a)–(d) in Figure 6.8. Starting high on one of these lines of fixed α, one is in an ergodic phase. On decreasing the noise level one crosses into a spin-glass phase. The temperature at which this will happen depends on α, as

$$T_g(\alpha) = 1 + \sqrt{\alpha}. \tag{6.47}$$

Below this line there is still no retrieval, $\mathbf{m}$=0. Yet the system is no longer fully ergodic, as is witnessed by the fact that q, of Eq. 6.29, is nonzero. The value of q develops continuously from zero as T decreases below the line T_g.

Moving down the vertical line (a) $\alpha > 0.138$, the reduction of noise does not bring about retrieval, down to T=0, where the value of q reaches saturation at q=1. This is consistent with the discussion of the previous section which concluded that at T=0 there is no retrieval for

such high values of α. On line (b), $0.05 < \alpha < 0.138$. As the noise is decreased, below the line T_g, the amplitude of the spin-glass order-parameter increases. Then a second line, $T_M(\alpha)$, is crossed. Below this line there are additional non-ergodic scenarios to the pure spin-glass. The network develops $2p$ $(=2\alpha N)$ meta-stable retrieval states, each with a macroscopic overlap, $m \neq 0$, with a single pattern. The choice of a particular attractor, upon the reduction of noise, depends on the network state of the system upon crossing the transition line T_M.

The value of T_M depends on α, and as α decreases the system can retrieve at higher levels of fast noise. This can be turned around to say that the critical α, α_c, is a decreasing function of T. As $T \to 1$, $\alpha_c \to 0$. This is but the familiar fact that above $T = 1$ a network cannot retrieve even a finite number of patterns. See e.g., Section 4.1.4. At its other end, the line T_M approaches the axis as $\alpha \to \alpha_c$. It does so linearly, according to

$$T_M(\alpha) \approx \frac{1}{C_0 \alpha_c}(\alpha_c - \alpha).$$

The constant $C_0 \approx 0.18$, which implies that T_M vanishes with the very steep slope of about 40.

Along the line T_M, the retrieval states appear with the same type of discontinuity as at T=0. The retrieval overlap m_M goes abruptly from zero to a finite value, as soon as the line is crossed. The value of the limiting overlap depends on the loading, i.e., $m_M = m_M(\alpha)$, which is a decreasing function of α. It tends to zero as $\alpha \to 0$, matching the finite-p situation in which the overlaps vanish continuously at the transition temperature $T = 1$.

When α is further decreased, to values below 0.051, a third transition takes place along a line like (c). Below T_M one crosses the line $T_c(\alpha)$, and below this line the retrieval (ferromagnetic) states become absolute minima of the free-energy. In other words, across this line it is the thermodynamic transition of first order that takes place. This line has no particular significance for our dynamical process. The same $2p$ patterns will remain attractors, alongside the no-retrieval spin-glass state. Note that as $\alpha \to 0$, $T_c(\alpha) \to 1$. This again corresponds to the results for finite p, where retrieval attractors are absolute minima upon their appearance at $T = 1$.

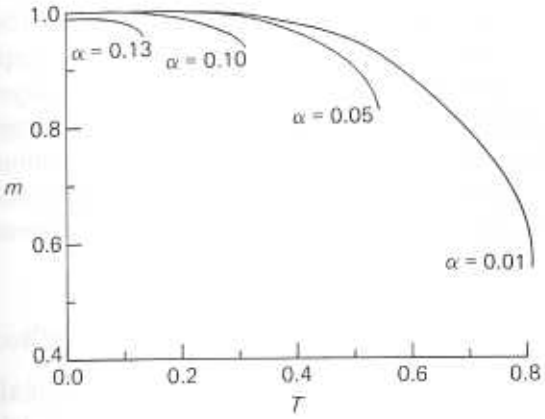


Figure 6.9: Retrieval quality (magnetization) vs temperature for several values of α. Each curve ends at a value of m equal to the discontinuity at T_M[1].

At still lower values of α, $\alpha < 0.03$, one may expect further lines to delimit regions in which spurious state attractors appear. The first of these is the line T_3 in Figure 6.8. It must terminate at T=0.46 when $\alpha \to 0$, as we have learnt in the finite-p case. It has a branch which goes all the way to T=1, but symmetric 3-mixtures are unstable for $T > 0.46$. We will not elaborate on them, since they limit retrieval rather than enrich it. Finally, crawling at the bottom of the diagram is a line, T_R, which is principally of technical interest. Below this line *replica symmetry* of the retrieval states no longer holds. This symmetry has been assumed in the derivation of the mean-field equations for the system. What it implies is that within the (magnetized) retrieval states, a fraction of the spins (neurons) freeze in a spin-glass fashion. The detailed structure of the randomly frozen unaligned spins has a very small impact on the various quantities of relevance for retrieval. Perhaps the most significant effect is the modification of α_c from 0.138 to 0.144 and of the retrieval quality at α_c from 0.967 to 0.983[16].

The retrieval quality m, at fixed α, decreases monotonically as T increases from T=0 to $T_M(\alpha)$. There is no noticeable effect upon crossing the line T_c. The behavior of the retrieval quality as a function of T for several values of α is shown in Figure 6.9. Each curve stops at $m = m_M(\alpha)$, which is the retrieval quality just before retrieval blackout.

The various properties of the network described in this section have been obtained by analytic or numerical investigation of Eqs. 6.32–6.33.

Many further results, mainly concerning the basins of attraction of the retrieval states as a function of α, as well as the properties of the spin-glass state beyond retrieval, have been studied by simulations. Those will be described in Section 6.5.2. Before describing some of the technical details involved in the analysis of the equations, we turn to one additional dimension which affects retrieval. This is the effect of *externally* imposed fields (PSP's), in contrast to those generated by the dynamics of the network.

6.4.2 Effect of external fields – thresholds and PSP's

The neuronal analog of external fields is either a set of thresholds, as described in Section 2.1.2, or a distribution of PSP's, which are fed into the neurons of the network from outside. Formally the effect of both is the same. Such neuronal variables can act in several roles:

- They can serve as an input mechanism, either for external stimuli or for communication between networks,

i.e they replace the imposition of an initial network state. If they are very large they will impose the state that is coupled to them. But even if they are not so strong, they enhance the tendency of the network to drift to an attractor that is correlated with them. Such was the case, for example, when chimes came in to be counted in Section 5.4.

- They can highlight certain memories, even beyond the saturation blackout.
- They may also intervene in some specialized learning. If thresholds are plastic, their modification may bring about a preferred retrieval of certain stored patterns.

We start by recalling the way external fields enter to affect the network's dynamics. A field that is coupled to patterns $1, \ldots, s$ contributes to neuron i a PSP of the form:

$$h_i = \sum_{\nu=1}^{s} h^\nu \xi_i^\nu .$$

This adds to the energy of the system the term:

$$H_h = -\sum_i \sum_{\nu=1}^{s} h^\nu \xi_i^\nu S_i .$$

See e.g., Eq. 3.14. The mean-field equations, Eqs. 6.32–6.34, have been written to allow for the possibility that such fields are present. It should be clear from the term added to the energy that the greater the value of some particular h^ν the higher the likelihood that the elements of the network enter the states $S_i = \xi_i^\nu$, since this is a way of lowering the energy, or of increasing the signal. Patterns coupled to fields we will call 'marked', and proceed to describe the retrieval properties of such patterns.

Single marked pattern at $T = 0$

For a field which acts on a single pattern ($s = 1$) in the absence of noise, the mean-field equations become:

$$m = \mathrm{erf}\left(\frac{m+h}{\sqrt{2\alpha r}}\right) \tag{6.48}$$

where m is the retrieval quality (magnetization) of the marked pattern. Eq. 6.43 becomes

$$C = \sqrt{\frac{2}{\pi \alpha r}} \exp\left(-\frac{(m+h)^2}{2\alpha r}\right) \tag{6.49}$$

and Eq. 6.42 for r remains unchanged, i.e., $r = (1-C)^{-2}$.

Analysis of these equations reveals that for large α there is a single solution, which is the spin-glass state. It has a small, non-zero value for m, which develops continuously from zero, linearly in the amplitude of the marking field. At the same time $q = 1 \neq m^2$, which indicates that this is a fully developed spin-glass random freezing, slightly polarized by the external field. In neural terms this is a situation in which a randomly selected subset of neurons fire bursts, while the rest are quiescent. It is the input which induces some correlations between the distribution of firing and refractory neurons and the pattern marked by it.

As α decreases, a meta-stable state appears when $\alpha = \alpha_c(h)$. This solution is stable while the spin-glass state persists, much like the situation at $h = 0$, described in Section 6.3.3 in connection with Figure 6.7. The *h–α phase diagram* is presented in Figure 6.10. For fixed h, as one crosses the line α_c, coming from above, the marked state can be retrieved with very good quality. The value of α at which the marked

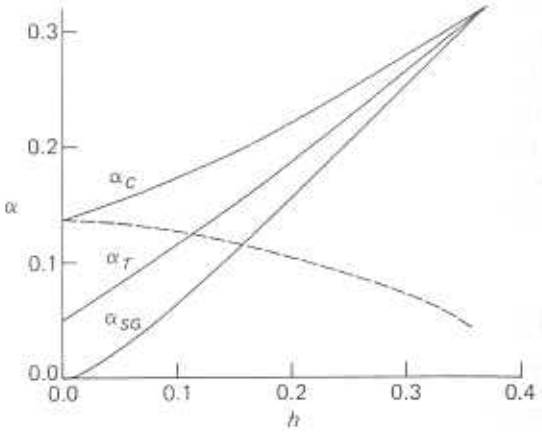


Figure 6.10: The h–α phase diagram. Below line α_c the marked state is retrievable; dashed line is the critical α for the retrieval of unmarked pattern; below α_T the retrieval state is absolute minimum and below α_{SG} there is no spin-glass state[1].

state becomes retrievable increases with α, starting at $\alpha_c(0)(= 0.138)$ and reaching 0.3. A marked pattern can therefore be retrieved even after the network has long been oversaturated. At the same time, the enhancement in the retrieval of the marked pattern generates extra noise on the unmarked ones. This is represented by the dashed line in Figure 6.10. Below $\alpha_c(0)$ the marking of a pattern may actually damage the retrieval of the other patterns since their critical α decreases. But, if the system is already overloaded, the highlighting of a pattern may allow the retrieval of that pattern. The other patterns cannot be damaged, as they had already been in the dark. Some additional *curios* are featured in Figure 6.10.

A stored pattern marked with $h = 0.2$ can be retrieved up to $\alpha = 0.22$. At this storage level the retrieval quality is $m = 0.95$, i.e., 2.5% errors. For the same value of h but $\alpha = 0.2$, way above the critical storage for $h = 0$, the errors amount to less than 1%. The most significant fact is that a memorized pattern, submerged by noise in an overloaded network, is retrieved with a very large overlap relative to the overlap of an attractor imposed by a field of the same strength coupled to a pattern which has not been in memory[1]. For example, if $h = 0.3$, the retrieval quality, m, of the marked pattern will be very

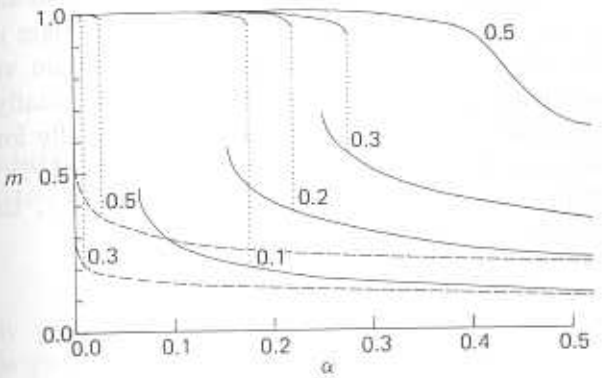


Figure 6.11: Retrieval quality vs α. The solid lines correspond to a marked pattern. Below $h = 0.37$ they are made of two parts – low-m polarized spin-glass and high-m retrieval. The curve with $h = 0.5$ is continuous. Dashed lines are corresponding m vs α for unmemorized marked patterns[1]).

near unity up to storage levels of $\alpha = 0.25$. In contrast, imposing on the over-saturated network a field coupled to a random pattern η_i,

$$h_i = h\eta_i,$$

with $\eta_i = \pm 1$, but uncorrelated with any of the memorized patterns, produces a negligible effect. For example, for the same amplitude (h=0.3) it will lead to an attractor with an overlap $m < 0.2$, for α as low as 0.02. In Figure 6.11 we plot the retrieval quality m vs α for several values of the amplitude of the external field acting on a single pattern.

6.4.3 Fields coupled to several patterns

An interesting phenomenon takes place when one marks by external fields several patterns. This can be done by adding to each neuron the PSP

$$h_i = h \sum_{\nu=1}^{s} \xi_i^\nu. \tag{6.50}$$

It turns out that the noise generated by one marked pattern on another is very effective in destroying the retrieval of all these patterns. As a consequence only a very small number of patterns, s in Eq. 6.50, can

be highlighted in this way, i.e., by enhancing their critical value of α_c. Moreover, beyond s=2! the enhancement of α_c is no longer monotonic in the strength of the marking field. Beyond a certain value of h, which decreases with increasing α, α_c drops and eventually becomes lower than $\alpha_c(h=0)$. In fact, α_c increases monotonically for s=1 and 2. For s=3, the maximal increase in α_c is only about 15% above the h=0 value. When the strength of the field reaches 0.1, the marked patterns will not be retrievable above $\alpha_c(h$=0)[1].

6.4.4 Some technical details related to phase diagrams

If one focuses attention on attractors which retrieve at most one pattern, then one can derive the various phase diagrams – (α–T) or (h–α) etc. – by numerical solution of the mean-field equations. They reduce to:

$$m = \int \frac{dz}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tanh[\beta(\sqrt{\alpha r}z + m + h)] \quad (6.51)$$

$$q = \int \frac{dz}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tanh^2[\beta(\sqrt{\alpha r}z + m + h)] \quad (6.52)$$

$$r = \frac{q}{(1-\beta+\beta q)^2}. \quad (6.53)$$

Much of the detail in the various phase diagrams was obtained in this way. Yet a fair amount of insight into the qualitative structure of these phase diagrams can be obtained analytically.

The spin-glass phase

Coming from the high temperature side with h=0, one has m=0. The first question is about the boundary of the spin-glass phase – the line T_g in Figure 6.8. At high-T, low β, the only solution of the equation for q, is $q=0$. The line T_g is that line in the α–T plane below which a solution with $q \neq 0$ appears. Anticipating that q will develop continuously from zero, the right hand side of Eq. 6.52 is expanded in powers of q to give, at lowest order:

$$q \approx \beta^2 \alpha r.$$

From Eq. 6.53 one finds that r is of first order in q and hence we have:

$$q = \frac{\beta^2 \alpha q}{(1-\beta)^2} + O(q^2).$$

To find the transition temperature T_g we equate the coefficients of the linear terms, which leads to

$$T_g = 1 + \sqrt{\alpha}. \quad (6.54)$$

One easily convinces oneself that for $T > T_g$ the only solution of Eq. 6.52 is q=0. Below, we expand in $T_g - T$, to find that

$$q \approx \beta^2 \alpha r \approx T_g - T. \quad (6.55)$$

Retrieval states

It can be shown that for $\alpha > 0$ the state with zero magnetization is stable, by calculating the corresponding susceptibility,

$$\chi^{\mu\nu} \equiv \frac{\partial m^\mu}{\partial h^\nu}\,|_{m=0} = \delta^{\mu\nu} \frac{\beta(1-q)}{1-\beta(1-q)},$$

which remains finite at all temperatures. The fact that retrieval states do nonetheless show up at low enough temperature is related to the discontinuous nature of the transition. Perusal of Figure 6.7 should make it clear that whether the transition takes place when the free-energy has the form (a), or the corresponding dynamical transition with the form (b), both old and new phases remain local minima even after the transition is over. In physical systems, this behavior underlies the phenomena of super-heating and super-cooling. It is absent in continuous transitions, such as the transition into the spin-glass state or the transition into the retrieval state in the finite-p case.

One implication of the above discussion is that it is not possible to study the form of the line T_M by analytic means, since there could be no expansion in m. Where one can make analytic progress is in the corner of the phase diagram near α=0 and T=1. There one would expect both q and the discontinuity in m (it vanishes as $\alpha \to 0$) to be small and the initial form of the lines T_M and T_c can be worked out, as we proceed to show. There are three small parameters t ($\equiv 1-T$), m

and q. The three equations 6.51–6.53 are expanded in powers of these parameters to give, respectively:

$$\begin{aligned} t &= \tfrac{1}{3}m^2 + \alpha r \\ q &= m^2 + \alpha r \\ r &= \frac{q}{(q-t)^2}. \end{aligned}$$

Note that a factor of m was divided out of the first equation, since we are looking for a non-zero solution for m. Eliminating q and r in terms of m and writing

$$\tau \equiv \frac{t}{\sqrt{\alpha}}, \qquad y \equiv \frac{2m^2}{3\sqrt{\alpha}}, \tag{6.56}$$

we arrive at a single equation for y, which has to be solved for positive values of y,

$$g(y) \equiv \tfrac{1}{2}y^3 - \tau y^2 + y + \tau = 0. \tag{6.57}$$

This equation has either two positive solutions, or none at all. The line T_M is determined by the disappearance of the two solutions. For low values of τ there are two positive solutions. Then, at $\tau = \tau_M$, the two solutions merge and at still higher τ there would be no positive solution. The condition which determines τ_M, and therefore T_M, is the joint vanishing of $g(y)$ and of its derivative. This is a set of two equations for τ_M and y_M, which gives $\tau_M = 1.95$ and hence

$$T_M = 1 - 1.95\sqrt{\alpha}.$$

The line T_c also emerges from the point $\alpha = 0$ and $T = 1$, in Figure 6.8. It is determined by the values of α and T for which the value of the free-energy at the non-zero solution of the mean-field equations equals the free energy at m=y=0. In addition to the equation for the zeroes of $g(y)$ we need an expansion of the free-energy, Eq. 6.36, with $\mathbf{h}$=0 and a single non-zero m, in powers of m, q and t. This leads to,

$$f(m) - f(0) = \tfrac{1}{2}\alpha\left(\ln y - \tau y - \tau y^{-1} + 2\tau + \tfrac{1}{4}y^2 + \tfrac{1}{2}\right) = 0$$

as the second equation, to go with $g(y)$=0. When the two are solved together the result is τ_c=2.6, or,

$$T_c = 1 - 2.6\sqrt{\alpha}.$$

The expressions for T_M and T_c as functions of α give the order in which these two phase boundaries emanate from the point α=0, T=1 in Figure 6.8.

6.5 Balance Sheet for Standard ANN

6.5.1 Limiting framework and analytic consequences

The entire discussion of the properties of ANN's presented so far was based on a simplified and idealized homogeneous biological system. In this section, we will list the assumptions we are aware of, and which have allowed a large variety of network properties to be exposed analytically. Then we will proceed to a recapitulation of the salient results so derived. In the following section, we will move on to domains which are more resistant to analytic inquiry. Those will be described by reference to systematic simulations, which will also be utilized to obtain further insight into the implications of some of the analytic results, especially where approximations could not be avoided.

1. Simplifications
   (a) Neurons are discrete two-state elements.
   (b) There is no internal dynamics to a neuron — PSP's disappear every cycle-time; thresholds are reestablished every cycle-time.
   (c) Synapses, excitatory and inhibitory, are distributed at random on the axons of each neuron.
   (d) In the absence of spikes there is no transmitter release and hence no PSP.
   (e) Dynamical temporal randomness in collecting PSP's and in generating spikes is replaced by synchronous or asynchronous updating procedures.
   (f) Full connectivity.
   (g) Symmetric connection matrix.

2. Idealizations
   (a) All neurons and all synapses carry their information with full fidelity, at whatever analog depth that may be required.
   (b) The memorized patterns are fully random and uncorrelated.

(c) Memories are stored in the synapses in the simple (Hopfield) form.

(d) All memories are stored with equal weights.

3. Homogenization

(a) There is no spatial structure in the neuronal assembly — no architecture.

(b) Every bit of a network activity pattern is as likely to change in a dynamical cycle as any other.

(c) There is no reduction of information — for N input bits there are N output bits.

Many of these simplifying assumptions will be relaxed when we discuss robustness in Chapter 7. The others ought to be subjects for further research. What transpires is the very remarkable fact that most of the features of the more realistic models are well captured by the simplified one. When the various simplifications and idealizations are lifted, one is still left with a homogeneous network. This homogeneity is here considered as a feature of a relatively small module within a vast structured network, such as the cortex. In the wider context, there are loosely connected modules, which are very strongly connected internally. Some of the modules perform various functions in parallel and others is series. The logical composition of the modules to form a comprehensive system has to be guided by cognitive psychology and it is a task we are only beginning to perceive. A glimpse of it has been suggested in our discussion of the network counting chimes. Yet, clearly, the properties of the individual homogeneous ANN will be essential determinants of the feasibility of a plausible architecture composed of such units.

The following is a list of properties of ANN's as drawn from Chapter 4 and from the present chapter.

1. **Noiseless ANN with finite number of memories.**

(a) Broken ergodicity — two types of dynamical development, attractors are cognitive events.

(b) All stored patterns as well as their reversed partners are stable, error correcting attractors.

(c) In a memory storing p patterns there are 3^p spurious attractors which are symmetric mixtures of the embedded patterns and of their reversed companions. In addition, there are spurious attractors which are non-symmetric mixtures.

(d) Synchronous and sequential dynamics have the same fixed-point attractors. In synchronous dynamics there are also two-cycle attractors. But those are irrelevant in that they are associated with unstable spurious states[34].

2. **The effect of noise on ANN with finite number of memories.**

(a) There are no stable states, but there are attractor distributions.

(b) If the noise is not too high, the network will wander in regions in its space of states in which the overlaps with memorized patterns fluctuate little about fixed values.

(c) The memorized patterns as well as their reversed states are the most stable attractors for all noise levels for which ergodicity is broken.

(d) As noise level mounts, the spurious states are gradually destabilized. First to be destabilized are the spurious states with the highest number of mixed states. The last to be destabilized, at T=0.46, are the symmetric 3-mixtures. At this point the retrieval quality (the average overlap) is still as large as $m = 0.97$.

(e) Above a certain noise level, T=1, the system becomes ergodic.

(f) Retrieval becomes inoperative before this noise level is reached, because as T increases toward T=1 the overlaps decrease continuously to zero.

3. **Noiseless ANN with extensive loading.**

(a) If the loading is higher than $0.14N$, there is no retrieval: All m^μ=0.

(b) Immediately below critical loading all memorized patterns and their reversed states are retrievable with quality better than 0.97, less than 1.5% errors.

(c) For $0.03 < \alpha < 0.14$ there are only two types of attractors: the memorized patterns and the spin-glass in which all overlaps with the memories are zero.

(d) Below α=0.03, discrete spurious states begin to appear.

(e) At α_c, the retrieval attractors disappear abruptly, and due to the symmetry between the different memories they disappear all together. This has been referred to as the 'blackout catastrophe'.

(f) Even above the critical storage, one is able to read patterns by highlighting them with a set of weak thresholds, or PSP's, correlated with the pattern.

(g) If one tries to highlight several patterns above α_c, they interfere and highlighting more than 3 patterns has no useful effect.

4. **Extensive storage in the presence of noise.**

(a) Fast synaptic noise, the analog of temperature, reduces the storage capacity of the network by adding to the slow (quenched) noise produced by the random overlaps. As $T \to 1$, the storage capacity goes to zero. Above T=1 no retrieval is possible.

(b) As for the finite-p case, noise eliminates spurious states by reducing their α_c.

(c) The disappearance of retrieval in the presence of noise is also discontinuous, but the discontinuity decreases as the level of noise increases.

(d) The addition of a very small amount of noise makes *replica symmetry* stable, for the retrieval states.

6.5.2 Finite-size effects and basins of attraction: simulations

All the properties of the network listed above have been obtained analytically within certain approximations. In the case of extensive storage of memories, the two most significant approximations have been the limit $N \to \infty$ and the approximation of replica symmetry. Even within these approximations there are certain properties which have not found satisfactory answers. Perhaps the most relevant one is the size of the *basins of attraction*. This is the size of the region of network

states around each memory within which all states are attracted by the dynamical process to a close neighborhood of a memory. Alternatively, and more in the spirit of this text, the basin of attraction is the size of a region around a memory in which all states are attracted to the memory within a prescribed time. Below we will provide some information on this question derived by simulation.

The approximation of infinite N has two implications. First, quantities which have finite values in this limit, such as retrieval overlaps, will be affected by fluctuations when N is finite. Those would be corrections of lower order in N, which are negligible in the limit, and consequently have not been captured by the theory at the level presented here. Second, there are regions in parameter space where the extensive properties vanish in that limit. There, the correction terms are the leading ones. Such, for example, is the situation with regard to the spin-glass state, both above and below the 'blackout', where all overlaps are calculated to be zero. Experience from spin-glass theory indicates that the attractor structure in the unmagnetized spin-glass is extremely complicated and quite resistant to analysis.

This brings us to the question of the effects of replica symmetry breaking. There has been some progress in the analysis of these effects, where they are expected to be small, i.e., in the retrieval states[16]. This has led to the result that α_c=0.144 rather than the 0.138 of the replica symmetric analysis. But only one step has been carried out in this direction and it took simulations to convince us that this value was quite accurate, thereby discouraging efforts to compute further corrections to the replica symmetric theory. It is the description of all these simulations to which we turn now.

Finite-size effects above and below α_c

The effects of the finite size of the system, as well as the possibility that further attractor structure may hide in lower order terms in N, have been estimated by simulations. Those were carried out on networks with N= 500, 1000, 2000, 3000 and $p = \alpha N$ stored random patterns, with α=0.14, 0.16. The synaptic efficacy matrix was of the type given by Eq. 6.1. For each network, the simulation is started with 200 initial states, each of which is a memorized pattern. Then a noiseless, asynchronous sequence of updatings is performed until the network reaches a state which does not vary under the dynamics.

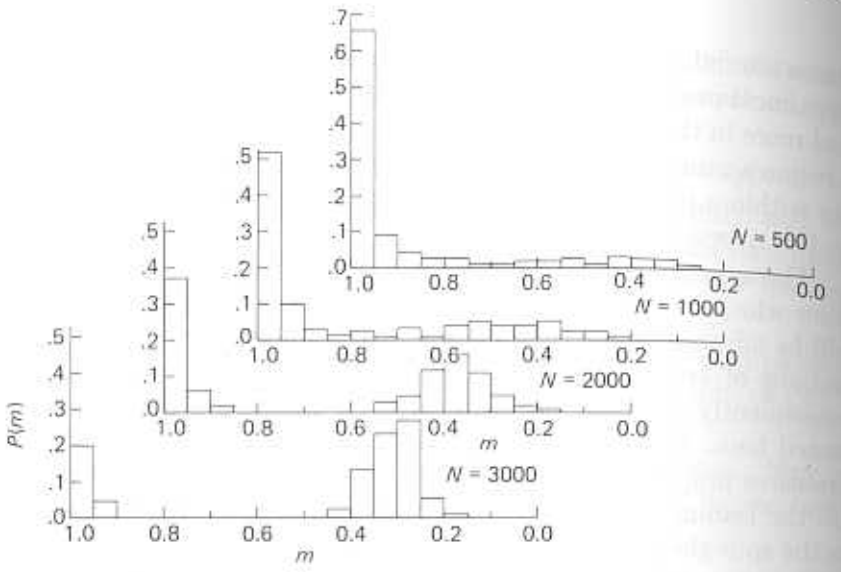


Figure 6.12: Four histograms of attractor statistics in networks with different numbers N of neurons and α=0.14. The dynamics is asynchronous with T=0 and each histogram represents 200 runs with a memory as initial state.

Upon reaching an attractor state, the overlap of the final state with the memory which served as the initial state was measured. These overlaps were then represented in a detailed histogram, by displaying the fraction of starts which terminated within five percentile bins of the overlap interval (0–1). In each of the Figures 6.12 and 6.13 there are four histograms, one for each value of N. The two figures show a significant size dependence of the performance of the network when all other parameters are kept fixed. In both, one observes a fair number of attractors with $m \neq 0$, which are not accounted for by the solutions of the mean-field equations. In Figure 6.12, with α=0.14, most of the attractors are near m=1. There are other attractors which are neither very close to m=1, as would have been expected if $0.14 < \alpha_c$, nor close to m=0, if 0.14 were to be above α_c. In Figure 6.13 one observes fewer attractors near m=1 and many more near m=0.35.

The remarkable difference between the two figures is in the variation with N. While in Figure 6.12 the distribution collapses into the high overlap bin as N goes from 500 to 3000, essentially eliminating all attractors with medium overlaps. In Figure 6.13 the opposite is the

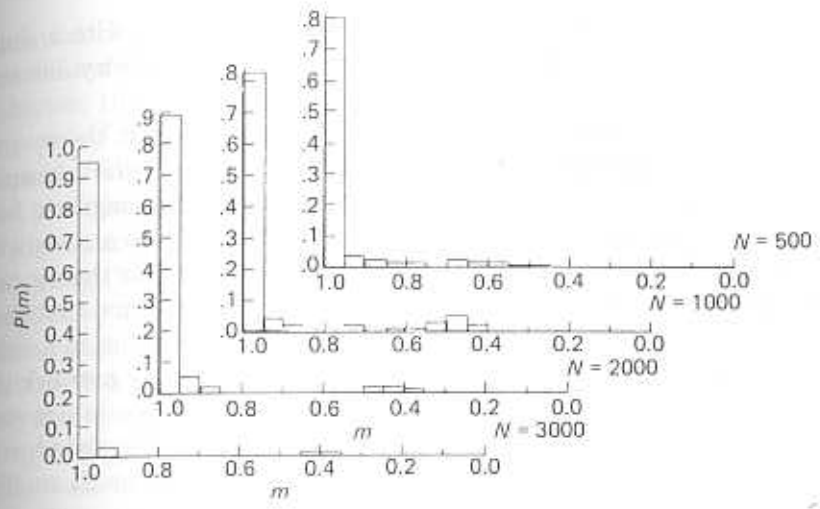


Figure 6.13: Four histograms of attractor statistics in networks with different numbers N of neurons and α=0.16. The dynamics is asynchronous with T=0 and each histogram represents 200 runs with a memory as initial state.

case. It is the bar in the high bin that shrinks as N increases, and the distribution is shifted entirely to the neighborhood of m=0.35. We have to keep in mind that the attractors that are here being probed are those that are reached from initial states that are stored memories. These simulations lead to the following conclusions:

1. The value of α_c is higher than 0.14.

2. The interpretation of α_c must be the storage level below which the high-m attractors dominate as N increases, and above which the high-m attractors disappear with increasing N.

The appearance of medium-m attractors in Figure 6.12 can be accounted for by invoking the finiteness of N and they do seem to disappear as $N \to \infty$. On the other hand, the distribution of attractors in Figure 6.13 is quite persistently concentrated at $m \approx 0.35$. It cannot be explained away as a finite-N effect. Instead, it is much more like a spin-glass phenomenon. In an unmagnetized spin-glass there is a vast number of meta-stable states, each of which is stable against the flipping of a few spins, but separated from each other and from lower energy states by relatively low barriers. Such states will be effective

attractors in our simulations, because the dynamics deals with a single neuron at a time. Yet they will not be detected by our theory, because the energy barriers grow less than linearly with N.

Attractor with finite overlaps, where replica symmetric theory predicts only zero overlaps, are like the *remnant magnetization* in spin-glasses[35]. When a spin-glass is polarized by a strong magnetic field and then the field is removed, the spin-glass will relax into a state with a finite *remnant magnetization*, where replica symmetric theory predicts no such stable states. There is, however, a difference that may interest the spin-glass theorist. The remnant magnetization in a spin-glass is usually about m=0.15. In the overloaded memory network the remnant overlaps are more like m=0.35. This difference is not well understood. In fact, if the initial state is taken to be uncorrelated with the stored patterns, then the remnant magnetization is much smaller – it is about 0.08 for α=0.16 and 0.12 for α=1[1].

The data of these simulations has been analyzed by writing

$$\Pr(m, \alpha, N) = A \exp[B(\alpha - \alpha_c)N]. \tag{6.58}$$

The three parameters α_c, A and B were then determined by a least squares fit to the entire set of simulation data. The result is α_c=0.144 ±0.009. The same question was addressed by analytical methods whereby the first step along the systematic treatment of *replica symmetry breaking* was undertaken. The gratifying, if somewhat mysterious, result is that this single step gives the value α_c=0.144[1,16].

Basins of attraction and retrieval times

The second question that has been addressed by simulations, for lack of a better tool[6], is a combined question. It consists of measuring the average sizes of the *basins of attraction* of the memories in the network as well as the retrieval times, as a function of the storage level α. Often these questions are considered unrelated, but in a biological system and even in a device there is likely to be a time limit on the process of recall or retrieval. In other words, if a stimulus has not led to an attractor within a physically prescribed amount of time, that stimulus will be ignored. This time window may be longer than the time necessary for retrieval of patterns by a stimulus at the extremities of the basin of

[6]Progress has begun to be made in this direction. See e.g., refs. [27,28].

attraction. In that case the question of the size of the basins decouples from that of the speed of recall. On the other hand, if the time window is shorter, then *it* will determine the size of the basins for all retrieval purposes of the network. Presently, we have no information on the relevant times involved. Consequently we will proceed to display the data as if the two issues were indeed separate.

Simulations were performed[12] on a network with 1000 neurons at two storage levels, 100 and 130 patterns, i.e., α=0.10 and 0.13, respectively. The dynamical process was noiseless (T=0) and asynchronous with a random updating sequence. It should be pointed out that basins of attraction as well as retrieval times are rather sensitive to the type of dynamics, as well as to the sequence in which neurons are selected for updating. The main significance of randomness in this context is in the indiscriminate treatment of 'correct' neurons and 'erring' neurons. The stimuli were produced from the stored patterns by flipping at random a given fraction of the spins. This produced stimuli with a given initial overlap, or Hamming distance, from the patterns. Each stimulus is then subjected to the dynamics, until the network reaches an attractor.

The results are presented in Figure 6.14. The data points are averages of the overlaps at the attractors reached from these stimuli (upper curves), and of the retrieval times (lower curves). Each point is an average over 400 initial stimuli with the same initial overlap. The squares correspond to α=0.10 and the circles to α=0.13. The size of the basins of attraction clearly shrinks with increasing α. For α=0.10, the initial overlap can decrease down to 0.5 (a Hamming distance of $0.25N$) and retrieval is almost perfect. It should be kept in mind that very high values of the average final overlap indicate that almost all 400 initial configurations ran into the attractor. Lower values of $\bar{m}$ reflect the fact that a sizable fraction of stimuli ran away from the attractor and into small overlap attractors. The value of the initial overlap for which significant decrease from $\bar{m}$=1 begins is, therefore, a good measure of the size of the basin of attraction. For α=0.13 such departures begin at m_0=0.75 (Hamming distance of $0.125N$).

The average convergence times $\bar{t}$ in the two lower curves seem to indicate a rapid increase of retrieval time either with initial distance from the pattern or with the storage level α. This is, however, misleading. The longer average times, like the lower final overlaps, are due to the increasing number of stimuli which do not flow to the memorized

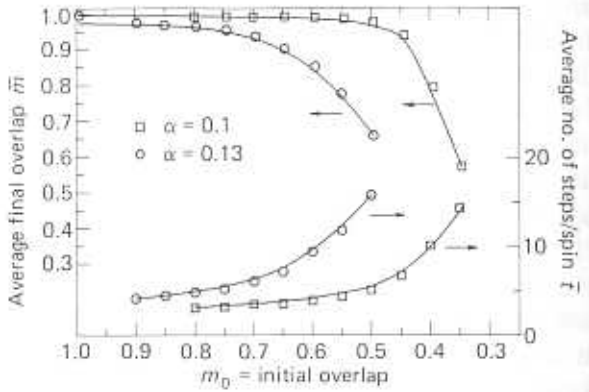


Figure 6.14: Average final overlaps (upper curve, left scale) and average retrieval times (lower curves, right scale) vs distance of stimulus from memorized pattern.

attractor, but instead flow away into spin-glass states. A meaningful comparison of retrieval times for different values of α is between points with equal average final overlap. Such pairs of points have essentially equal retrieval times.

6.6 Beyond the Memory Blackout Catastrophe

6.6.1 Bounded synapses and palimpsest memory

The main reason for the total blackout of the network memory upon oversaturation is easily identified. The total symmetry between the stored memories leads to a situation in which it is impossible to lose memories selectively[36]. If noise due to excessive storage obscures one memory, it will typically obscure all memories. This led to a few suggestions for ANN's with synaptic efficacies which are generated by combining patterns with unequal weights. Such networks can serve as generic examples of selective erasure of patterns from memory by newly arrived information — 'palimpsests'. One suggestion[7] was already mentioned in our list of analytic results, Section 6.1.4. We shall come back to that calculation, but first we turn to a much simpler consideration, which had served to stimulate it[36,37].

The basic idea is to view the synaptic structure Eq. 6.1 as the result

of some unspecified learning process. In this process, J_{ij} is modified at each step by the addition of a term

$$\Delta J_{ij} = \frac{1}{N}\xi_i^\mu \xi_j^\mu \tag{6.59}$$

at the μ'th step. As the number of terms in J_{ij} increases, so does the overall noise level in the system, while the new term is added with the same amplitude. The general level of noise in a network storing p patterns can be characterized by

$$K(p) = \langle\langle (J_{ij}^2)\rangle\rangle - \langle\langle J_{ij}\rangle\rangle^2. \tag{6.60}$$

The imprinting of the additional pattern μ is similarly measured by the quantity

$$k(\mu) = \langle\langle (\Delta J_{ij}^2)\rangle\rangle - \langle\langle \Delta J_{ij}\rangle\rangle^2. \tag{6.61}$$

In the standard case of Eq. 6.1, we would have

$$k(\mu) = \frac{1}{N^2},$$

which is independent of μ.

If the different contributions ΔJ_{ij} can be considered uncorrelated, then for large N and p one has the equation:

$$K(p) - K(p-1) = k(p), \tag{6.62}$$

which can also be written as

$$\frac{dK(t)}{dt} = k(t). \tag{6.63}$$

The required imprinting strength, k, of an additional stored pattern is defined by demanding that the pattern be retrievable, with a quality which is set by convention to be 0.97, against a given background of synaptic noise. It turns out, using numerical simulations, that this can be assured if

$$\frac{k}{K} = \frac{\epsilon^2}{N} \tag{6.64}$$

with $\epsilon \approx 2.5$.

Coming back to the standard model, the imprinting strength is independent of the pattern number ("time"). The solution of Eq. 6.63 is simply

$$K(t) = kt$$

and Eq. 6.64 becomes:

$$\frac{k}{K(p)} = \frac{k}{kp} = \frac{\epsilon^2}{N},$$

which reads

$$\alpha_c = \frac{p_c}{N} = \frac{1}{\epsilon^2}.$$

If we recall that the appropriate value for ϵ is about 2.5, then α_c is found to be about 0.16.

A simple extension of the standard model which can cope with the blackout catastrophe is 'learning within bounds'. In this model, described in learning language, when a new pattern is learned the synaptic efficacies would be modified as in the standard model. Every synapse is either modified by a term of the form 6.59 or, if upon modification its absolute value would cross some fixed bound, it becomes equal to the bound. In other words the relation between the values of a given synapse before and after modification is[37]:

$$J_{ij}(p) = f\left(J_{ij}(p-1) + C\xi_i^p\xi_j^p\right) \tag{6.65}$$

where the function f is defined by

$$\begin{array}{rcrcr} f(x) & = & -A & \text{for} & x < -A \\ f(x) & = & x & \text{for} & -A < x < A \\ f(x) & = & A & \text{for} & x > A. \end{array} \tag{6.66}$$

Quite clearly, if the strength of the modification C, in Eq. 6.65, is very small compared to the bound A, then the blackout at α=0.14 would take place before the existence of the bounds had been felt. On the other hand, if the amplitude is large, it is the last pattern to be added which will be best stored, irrespective of how late in the game it arrives. In this case, there is no blackout, but the last pattern may obscure all the previous ones. After all, those synapses that are found at their bounds have values which are favorable to the pattern being

added. The synapses which can be modified will be modified significantly, so that previously stored patterns are liable to be much affected. This implies also that in the process different patterns are stored non-uniformly. In between these extreme ratios of C to A there should be values for which a number of the most recently added patterns are retrievable, while previous ones are erased.

Typically, the sizes of the synapses in a standard ANN with storage $p = \alpha N$ are of order $N^{-\frac{1}{2}}$, since each synapse is a sum of αN $\pm$1's divided by N. The bound A should itself, therefore, be of the same order. If it is taken to be

$$A = \frac{1}{\sqrt{N}}$$

then

$$K(p) \approx 1/N,$$

and Eq. 6.64 would determine the way the imprinting strength must depend on N. Given that there is a critical value for ϵ, which optimizes the number of recent retrievable patterns, we have:

$$k_c = \frac{\epsilon_c^2}{N^2}.$$

The value of ϵ_c has been determined by numerical simulation[36,37]. The results are reproduced in Figures 6.15 and 6.16.

In Figure 6.15 for ϵ=1 the number of retrievable patterns reaches some maximum as a function of storage α, and then decreases to zero as α increases beyond 10%. This is a low ϵ case, which is just the behavior one would expect in the standard ANN. As ϵ increases to 2.7 the number of recent retrievable patterns stabilizes and remains at about 2.5% of N. Figure 6.16 presents evidence that an optimal ϵ exists. Its value is independent of N and is somewhere around ϵ=3. The peak value of the number of retrievable patterns with $m > 0.97$ varies with N and seems to tend asymptotically to $0.016N$.

The statistical mechanician should be moved by the apparent phase transition which Figure 6.16 exhibits. The retrievable fraction of patterns behaves like an order-parameter which vanishes for $\epsilon < 1.2$ and then grows very rapidly as a function of ϵ. As such, it is rather special as it is not monotonic. Whether it is a real phase transition can be ascertained only by an analytical study. This precise model has not

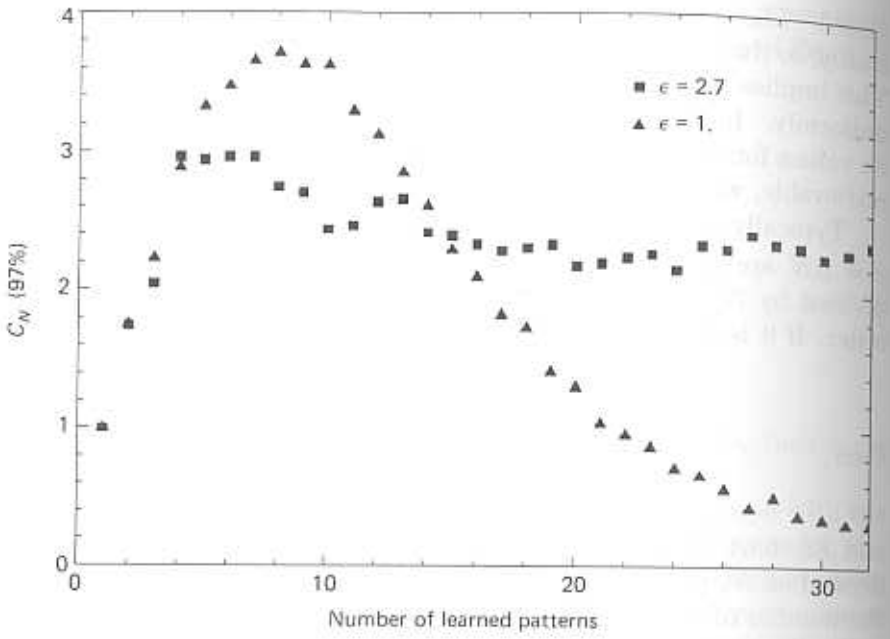


Figure 6.15: Number of most recent patterns, retrievable with quality better than 97%, vs total storage in a network with 100 neurons, for two values of ϵ. (After ref. [36], by permission.)

yet found an analytical treatment. But, a very closely related model mentioned in Section 6.1.4, in which patterns have an exponentially decreasing weight, from last to first, has been solved[7]. See e.g., Appendix 6.7.3. It does undergo a real phase transition as a function of ϵ.

6.6.2 The 7 ± 2 rule and palimpsest memories

It has been repeatedly found that if humans are presented with lists of items, e.g., numbers, they can recall such lists soon after presentation provided they consist of seven plus or minus two items. See e.g., [38]. If it were the mechanism of 'learning within bounds' which is responsible for this phenomenon of short term memory, then the results of the previous section, giving $0.016N$ as the optimal recall of recent memories, would indicate that the relevant network should have about 500 neurons. Moreover, since the N characterizing the network,

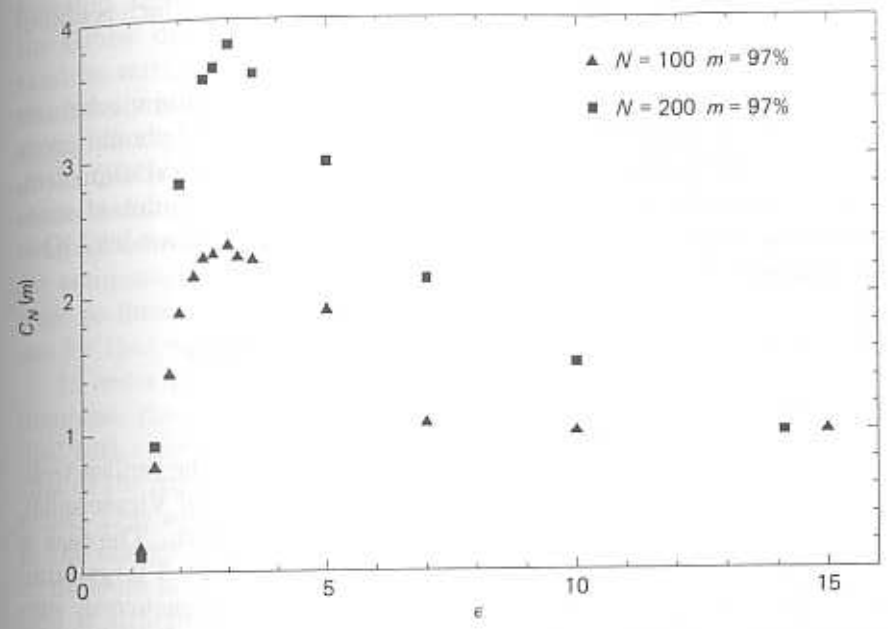


Figure 6.16: Number of most recent patterns with retrieval quality better than 97% vs the imprinting strength ϵ. (After ref. [36], by permission.)

for purposes of storage capacity, is the connectivity of the network, the implication of such experiments would be that 500 is the level of connectivity in that part of the brain which is responsible for the recall. This number, though not outlandish, seems somewhat low, which has motivated a preference for networks which do not operate in their optimal parameter range. In other words, one could choose, for example, in Figure 6.16 a value of ϵ which gives a lower value of C requiring higher connectivity. Note the interesting fact that this attitude goes against the grain of typical tendencies in the field, which emphasize maximization of storage capacity.

If such a model is to capture an element of brain reality and to provide meaningful tests for the theory, it must deal with two issues:

- The experiments involve the memorization and recall of ordered temporal sequences of patterns.

- The experiments involve learning, as well as recall, which is absent in the model.

There has been a recent proposal for a dynamical learning mechanism which keeps synapses within bounds[39]. If this proposal should prove effective and if it will be possible to extend it to temporal sequences, then both points will have been answered and this account of short term memory may become a first direct contact between model ANN's and psychological phenomenology.

6.7 Appendix: Replica Symmetric Theory

6.7.1 The replica method

For an extensive discussion of the ideas in and around the replica technique one should turn to the book of Mezard, Parisi and Virasoro[33]. Here, for the sake of completeness, we bring a bare sketch. The task is to compute the equilibrium properties of a system with a large number of degrees of freedom which depends on a set of *quenched, random* variables. When the number of stored patterns was kept finite as $N \to \infty$ things were simple. It was shown in Appendix 4.7.1 that the free-energy, for example, could be calculated for a **given** realization of the random patterns. Then, as $N \to \infty$, *self-averaging* produced an averaging over the distribution of patterns. This was due to the fact that the finiteness of the number of patterns ensured the appearance of each one of them, when sampled an infinite number of times, with a frequency proportional to its probability.

When p is of order N, this is no longer the case. Yet the properties of the system can be of interest only if they do not depend significantly on the particular detailed realization of the quenched randomness. Such insensitivity must be expressed by the fact that *typical* realizations of the randomness would lead to the same results. Or, in other words, that the relevant properties of the system must be sharply peaked when viewed over the space of the various configurations of the randomness. If this *is* the case, then one should be able to compute them by averaging over the random variables, as if there was an ensemble of systems each with a different realization of the randomness, distributed according to the distribution of the random variables.

The question then becomes: what quantities should be so averaged? The simplest possibility would have been if this were the partition

function, since the random variables usually appear in a simple form in the Gibbs' distribution. But this would be equivalent to treating the random variables as *annealed*, namely as variables which participate in the fast dynamics of the system, just like the spins themselves, for example. A key quantity is the free-energy which generates the moments of the Gibbs' distribution. If this quantity is known for any realization of the randomness, then the other thermodynamic quantities can be computed from it. See e.g., Section 3.4.2. This is a natural candidate for computation by averaging over the randomness. If this quantity were to fluctuate much, then statistical mechanics would be of little use for that system.

In order to perform the averaging of the free-energy over the randomness, the *replica method* has been invented[40]. It is intended to deal with cases in which the averaging of the Gibbs weight is simple but that of the free-energy – the logarithm of the partition function – is difficult. This would typically be the case if the random variables appear in the energy linearly or quadratically, or if they appear uncorrelated at different sites. In those cases, the averaging over the partition function is straight-forward. The observation is then that the logarithm of the partition function can be represented in the form of a partition function, *albeit* of a rather peculiar system. The idea is based on the following limiting form[40]:

$$\lim_{n\to 0} \frac{Z^n - 1}{n} = \ln Z. \tag{6.67}$$

It implies that in order to obtain the mean of $\ln Z$ one can average Z^n, because the additional terms on the left hand side do not depend on the random variables. The hitch is in taking the limit. The average can be performed for any integer n, but not for n=0. One must interpret the procedure to be a computation of the mean of the left hand side as a function of n, and then the limit is reached by an *analytic continuation*. This is the desired quantity. The formal aspects of this analytic procedure leave much to be desired, but this does not render the technique any less useful.

Next, one notes that Z^n is itself like a partition function of n identical systems which, for any given set of random variables, do not interact. These are the *replicas*. They come about in the following way. Suppose that the energy of a spin system is written as $E(\{x\}, \{S\})$,

indicating a dependence on the state of the spins and on a particular set of random variables $\{x\}$. One can write

$$\begin{aligned} Z^n(\{x\}) &= (Tr_S \exp[-\beta E(\{x\},\{S\})])^n \\ &= Tr_{S^1} \exp[-\beta E(\{x\},\{S^1\})] \ldots Tr_{S^n} \exp[-\beta E(\{x\},\{S^n\})] \\ &= Tr_{S^1,\ldots,S^n} \exp[-\beta \mathcal{E}(\{x\},\{S^1,\ldots,S^n\})], \end{aligned}$$

where

$$\mathcal{E}(\{x\},\{S^1,\ldots,S^n\}) = \sum_{\alpha=1}^{n} E(\{x\},\{S^\alpha\})$$

is an energy of a system of $n \times N$ spins. The trace is over the entire set of spin variables.

Once the averaging over the random variables x is carried out, the various replicas of the system begin to interact via the correlations implied by the presence of the same random variables in every replica. The next appendix is an example of the process, as well as of the ultimate limit $n \to 0$.

6.7.2 The free-energy and the mean-field equations

We now proceed to employ the method of Appendix 6.7.1 to the computation of the free-energy of a system with a standard coupling matrix, Eq. 6.1, in which

$$p = \alpha N.$$

In this framework Eq. 3.46 for the free-energy becomes

$$f = -\frac{1}{\beta} \lim_{n\to 0} \lim_{N\to\infty} \frac{1}{nN} \left(\langle\langle Z^n \rangle\rangle - 1 \right). \tag{6.68}$$

It will be a function of the temperature, of the condensed (extensive) overlaps m^μ and of the order-parameters q and r described in Section 6.3.1. The double brackets denote, as usual, the averaging over the quenched disorder which in this case are the stored patterns.

We concentrate therefore on the computation of $\langle\langle Z^n \rangle\rangle$. It is written as:

$$\langle\langle Z^n \rangle\rangle = \left\langle\left\langle Tr_{S^\rho} \exp\left(\frac{\beta}{2N} \sum_{ij\mu\rho} (\xi_i^\mu S_i^\rho)(\xi_j^\mu S_j^\rho) - \tfrac{1}{2}\beta p n + \beta \sum_{\nu=1}^{s} h^\nu \sum_{i\rho} \xi_i^\nu S_i^\rho \right) \right\rangle\right\rangle.$$

Note that the trace is over the configurations of the n-fold replicated system, labelled by $\rho = (1,\ldots,n)$. All the sums in the exponential are unrestricted and the second term, $\frac{1}{2}\beta pn$, corrects for the presence of the self-interaction in the first. The last term in the exponent represents the effect of s external fields, each coupled to one of a finite number of patterns which are candidates for condensation, as explained at the end of Section 3.4.2.

As usual, the quadratic terms in the exponential are linearized by a pn-fold gaussian transformation. See e.g., Eq. 3.50 and Appendix 4.7.1. It leads to

$$\begin{aligned} \langle\langle Z^n \rangle\rangle &= e^{-\beta pn/2} Tr \int_{-\infty}^{\infty} \prod_{\mu\rho} \frac{dm_\rho^\mu}{\sqrt{2\pi}} \\ &\left\langle\left\langle \exp \beta N \left(-\tfrac{1}{2} \sum_{\mu\rho} (m_\rho^\mu)^2 + \sum_{\mu\rho} m_\rho^\mu \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu S_i^\rho \right) \right.\right. \\ &\left.\left. \times \exp \beta N \left(-\tfrac{1}{2} \sum_{\nu\rho} (m_\rho^\nu)^2 + \sum_{\nu\rho} (m_\rho^\nu + h^\nu) \frac{1}{N} \sum_{i=1}^{N} \xi_i^\nu S_i^\rho \right) \right\rangle\right\rangle. \end{aligned}$$

Note that on the right hand side there are pn integrations but the sum in the exponent has been separated into two parts. One sum over μ includes the indices of the $p - s$ patterns which will not become extensive – they will remain of order $O(N^{-1/2})$. The other sum is over the patterns $\nu(= 1,\ldots,s)$. These are the candidates for condensation, hence symmetry breaking and non-ergodicity. This is indicated also by the *symmetry breaking* fields h^ν which accompany them. The number of condensed patterns must remain finite as $N \to \infty$, hence, $p - s \approx p$.

Next we average Z^n over the high ξ_i^μ. The first exponential becomes, after a simple rescaling of variables by $\sqrt{\beta N}$ to ensure a safe thermodynamic limit:

$$\exp\left(-\tfrac{1}{2} \sum_{\mu\rho} (m_\rho^\mu)^2 + \sum_{i\mu} \ln\cosh\left(\sqrt{\frac{\beta}{N}} \sum_{\rho} m_\rho^\mu S_i^\rho \right) \right).$$

The expectation is that the rescaled m_ρ^μ, being random, will be of order 1 and the first term in the exponent will become of order N by virtue of the sum over μ. The second term becomes extensive for a more subtle reason. The argument of the cosh tends to zero for large N and the

cosh itself to 1. The leading term in the expansion of the ln cosh is the quadratic one which is of order N^{-1}, but the sums over i and μ make this term of order N again. This suggests that one needs only the quadratic term in all, because the higher order terms in the expansion of the ln cosh lead to terms which become relatively negligible as $N \to \infty$.

The first exponential can therefore be written as

$$\exp\left(-\tfrac{1}{2}\sum_{\mu\rho}(m_\rho^\mu)^2 + \frac{\beta}{2N}\sum_{i\mu\rho\sigma} m_\rho^\mu S_i^\rho m_\sigma^\mu S_i^\sigma\right),$$

where we see the first indications of an interaction between the different replicas ρ and σ, brought about by the averaging process. Since the result is quadratic in the 'high' m's, they can be integrated out. The matrix of coefficients is the same for all values of μ. It is

$$K_{\rho\sigma} = \delta_{\rho\sigma} - \frac{\beta}{N}\sum_i S_i^\rho S_i^\sigma .$$

Note that the diagonal elements are numbers, $(1-\beta)$, while the off-diagonal ones depend on the spin variables. The integral over every set of m_ρ^μ, with fixed μ, is just the square root of the inverse of the determinant of this matrix. And this has to be raised to the p-th power. Thus the integral over the 'high' m's can be written as

$$\int_{-\infty}^{\infty}\prod_{\mu\rho}\frac{dm_\rho^\mu}{\sqrt{2\pi}}\exp\beta N\left(-\tfrac{1}{2}\sum_{\mu\rho}(m_\rho^\mu)^2 + \sum_{\mu\rho} m_\rho^\mu \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu S_i^\rho\right)$$

$$= [\det\mathcal{K}]^{-p/2} = \exp\left(-\frac{p}{2}Tr\ln\mathcal{K}\right),$$

where we have used a well known identity to express the determinant, in which the trace refers to matrix elements diagonal in the replica indices.

To deal with the fact that the off-diagonal elements of the symmetric matrix $\mathcal{K}$ are functions of S_i, one introduces two sets of $n(n-1)$ variables each, $q_{\rho\sigma}, r_{\rho\sigma}$, to write,

$$\exp\left(-\tfrac{1}{2}pTr\ln\mathcal{K}\right)$$
$$= \int\prod_{(\rho\sigma)} dr_{\rho\sigma}\prod_{(\rho\sigma)} dq_{\rho\sigma}\exp\left(-\tfrac{1}{2}pTr\ln[(1-\beta)\mathcal{I} - \beta\mathcal{Q}]\right)$$
$$\times\exp\left(-\tfrac{1}{2}N\alpha\beta^2\sum_{\rho\sigma} r_{\rho\sigma}q_{\rho\sigma} + \tfrac{1}{2}\alpha\beta^2\sum_{i\rho\sigma} r_{\rho\sigma}S_i^\rho S_i^\sigma\right). \qquad (6.69)$$

The notation $(\rho\sigma)$ implies that $\rho < \sigma$. The matrix $\mathcal{Q}$ is a symmetric numeric matrix with zero diagonal, and $\mathcal{I}$ is the unit matrix. The key to the last equation is that when the variables $r_{\rho\sigma}$ are integrated out they leave behind

$$\delta\left(q_{\rho\sigma} - \frac{1}{N}\sum_i S_i^\rho S_i^\sigma\right)$$

and then, when the q's are integrated over, the original form of the matrix $\mathcal{K}$ is restored. As usual $\alpha = p/N$.

When Eq. 6.69 is inserted in the expression for $\langle\langle Z^n\rangle\rangle$ one finds:

$$\langle\langle Z^n\rangle\rangle = \exp(-\tfrac{1}{2}\beta pn)\int\prod_{(\rho\sigma)} dr_{\rho\sigma}\prod_{(\rho\sigma)} dq_{\rho\sigma}$$
$$\exp N\left[-\tfrac{1}{2}\beta\sum_{\nu\rho}(m_\rho^\nu)^2 - \tfrac{1}{2}\alpha Tr\ln[(1-\beta)\mathcal{I} - \beta\mathcal{Q}] - \tfrac{1}{2}\alpha\beta^2\sum_{\rho\sigma} r_{\rho\sigma}q_{\rho\sigma}\right.$$
$$\left. + \left\langle\left\langle\ln Tr\exp\left(\tfrac{1}{2}\alpha\beta^2\sum_{\rho\neq\sigma} r_{\rho\sigma}S^\rho S^\sigma + \beta\sum_{\nu\rho}(m_\rho^\nu + h^\nu)\xi^\nu S^\rho\right)\right\rangle\right\rangle_\xi\right]. \qquad (6.70)$$

The term in double brackets calls for a word of explanation. It is basically the consequence of self-averaging, which is brought about by the fact that s is finite. In other words, after having performed the average over the $p-s$ uncondensed patterns, the part that depends on the spins S_i^ρ and the quenched random patterns ξ_i^ν is

$$\left\langle\left\langle Tr_{S^\rho} \exp\left(\beta\sum_{i\nu\rho}(m_\rho^\nu + h^\nu)\xi_i^\nu S_i^\rho + \frac{1}{2}\alpha\beta^2\sum_{i\rho\neq\sigma} r_{\rho\sigma}S_i^\rho S_i^\sigma\right)\right\rangle\right\rangle$$

$$= \left\langle\left\langle \prod_i Tr_{S^\rho} \exp\left(\beta\sum_{\nu\rho}(m_\rho^\nu + h^\nu)\xi_i^\nu S^\rho + \frac{1}{2}\alpha\beta^2\sum_{\rho\neq\sigma} r_{\rho\sigma}S^\rho S^\sigma\right)\right\rangle\right\rangle$$

$$= \left\langle\left\langle \exp\left[\sum_i \ln Tr_{S^\rho} \exp\left(\beta\sum_{\nu\rho}(m_\rho^\nu + h^\nu)\xi_i^\nu S^\rho + \frac{1}{2}\alpha\beta^2\sum_{\rho\neq\sigma} r_{\rho\sigma}S^\rho S^\sigma\right)\right]\right\rangle\right\rangle$$

The above transformations should be self-explanatory. In order to arrive at Eq. 6.70 one then applies self-averaging to the sum in the exponent. The result is an average over the ξ's in the exponent. This is independent of the ξ's, and the external average, represented by the double brackets, becomes spurious. As a consequence, the double brackets move into the exponent. The next line in the above chain of equations is

$$\exp N\left\langle\left\langle \ln Tr_{S^\rho} \exp\left(\beta\sum_{\nu\rho}(m_\rho^\nu + h^\nu)\xi^\nu S^\rho + \tfrac{1}{2}\alpha\beta^2\sum_{\rho\neq\sigma} r_{\rho\sigma}S^\rho S^\sigma\right)\right\rangle\right\rangle_\xi ,$$

where the subscript ξ, which is usually omitted, is a reminder that the double brackets are an average over the discrete ξ's.

The free-energy, Eq. 6.68, can be evaluated directly from Eq. 6.70. The main dependence on N is the overall factor in the exponent. It follows that the free-energy can be obtained by the saddle-point technique. At the saddle-points the values of m^ν, q and r are just the averages defined by the order-parameters Eqs. 6.28, 6.29 and 6.31. In terms of these parameters the free-energy reads,

$$f = \tfrac{1}{2}\alpha + \frac{1}{2n}\sum_{\nu\rho}(m_\rho^\nu)^2 + \frac{\alpha}{2\beta n} Tr\ln[(1-\beta)\mathcal{I} - \beta\mathcal{Q}]$$

$$+ \frac{\alpha\beta}{2n}\sum_{\rho\sigma} r_{\rho\sigma}q_{\rho\sigma} - \frac{1}{\beta n}\langle\langle \ln\left(Tr_{S^\rho}\exp(\beta\mathcal{H}_\xi)\right)\rangle\rangle,$$

where the right hand side should be understood as the limit $n \to 0$. That such a limit may exist can be made plausible by observing that:

- in the second term the summation over ρ could make the sum of order n;
- the third term, the Tr, is a trace of a matrix of order $n \times n$, which has n elements in its diagonal, and could be of order n;
- the fourth term has $n(n-1)$ terms in the sum, which in the limit could be proportional to $-n$ (sic).
- the last term is itself like a free-energy of a system with an energy $\mathcal{H}_\xi$, which can be read off Eq. 6.70 to be

$$\mathcal{H}_\xi = \sum_{\nu\rho}(m_\rho^\nu + h^\nu)\xi^\nu S^\rho + \frac{\alpha\beta}{2}\sum_{\rho\neq\sigma} r_{\rho\sigma}S^\rho S^\sigma. \qquad (6.71)$$

This is a free-energy of a system in the replica space and the total number of variables is n. Such a free-energy should be 'extensive' and hence proportional to n.

All the above are just plausibility arguments in an implausible exercise in a space of matrices with zero dimensions. The stationary states of the system are obtained by varying f with respect to all order-parameters.

Replica symmetry

To make further progress, one must make some simplifying assumptions about the nature of the matrices in replica space. The simplest such assumption is replica symmetry. It is defined to read

$$\begin{aligned} m_\rho^\nu &= m^\nu \\ q_{\rho\sigma} &= q \qquad \rho \neq \sigma \\ r_{\rho\sigma} &= r \qquad \rho \neq \sigma. \end{aligned} \qquad (6.72)$$

With these matrices, the computation of the free-energy can be completed. The result is

$$\begin{aligned} f &= \tfrac{1}{2}\alpha + \tfrac{1}{2}\sum_{\nu\rho}(m^\nu)^2 + \frac{\alpha}{2\beta}\left(\ln[1-\beta(1-q)] - \frac{\beta q}{1-\beta(1-q)}\right) \\ &+ \frac{\alpha\beta r}{2}(1-q) - \beta^{-1}\int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}} \\ &\exp(-\tfrac{1}{2}z^2)\langle\langle \ln 2\cosh\beta[\sqrt{\alpha r}z + \sum_\nu (m^\nu + h^\nu)\xi^\nu]\rangle\rangle, \end{aligned} \qquad (6.73)$$

which is just Eq. 6.36.

We now proceed to highlight some of the intermediate steps involved in reaching Eq. 6.73. Three non-trivial limiting procedures are involved:

- The matrix $(1-\beta)\mathcal{I} - \beta\mathcal{Q}$ has one eigen-value $1-\beta-(n-1)\beta q$ and $n-1$ eigen-values $1-\beta(1-q)$. Hence,

$$\begin{aligned}
&\lim_{n\to 0}\frac{1}{n}Tr\ln[(1-\beta)\mathcal{I}-\beta\mathcal{Q}] \\
&\quad=\lim_{n\to 0}\frac{1}{n}[(n-1)\ln[1-\beta(1-q)]+\ln[1-\beta-(n-1)\beta q]] \\
&\quad=\ln[1-\beta(1-q)]-\frac{\beta q}{1-\beta(1-q)}.
\end{aligned} \tag{6.74}$$

- With Eqs. 6.72

$$\lim_{n\to 0}\frac{1}{n}\sum_{\rho\neq\sigma} r_{\rho\sigma}q_{\rho\sigma} = -rq.$$

- The 'free-energy' in the replica universe is,

$$\begin{aligned}
&\frac{1}{n}\langle\langle\ln\left(Tr_{S^\rho}\exp(\beta\mathcal{H}_\xi)\right)\rangle\rangle \\
&\quad=\frac{1}{n}\left\langle\left\langle\ln Tr_{S^\rho}\exp\left(\tfrac{1}{2}\alpha\beta^2 r\sum_{\rho\neq\sigma}S^\rho S^\sigma-\beta\sum_{\nu\rho}(m^\nu+h^\nu)\xi^\nu S^\rho\right)\right\rangle\right\rangle \\
&\quad=-\tfrac{1}{2}\alpha\beta^2 r+\frac{1}{n}\left\langle\left\langle\ln\int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}\right.\right. \\
&\left.\left.\exp\left[-\tfrac{1}{2}z^2+n\ln 2\cosh\left(\beta\sqrt{\alpha r}z+\sum_\nu(m^\nu+h^\nu)\xi^\nu\right)\right]\right\rangle\right\rangle \\
&\quad\overset{n\to 0}{\to}-\tfrac{1}{2}\alpha\beta^2 r+\left\langle\left\langle\ln 2\cosh\left(\beta\sqrt{\alpha r}z+\sum_\nu(m^\nu+h^\nu)\xi^\nu\right)\right\rangle\right\rangle.
\end{aligned}$$

Note that in the second line we have been careful about the diagonal term in the replica interaction term. The resulting square has then been linearized by the gaussian transform, leading to a noise term with a gaussian distribution. In the last line, the double brackets stand for the gaussian average over z as well as for the discrete average over the condensed ξ's.

Mean-field equations

Varying f of Eq. 6.73 with respect to m^ν is straight-forward, since the m's enter the cosh only. It leads to

$$m^\nu=\left\langle\left\langle\xi^\nu\tanh\left(\beta\sqrt{\alpha r}z+\sum_\nu(m^\nu+h^\nu)\xi^\nu\right)\right\rangle\right\rangle \tag{6.75}$$

with the double brackets expressing the two types of averages together.

The variation with respect to r leads to

$$\begin{aligned}
&\tfrac{1}{2}\alpha\beta^2(q-1) \\
&\quad=-\frac{\beta\sqrt{\alpha}}{2\sqrt{r}}\int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}} \\
&\qquad\exp\left(-\tfrac{1}{2}z^2\right)z\left\langle\left\langle\tanh\left(\beta\sqrt{\alpha r}z+\sum_\nu(m^\nu+h^\nu)\xi^\nu\right)\right\rangle\right\rangle_\xi,
\end{aligned}$$

where we have separated the two averages temporarily. Integrating by parts, one arrives at Eq. 6.33. Finally, varying f with respect to q one arrives directly at Eq. 6.34.

6.7.3 Marginal storage and palimpsests

A somewhat simplified version of 'learning within bounds' is a model which can be treated analytically[36]. In this model, labelled 'marginalist learning', one confirms all the qualitative results which were mentioned in Section 6.1.4. The model is defined by the synaptic matrix:

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\Lambda\left(\frac{\mu}{N}\right)\xi_i^\mu\xi_j^\mu. \tag{6.76}$$

with

$$\Lambda(x)=\epsilon\exp\left(-\tfrac{1}{2}x\epsilon^2\right).$$

The most recently acquired pattern is $\mu=1$ and the oldest is $\mu=p$. The parameter ϵ plays here the same role as it did in Section 6.6.1. Large ϵ implies dominant weights for the most recent patterns – p=1. Note that this choice of the weight function Λ obeys the normalization

$$\int_0^\infty dt\Lambda^2(t)=1.$$

The order-parameters $\mathbf{m}$ and q are defined just like before, while the definition of r undergoes a modulation by the weights:

$$r \equiv \left\langle\left\langle \sum_{\mu=s+1}^{p=\alpha N} \Lambda^2(\mu/N)\langle m^\mu\rangle^2 \right\rangle\right\rangle . \tag{6.77}$$

Since the couplings remain symmetric, a thermodynamic analysis is possible. The mean-field equations for the order-parameters, with $\mathbf{h}$=0 become:

$$m^\nu = \left\langle\left\langle \xi^\nu \tanh\beta\left[\sqrt{r}\,z + \sum_{\mu=1}^{s}\Lambda(\mu/N)m^\mu\xi^\mu\right]\right\rangle\right\rangle \tag{6.78}$$

$$q = \left\langle\left\langle \tanh^2\beta\left[\sqrt{r}\,z + \sum_{\mu=1}^{s}\Lambda(\mu/N)m^\mu\xi^\mu\right]\right\rangle\right\rangle \tag{6.79}$$

$$r = \int_0^\infty dt\Lambda^2(t)\frac{q}{[1-\beta(1-q)\Lambda(t)]^2}. \tag{6.80}$$

Note that we have absorbed a factor of α in r, which is reflected in the absence of α in the coefficient of the gaussian variable. The notation remains exactly as in Section 6.3.1.

In the zero temperature limit, we focus on the retrieval state with a single non-vanishing overlap $m^\nu = m$, namely the one that had been added $p = \alpha N$ stages before the last one. It will set a lower bound on the retrieval quality of more recent memories. Eq. 6.41 is replaced by

$$m = \mathrm{erf}\left(\frac{m\Lambda(\alpha)}{\sqrt{2r}}\right). \tag{6.81}$$

Eq. 6.42 becomes

$$r = \int_0^\infty dt\frac{\Lambda^2(t)}{[1-C\Lambda(t)]^2} \tag{6.82}$$

and Eq. 6.43 becomes

$$C = \sqrt{\frac{2}{\pi r}}\exp\left(-\frac{\Lambda^2(\alpha)m^2}{2r}\right). \tag{6.83}$$

The single equation 6.44 is a special case of the two coupled equations

$$y\exp(-y^2) = \tfrac{1}{2}\sqrt{\pi}\Lambda(\alpha)C\,\mathrm{erf}(y) \tag{6.84}$$

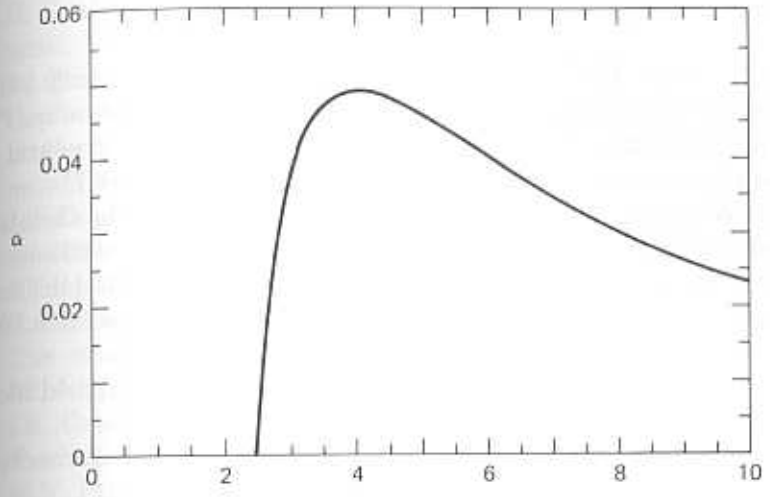


Figure 6.17: Fraction of retrievable patterns vs the parameter ϵ – the weight enhancement of the most recent patterns. (From ref. [7], by permission.)

$$\exp(-2y^2) = \tfrac{1}{2}\pi C^2\int_0^\infty dt\frac{\Lambda^2(t)}{[1-C\Lambda(t)]^2} \tag{6.85}$$

where now

$$y = \frac{m\Lambda(\alpha)}{2r}.$$

Note that to recover the equations of the standard model, one must set

$$\Lambda(t) = \epsilon,$$

for $0 < t < \tau$ and Λ=0 otherwise; the normalization implies $\tau = 1/\epsilon^2$ and counting the number of memorizable patterns one has $\tau = \alpha$.

These two equations have been solved numerically, and the outcome is the dependence of the memorization and storage capacity on the value of ϵ. For $\epsilon < \epsilon_c \approx 2.5$ the equations have no solution with $m \neq 0$ for any α. This is the phase transition mentioned in Section 6.6.1. The optimal value of ϵ is $\bar{\epsilon}$=4.108. For this value of ϵ, the storage capacity is α_c=0.049, which implies that $0.049N$ most recent patterns can be retrieved with less than 1.5% error. The appearance of the phase transition is presented in Figure 6.17.

Bibliography

[1] D.J. Amit, H. Gutfreund and H. Sompolinsky, Storing infinite number of patterns in a spin-glass model of neural networks, *Phys. Rev. Lett.*, **55**, 1530(1985) and Statistical mechanics of neural networks near saturation, *Annals of Physics*, **173**, 30(1987).

[2] G. Weisbuch and F. Fogelman-Soulié, Scaling laws for the attractors of Hopfield networks, *J. Physique Lett.*, **2**, 337(1985).

[3] D.J. Amit, H. Gutfreund and H. Sompolinsky, Information storage in neural networks with low levels of activity, *Phys. Rev.* **A32**, 1007(1987).

[4] E. Gardner, Structure of metastable states in the Hopfield model, *J. Phys.*, **A19**, L1047(1986).

[5] E. Gardner, Maximum storage capacity in neural networks, *J. Phys.*, **18A**, L1047(1986).

[6] B. Derrida, E. Gardner and A. Zippelius, An exactly soluble asymmetric neural network model, *Europhys. Lett.*, **4**, 167(1987).

[7] M. Mezard, J.-P. Nadal and G. Toulouse, Solvable models of working memories, *J. Physique*, **47**, 1457(1986).

[8] P. Baldi and S. Venkatesh, Number of stable points for spin-glasses and neural networks of higher orders, *Phys. Rev. Lett.*, **58**, 913(1987).

[9] E. Gardner, N. Stroud and D.J. Wallace, Training with noise and the storage capacity of correlated patterns in a neural network model, in R. Eckmiller, ed., *Neural Computers: From Computational Neuroscience to Computer Design* (Springer-Verlag, Heidelberg, in press).

[10] B.M. Forrest, Content-addressability and learning in neural networks, *J. Phys.*, **21A**, 245(1988).

[11] J. Von Neumann, *The Computer and the Brain* (Yale University Press, New Haven, 1958).

[12] D.J. Amit, Neural networks – achievements, prospects, difficulties, in W. Guttinger ed. *The Physics of Structure Formation*, (Springer-Verlag, Berlin, 1987).

[13] L. Standing, Learning 10,000 pictures, *Quarterly Journal of Experimental Psychology*, **25**, 207(1973).

[14] J. Buhmann and K. Schulten, Noise-driven association in neural networks, Techniche Univesität München, preprint (1987).

[15] H. Gutfreund, Neural networks with hierarchically correlated patterns, *Phys. Rev.*, **A37**, 570(1988).

[16] A. Crisanti, D.J. Amit and H. Gutfreund, Saturation level of the Hopfield model for neural network, *Europhys. Lett.*, **2**, 337(1986).

[17] E. Gardner and B. Derrida, Optimal storage properties of neural network models, *J. Phys.*, **21**, 271(1988).

[18] J.J. Hopfield, Neural networks and physical systems with emergent selective computational abilities, *Proc. Natl. Acad. Sci. USA*, **79**, 2554(1982).

[19] R.J. McEliece, E.C. Posner, E.R. Rodemich and S.S. Venkatesh, The capacity of the Hopfield associative memory, *IEEE Trans. IT*, **33**, 461(1987).

[20] I.S. Gradsteyn and I.M. Ryzhik, *Tables of Integrals Series and Products* (Academic Press, New York, 1965).

[21] M.V. Feigelman and L.B. Ioffe, The augmented models of associative memory: asymmetric interaction and hierarchy of patterns, *Int. Jour. of Mod. Phys.*, **B1**, 51(1987).

[22] E. Amaldi and S. Nicolis, Stability-Capacity diagram of a neural network with Ising couplings, Rome University preprint no. 642.

[23] E. Gardner, Multiconnected neural network models, *J. Phys.*, **20A**, 3453(1987).

[24] L.F. Abbott and Y. Arian, Storage capacity of generalized networks, *Phys. Rev.*, **A36**, 5091(1988).

[25] D. Horn and M. Usher, Capacities of multiconnected memory models, *J. Phys. France* **49**, 389(1988).

[26] E. Gardner, The phase space of interactions in neural network models, *J. Phys*, **A21**, 257(1988).

[27] W. Krauth, J.-P. Nadal and M. Mezard, The roles of stability and symmetry in the dynamics of neural networks, *J. Phys.*, **21A**, 2995(1988).

[28] T.B. Kepler and L.F. Abbott, Domains of attraction in neural networks, *J. Phys. France*, **49**, 1657(1988).

[29] S.F. Edwards and P.W. Anderson, Theory of spin-glasses, *J. Phys.*, **F5**, 965(1975).

[30] S. Kirkpatrick and D. Sherrington, Infinite-ranged models of spin-glasses, *Phys. Rev.*, **B17**, 4384(1978).

[31] S.F. Edwards and F. Tanaka, The ground state of a spin glass, *J. Phys.*, **F10**, 2471(1980).

[32] D.J. Thouless, P.W. Anderson, and R.G.. Palmer, Solution of 'Solvable model of a spin glass', *Phil. Mag.*, **35**, 593(1977).

[33] M. Mezard, G. Parisi and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[34] D.J. Amit, The properties of models of simple neural networks, in L. van Hemmen and I. Morgenstern eds. *Heidelberg colloquium on Glassy Dynamics* (Springer-Verlag, Heidelberg, 1987).

[35] W. Kinzel, Remanent magnetization of the infinite range Ising spin glass, *Phys. Rev.*, **B33**, 5086(1985).

[36] J.P. Nadal, G. Toulouse, J.P. Changeux and S. Dehaene, Networks of formal neurons and memory palimpsests, *Europhys. Lett.*, **1**, 535(1986).

[37] G. Parisi, A memory which forgets, *J. Phys.*, **A19**, L617(1986).

[38] F.E. Bloom, A. Lazerson and L. Hofstadter, *Brain, Mind and Behavior* (W.H. Freeman, San Francisco, 1985), p. 191.

[39] S. Shinomoto, Memory maintenance in neural networks, *J. Phys.*, **20A**, L1305(1987).

[40] V.J. Emery, Critical properties of many component systems, *Phys. Rev.*, **B11**, 239(1975).

7

Robustness – Getting Closer to Biology

7.1 Synaptic Noise and Synaptic Dilution

7.1.1 Two meanings of robustness

In Section 1.2.1, we listed some of the simplifying assumptions involved in the construction of the models discussed so far. Many more assumptions may have been detected by the reader along the way. No amount of lifting of simplifications will closely approximate the full glory of an assembly of real live neurons. Yet, as the grossest assumptions are replaced by more realistic ones and as the model is modified to account for more complex types of behavior without a significant loss in its basic functional features and in its effectiveness, the model gains in plausibility. To recapitulate our general methodological point of view: The lifting of simplifications is not performed as an end in itself. If the more complicated system functions in a qualitative way that can be captured by the simplified system, then the complication is removed and analysis continues with the simplified system.

We shall recognize two types of robustness, related to two types of results:

1. Robustness of specific properties to perturbation of the underlying parameters.

2. Robustness of general features to modifications required by more complex functions.

Since this chapter will be primarily concerned with robustness of the first kind we start by giving examples of situations of the second kind.

In Chapter 5, we have already encountered robustness of general features such as the existence of attractors (rather quasi-attractors), error correction (content addressability), or the control of the attractors via synaptic construction. These features have been found to be robust when the network was modified to store temporal sequences. Similarly in Section 8.2, we exhibit a network which can perform at low spatial levels of neural activity. As we shall see, lower levels of mean neuronal activity imply correlations between the stored patterns, which opens up the wider issue of correlated patterns to be discussed in the next chapter. Some cases have been studied in detail, and they reveal that the general features mentioned above are robust when the required modifications are introduced.

Next we turn to a list of constraints, of the first kind, whose relaxation will be discussed in the rest of this chapter.

1. *Synaptic noise.* The neatly engineered set of synaptic efficacies, Eq. 4.3, is an idealization. Typically, each synapse will deviate from its 'Hebbian' value by a random amount.

2. *Full connectivity.* Given a well engineered set of synapses, how would the performance of the network be affected by the **random symmetric** elimination of synapses?

3. *Analog depth of synapses.* It may be the case that synapses can hold only a narrow spectrum of values, perhaps even as low as two (just a sign), or three (a sign and an absence of connection).

4. *Symmetry.* Synaptic symmetry is anathema to neurobiologists and a joy for physicists. One can have a well engineered synaptic matrix which is then **randomly and asymmetrically** diluted, by severing synapses.

5. *Synaptic specificity.* This is *Dale's law*, namely the observation that neurons typically emit one kind of neurotransmitter and hence a neuron emits synapses which are all excitatory or all inhibitory. This situation may also be arrived at by **random** modification of the standard model.

6. *Spike rates and neuronal memory.* Bursts – or attractors – when they appear in the cortex, have spike rates which are significantly lower than 500 per second, as would be implied by the standard model. Part of this slowdown may be attributed to the *relative refractory period* which, in turn, implies internal neuronal dynamics with memory.

These various modifications will be studied in the following sections. We will focus mainly on questions of the robustness of detail such as the effects on storage capacity, on retrieval quality, spurious states etc. In some cases, there are analytic results. In others, we will have to resort to simulations. Not all the answers are in and much more work has still to be done.

7.1.2 Noise in synaptic efficacies

Noise enters neural networks in many different forms. Two of them we have already encountered. The first is the *fast noise*, or temperature, which affects the dynamics at every step. The second has been the *slow noise*, imported by the random overlaps among the memorized patterns, which eventually turns the network into a spin-glass. We have seen in the preceding chapters that the network can function rather effectively in the presence of these two types of noise, as long as they remain below certain limits. At times, their presence has even led to improved performance. Here we shall consider the effects of an additional type of noise. It is the random departures of the synaptic efficacies from their standard 'Hebbian' form. In a double *tour de force*, Sompolinsky[1] has analytically computed the properties of such networks and has further shown that from this solution one can arrive at the properties of a whole variety of networks which depart from the standard model.

To be more specific, let us suppose that the synaptic efficacies have the form:

$$T_{ij} = J_{ij} + \eta_{ij} \qquad i \neq j. \tag{7.1}$$

Here J_{ij} is the standard efficacy, Eq. 4.3, which has ensured the satisfactory performance of the standard model. The additional term η_{ij} is a random variable, uncorrelated to any of the stored patterns. It has a gaussian distribution with mean and mean square deviation given by:

$$[\eta_{ij}] = 0,$$
$$[\eta_{ij}^2] = \frac{\eta^2}{N}. \tag{7.2}$$

The square brackets denote averaging over the distribution of the new noise, introduced by the random destruction of synapses. If η is of order unity, then the typical change in the efficacy of a synapse is $O(N^{-1/2})$. This value is seen in the right perspective when one recalls that the usual synaptic efficacy J_{ij} is of order N^{-1} when the number of stored patterns p is finite and becomes itself of order $N^{-1/2}$ only when p becomes proportional to N, namely $p = \alpha N$.

Two contextual comments are in order. Essentially the same model, Eq. 7.1, has been proposed[2] with a rather different purpose in mind. Starting from the opposite limit, in which the synaptic efficacies are purely random, the system is a spin-glass of the SK type[3]. Such a system possesses exponentially many attractors and cannot function as an associative memory. Learning can then proceed by *selection* – strengthening the valleys near persistent external stimuli and in the process pruning many of the other minima. The modifications due to the stimuli, proposed in this approach, consist of addition of 'Hebbian' terms to the purely gaussian synapses.

The second remark is to point out that couplings of the form Eq. 7.1 have been discussed in the original study of the SK-model[3]. In that important article, the spin-glass was discussed in the presence of a net ferromagnetic interaction. This is fully equivalent to the model defined by Eq. 7.1 if the network stores a single pattern. The immediate lesson is that it is the **square** of the random interaction, which itself vanishes on the average, that competes with the ferro-magnetic, ordering force. This explains why in the present case the magnitude of the synaptic noise can be so much larger than the magnitude of the Hopfield term.

The main features of this network, at T=0, are summarized in the following two figures. Figure 7.1 describes the *phase diagram* of the network in a plane of noise amplitude, η, and loading level α $(= p/N)$. The line in the figure represents

$$\alpha_c = \alpha_c(\eta),$$

below which retrieval states exist. Above this line there are only spin-glass states, which are uncorrelated with the stored patterns. This diagram should be compared with the T–α phase diagram of Figure 6.1. It expresses the fact that upon increasing the noise level, the storage capacity decreases. Beyond a critical value of the noise

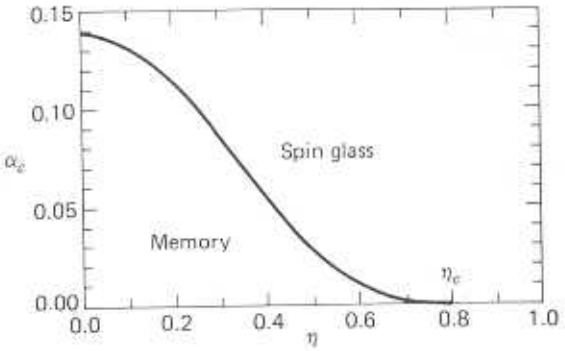


Figure 7.1: Phase diagram of the standard ANN in the presence of synaptic noise. The curve of α_c vs η delimits a region of retrieval from a spin-glass region. (After ref. [1], by permission).

$$\eta_c = \sqrt{\frac{2}{\pi}} \approx 0.8,$$

no retrieval is possible even at finite p. This corresponds to the critical temperature T_c=1, at which $\alpha_c(T)$ vanishes in Figure 6.1.

Figure 7.2 presents the retrieval quality of the network at T=0 – the overlap with a memorized pattern at the attractor – in the presence of noise of amplitude η. The lower curve, m_c, is the overlap of the attractor just below $\alpha_c(\eta)$, which is the highest possible storage for the corresponding value of η. It is, therefore, the lowest retrieval quality for that noise level. For comparison we have the upper curve, denoted by m_0, which is the retrieval quality of the same network at α=0, i.e., at finite p. In the presence of noise, retrieval is not perfect, i.e., $m < 1$, even in this limit. We return to this question in Section 7.5.1.

First let us observe that Figure 7.1 implies that the network is rather robust to noise. To appreciate this robustness, recall that for a given value of η a typical synapse is subjected to a noise level of magnitude $\eta/\sqrt{N}$, while the imprinted memories contribute each a term $1/N$. A network of 10,000 neurons, for example, with an $\eta \approx 0.2$ can tolerate typical noise on each synapse which is **20 times** greater than the contribution of a single memory to the magnitude of the synapse. That network can still effectively store and retrieve better than $0.1N$, or 1000 patterns. Another perspective on the results represented in the two figures is obtained as follows:

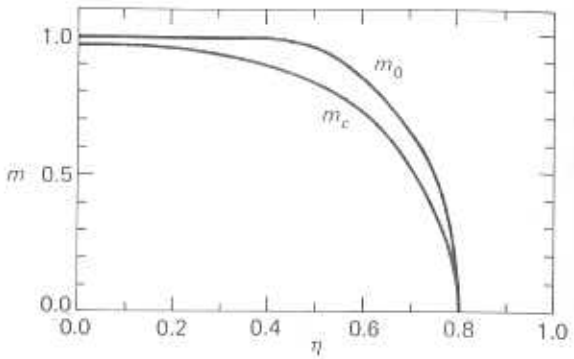


Figure 7.2: Retrieval quality – overlap of attractor with stored pattern – vs synaptic noise level η. The curves marked m_c and m_0 correspond to $\alpha = \alpha_c$ and α=0, respectively. (After ref. [1], by permission.)

Suppose that the biological system can tolerate at most 2.5% retrieval errors, i.e., it requires $m > 0.95$. We then read off Figure 7.2 that the noise level must be kept to $\eta < 0.22$. Proceeding with this value of η to Figure 7.1 one concludes that the fractional storage can be as high as α_c=0.12, compared with $\alpha_c \approx 0.14$ in the perfect network.

Synaptic noise proportional to root α

A particularly useful and interesting case to study is when

$$\eta = \eta_0 \sqrt{\alpha}. \tag{7.3}$$

In this case, the noise is proportional to $1/N$ if p is finite and becomes of order $1/\sqrt{N}$ only when $p = \alpha N$ with finite α. The results for this case can be read off the two figures 7.1 and 7.2, by a redefinition of the noise variable. The consequences are rather surprising. They are represented in Figure 7.3, where the variable η_0 was replaced by $\eta_0^2/(1+\eta_0^2)$ for convenience.

What one observes is that only for very large values of η_0 does the retrieval quality begin to decrease significantly from its value in the absence of noise, η_0=0. Only for $\eta_0 \to \infty$ do $m_c(\eta_0)$ and $\alpha_c(\eta_0)$ go to zero. This can be readily understood from the previous two figures in which we read that both $m_c(\eta_0)$ and $\alpha_c(\eta_0)$ vanish at a finite value of η. But when α vanishes at a finite η, Eq. 7.3 implies that $\eta_0 = \infty$.

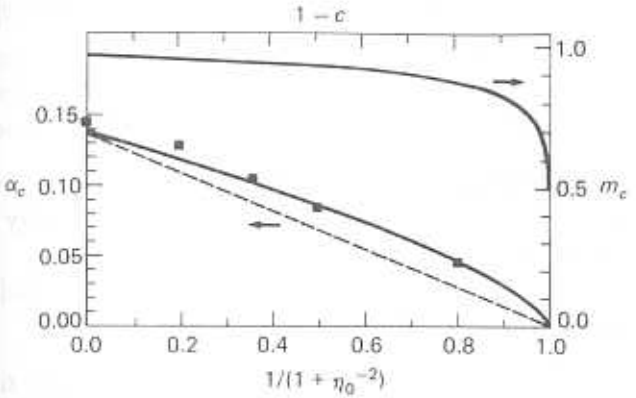


Figure 7.3: Storage capacity (α_c) and retrieval quality (m_c) as functions of the amplitude η_0 of noise proportional to $\sqrt{\alpha}$ (bottom abcissa) or dilution fraction (top abcissa). The dashed line is signal-to-noise estimate. Squares are simulation results. (After ref. [1], by permission.)

For $\alpha \to 0$, perfect retrieval is recovered, because both sources of noise – from the random overlaps between patterns and from the random variables added to the synapses – are put out simultaneously.

The storage capacity follows quite closely the straight line

$$\alpha_c(\eta_0) = \frac{\alpha_c(0)}{1+\eta_0^2}.$$

This is just the reduction of the storage capacity that would have resulted from elementary signal-to-noise considerations like those of Section 4.3.1. In fact, one finds from an extension of such a calculation to the present case that to ensure perfect retrieval of a random pattern the storage is limited by:

$$\alpha < \frac{1}{2(1+\eta_0^2)\ln N}, \tag{7.4}$$

while to ensure perfect retrieval of p patterns, the condition is:

$$\alpha < \frac{1}{4(1+\eta_0^2)\ln N}. \tag{7.5}$$

The same results can be derived long hand, by performing a mean-field

calculation of the free-energy.[1] One result of such analysis[1] is that for small values of α the relative number of errors is

$$\frac{N_{err}}{N} = \tfrac{1}{2}(1-m) \approx \exp\left(-\frac{1}{2\alpha(1+\eta_0^2)}\right). \tag{7.6}$$

This is the analog of Eq. 6.18. From it one derives, as in Section 6.2.1, the two bounds, Eqs. 7.4 and 7.5, on the storage capacity if perfect retrieval is required.

7.1.3 Random symmetric dilution of synapses

The first application of the above results is to the study of the performance of a standard network with randomly disconnected synapses. Starting with a network with synapses of the form Eq. 4.3, one zeroes synaptic efficacies symmetrically, until the average number of non-zero connections per neuron becomes Nc. The process keeps the synaptic matrix symmetric. We write:

$$T_{ij} = \frac{c_{ij}}{Nc}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu \qquad i \neq j, \qquad T_{ii} = 0. \tag{7.7}$$

The dilution coefficients $c_{ij}(=c_{ji})$ take the values 1 and 0 with probabilities c and $1-c$, respectively. Eq. 7.7 implies that on the average each neuron will remain connected to Nc others. Here we will restrict ourselves to the case where Nc is still of order N – the network remains densely connected. Note also that a factor c was introduced in the normalization of the synaptic matrix in order to have a fixed size for the PSP on a neuron when the network is in a pattern.

It is at this stage that Sompolinsky's idea[1] manifests its power. The synaptic matrix Eq. 7.7 can be rewritten as a fully connected Hopfield model with synaptic noise. One can write

$$T_{ij} = J_{ij} + \delta J_{ij} \tag{7.8}$$

where J_{ij} is the usual synaptic matrix Eq. 4.3, which is also Eq. 7.7 when c=1. Clearly,

$$\delta J_{ij} = \left(1 - \frac{c_{ij}}{c}\right) J_{ij} \tag{7.9}$$

[1] Recall that the noise has not affected the symmetry of the synapses, and thermodynamic analysis, along the lines of Chapter 6, is legitimate.

is a fluctuating random variable added to each synapse. For large N, its distribution is gaussian[1]. Its mean is zero, since the two factors in this expression are independent and each has zero mean. The variance of this noise, its mean square, can be computed directly to be

$$[\delta_{ij}^2] = \frac{\alpha(1-c)}{Nc} \tag{7.10}$$

and this has just the form of Eq. 7.3, with

$$\eta_0 = \sqrt{\frac{1-c}{c}}. \tag{7.11}$$

The properties of a randomly, symmetrically, diluted network, which remains densely connected, can be now simply read off Figure 7.3 by expressing the variable $1/(1+\eta_0^{-2})$ in terms of the dilution parameter $1-c$. From Eq. 7.11 one has

$$1/(1+\eta_0^{-2}) = 1-c.$$

This is expressed by providing Figure 7.3 with a second abscissa, for $1-c$, at its top. Thus one has a full solution of the symmetric dilution problem.[2] The theoretical predictions of this solution have been tested by numerical simulations[1] on networks with 500 and 1,500 neurons. The results are presented in the figure as full squares. The slight deviations can be attributed to the small effects of *replica symmetry breaking*.

A few concluding remarks: The dashed line in the new context is the naive estimate:

$$\alpha_c(c) = c\alpha_c(c=1).$$

The actual capacity of the diluted network is somewhat higher. The quality of retrieval is very mildly impaired by symmetric dilution. All patterns are retrieved with an overlap higher than 90% for dilution level of up to 0.8, when each neuron remains connected to 20% of the other neurons in the network. At 40% dilution the maximal number of errors, below saturation, is 2.5%. The storage capacity, α_c, goes down from 0.14 to 0.1.

[2] Within the approximation of replica symmetry.

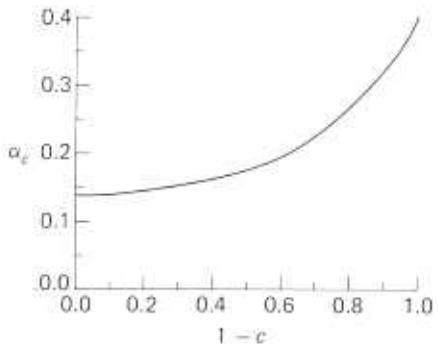


Figure 7.4: Storage capacity relative to the number of surviving synapses per neuron (α_c/c) vs the level of dilution ($1-c$).

It is natural to redefine the storage capacity as the ratio of the maximal number of patterns to the average connectivity of the network, i.e.,

$$\bar{\alpha} \equiv \frac{p}{Nc} = \frac{\alpha_c}{c}. \tag{7.12}$$

This storage capacity is plotted in Figure 7.4, which is again nothing but a redrawing of Figure 7.3. The storage capacity relative to the mean connectivity of the diluted network is higher than the storage capacity of the fully connected network. This effect becomes more pronounced when dilution becomes more extreme. See e.g., Section 7.3.4, below.

One can therefore safely conclude that dilution of a well organized network, if it is random, leaves a network of essentially undiminished performance up until rather high dilution levels. This has been verified for symmetric dilution that leaves the network densely connected. In practical terms, it implies that if a device is constructed along the lines of the standard model then the production of the synapses can be of rather low quality and the network will perform effectively. Some of the technical details involved in deriving these results can be found in Appendix 7.5.1.

7.2 Non-Linear Synapses & Limited Analog Depth

7.2.1 Place and role of non-linear synapses

The third item on our list in Section 7.1.1 has been the question of the *analog depth* of the synapses, namely the number of possible values that a biological or artificial synapse can actually maintain. The standard model, in its domain of retrieval, is based on synapses which can discriminate between $2p+1 = O(N)$ different values. This is not very likely to be a property of a biological synapse, though the biological constraint is not very well known. On the other hand, early prototypes of artificial hardware neural networks, to be discussed at some length in Chapter 10, indicate that such networks should and can operate with the extreme limitation that a synapse will be able to carry two or three different values.

But for the network to actually store and retrieve information, the distribution of the synaptic values must be programmed, or learned. In the present context, the most natural way of organizing the synaptic matrix would be to start from the familiar, generic standard model and to *clip* the synapses by preserving only their signs. In other words, the clipped synaptic matrix will be

$$T_{ij} = \mathrm{sign}(J_{ij})$$

with J_{ij} given, as usual, by Eq. 4.3. This is a rather extreme case of clipping. After all, the standard synapses storing p patterns have a distribution of magnitudes distributed around $\sqrt{p}$ and it seems that a more faithful clipping procedure would be to give the value zero to those synapses whose magnitude is smaller than some low cutoff in the standard model and ± 1 to the synapses of magnitude larger than the cutoff. We shall discuss both cases below.

We start with the normalization of the clipped T_{ij}'s. In the standard J_{ij}, if p is large enough, then each synaptic element is typically of magnitude $\sqrt{p}$. The overall magnitude of the synaptic matrix is of no concern if the system is noiseless. But it is just this scale which determines the range of noise in which retrieval is possible. In order to keep that range fixed as p varies, the appropriate scaling of the couplings will be:

$$T_{ij} = \frac{\sqrt{p}}{N}\mathrm{sign}(J_{ij}) \tag{7.13}$$

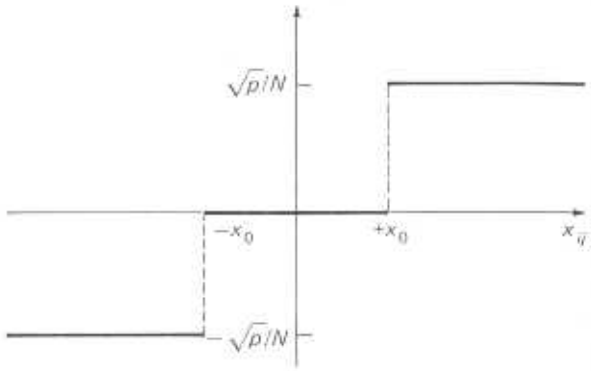


Figure 7.5: The functional form of the efficacy of the 3-state synapse vs. the corresponding value of the standard efficacy.

which sets the magnitude of each synapse equal to that of the typical synapse in the standard model.

The extension of 7.13 to the case of a 3-state synapse can be formulated as follows:

$$T_{ij} = \frac{\sqrt{p}}{N} F(x_{ij}) \tag{7.14}$$

with

$$x_{ij} = \frac{1}{\sqrt{p}} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{7.15}$$

which is a variable of typical magnitude 1, and F is

$$F(x) = \left\{ \begin{array}{ll} 0 & \text{if } |x| < x_0 \\ \text{sign}(x) & \text{if } |x| > x_0 \end{array} \right\}. \tag{7.16}$$

The form of F is sketched in Figure 7.5. The limit x_0=0 recovers the simple clipping of Eq. 7.13.

Both versions of clipping, 7.13 and 7.14, are special cases of a synaptic matrix which has a general non-linear function for F. It is another remarkable result of ref. [1] that this general case of non-linear storage prescription can be reduced again to one of standard synapses with noise.

7.2.2 Properties of networks with clipped synapses

2-state clipping

- A network with two-state clipped synapses is equivalent to a standard network with noise of variance η_0^2=0.57. Inspection of Figure 7.3 for this value of η_0 gives

$$\alpha_c \approx 0.1, \qquad m_c \approx 0.95.$$

In other words, this extreme type of reduction in the *analog depth* of the synapses has a rather mild effect again. The storage capacity is reduced by about 30%. The retrieval quality at saturation involves 2.5% errors compared to the 1.5% errors of the intact network.

- The capacity for retrieval without errors can be read from Eq. 7.4. This capacity is decreased by a factor of $2/\pi$ (See also ref. [4]) and becomes

$$\alpha < \frac{1}{\pi \ln N}.$$

It is amusing to note that the reduction of the absolute storage capacity of the network is also rather close to $2/\pi$.

A threefold equivalence has been established between

- The theoretical, replica-symmetric, results.
- The dilute system with $c = 2/\pi$.
- The clipped network.

All three are compared in Figure 7.6. Most remarkable is the accord between the simulations of the clipped and diluted networks for such detailed quantities as retrieval quality. The agreements with the theoretical results is fine until the theoretical saturation. Beyond this point the actual network is much affected by finite size effects, which are significant even at N=1,000. Compare Section 6.5.2. Yet, it is clear that beyond α=0.1 the retrieval quality decreases quite rapidly with α.

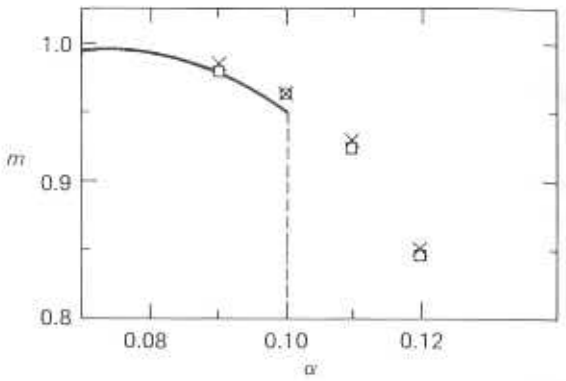


Figure 7.6: Three-way comparison of retrieval quality vs storage level α. Curve is the replica-symmetric mean-field theory; $\times$ and $\square$ are simulation results for clipped and diluted ($c = 2/\pi$) networks, respectively, with 1,000 neurons. (After ref. [1], by permission.)

Properties of networks with three-way synapses

Another, richer, application of the equivalence of non-linearity in the storage prescription to noise is the case of 3-state clipping introduced in Eq. 7.16, above. It has been studied also in ref. [5]. This version is a combination of dilution and clipping. Synapses whose standard value is less than x_0 are cut off and hence as x_0 increases, the connectivity of the network becomes more dilute. On the other hand, those synapses which remain, are clipped.

- The remaining connectivity is related to the cutoff parameter x_0 of Eq. 7.16 by

$$c_0 = 1 - \mathrm{erf}\left(\frac{x_0}{\sqrt{2}}\right). \tag{7.17}$$

- The network is equivalent to a standard network with a relative noise parameter:

$$\eta_0^2 = \tfrac{1}{2}\pi c_0 e^{x_0^2} - 1, \tag{7.18}$$

which is a non-monotonic function of the cutoff x_0. This is a result of the fact that as x_0 increases the concentration of surviving bonds, c_0, decreases while the exponential factor multiplying it increases. As a consequence, the equivalent noise increases at low values of x_0 and then

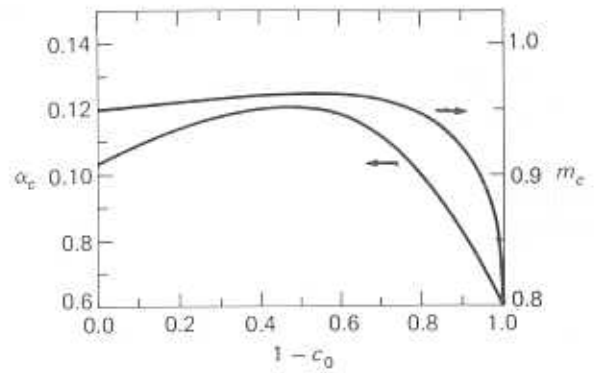


Figure 7.7: Storage capacity and retrieval quality vs dilution level resulting from 3-state clipping. (After ref. [1], by permission.)

starts to decrease. It leads to a corresponding non-monotonic behavior of the network performance as a function of x_0, which is shown in Figure 7.7.

In this figure the storage capacity and the retrieval quality are plotted vs the level of dilution, which according to Eq. 7.17 is a monotonic function of x_0. One observes that

- there is an optimal clipping cutoff, or dilution. For $x_0 \approx 0.62$, where $c_0 \approx 0.63$, one finds

$$\alpha_c \approx 0.12, \qquad m_c \approx 0.96.$$

In other words, eliminating 37% of the bonds, albeit the weak ones, and clipping all the other synaptic efficacies leads to a mere decrease of storage capacity from 0.138 to 0.12 and in retrieval quality from 0.97 to 0.96.

7.2.3 Non-linear storage and the noisy equivalent

High storage level

For the network to function as an associative memory, the function F has to be subjected to a few constraints. These constraints, the details of which will be sketched below, can be summarized as follows:

- To make possible the stability of the patterns we must have

$$J \equiv \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} \frac{dF(x)}{dx} \equiv \left\langle\left\langle \frac{dF(x)}{dx} \right\rangle\right\rangle > 0. \tag{7.19}$$

- To have equal overall weight for excitatory and inhibitory synapses one should have:

$$\langle\langle F(x) \rangle\rangle = 0. \tag{7.20}$$

- The variance of F must be finite, namely

$$\bar{J}^2 \equiv \langle\langle F^2(x) \rangle\rangle < \infty. \tag{7.21}$$

Note that the standard model has $F(x) = x$, which gives J=1 and $\bar{J}^2$=1. If the standard synaptic matrix is scaled by a factor J_H, then $J = J_H$ and $\bar{J}^2 = J_H^2$.

The major technical result[1] is that the network with the non-linear storage prescription is equivalent to a standard Hopfield (linear) network with a synaptic scale factor $J_H = J$, of Eq. 7.19, with a gaussian noise of zero mean and a variance η^2/N given by:

$$\eta^2 = \alpha(\bar{J}^2 - J^2). \tag{7.22}$$

It has the form Eq. 7.3 with one proviso: When the amplitude of the standard part of the model is not unity but J_H, then η is replaced by

$$\zeta \equiv \frac{\eta}{J_H} \tag{7.23}$$

as the relevant parameter. This is intuitively clear since it is the relative strength of the noise term to the 'Hebbian' component which counts. Hence, the solution of the two clipped networks, at high loading, is just a matter of transcribing Figure 7.3. This we do in the next section.

Equivalence of 2-state clipped synapses

The application of the results of Section 7.2.2 to the case of 2-state clipping of synaptic efficacies, as described by Eq. 7.13 proceeds as follows:

Using the functional form 7.13, one computes the two parameters J and $\bar{J}$ via Eqs. 7.19 and 7.21. Now, for $F(x) = \text{sign}(x)$, the derivative of F is $2\delta(x)$. Hence Eq. 7.19 gives

$$J = \sqrt{\frac{2}{\pi}}.$$

Similarly, the square of the sign-function is unity. Hence, the normalization of the gaussian distribution leads, via Eq. 7.21, to

$$\bar{J}^2 = 1.$$

Using Eqs. 7.22 and 7.23, one arrives at the second parameter of the equivalent noisy model, namely

$$\eta^2 = \alpha\left(1 - \frac{2}{\pi}\right),$$

from which one concludes that

$$\zeta_0^2 \equiv \frac{\zeta^2}{\alpha} = \frac{\bar{J}^2 - J^2}{J^2} = \tfrac{1}{2}\pi - 1.$$

The clipped network is therefore equivalent to a standard network with noise of relative strength $\eta_0^2 \approx 0.57$.

Equivalence of three-state synapses

For p large, the values of the variable x in F, corresponding to the standard synaptic efficacies (see e.g., Eq. 7.15), are those of a random walk with steps $\pm 1/\sqrt{p}$. It is a random variable of gaussian distribution with zero mean and unit variance. It follows that the fraction of non-vanishing synapses compared to the fully connected network is[1]:

$$c_0 = 2\int_{x_0}^{\infty} \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} = 1 - \text{erf}\left(\frac{x_0}{\sqrt{2}}\right). \tag{7.24}$$

But this is not random dilution! It is correlated with the way information about the patterns is stored in the synaptic efficacies of the standard model. Consequently, one proceeds along the lines of the non-linear prescription, rather than those of random dilution.

The first step again is to evaluate the two parameters J and $\bar{J}$. This time

$$\frac{dF(x)}{dx} = \delta(x + x_0) + \delta(x - x_0)$$

and hence

$$J = \sqrt{\frac{2}{\pi}} \exp\left(-\tfrac{1}{2}x_0^2\right).$$

On the other hand,

$$\bar{J}^2 = \int_{-\infty}^{-x_0} \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} + \int_{x_0}^{\infty} \frac{dx}{\sqrt{2\pi}} e^{-x^2/2} = c_0.$$

Using Eq. 7.23 one has[1]

$$\zeta_0^2 = \frac{\eta^2}{\alpha} = \frac{\pi}{2} c_0 e^{x_0^2} - 1.$$

7.2.4 Clipping at low storage level

The performance of the system at low storage level – finite p – is very robust. A general procedure for analyzing this problem has been outlined in ref. [1]. Here we will restrict ourselves to a few representative examples, using direct considerations of stability of patterns, similar to those described in Section 4.3.1.

Suppose that the synaptic efficacies are given by a clipped standard version, Eq. 7.13. The relative sign of the local field (PSP) on neuron i and the state of this neuron, when the network is in the state corresponding to pattern number 1, i.e., $S_i = \xi_i^1$, is given by the sign of their product

$$h_i S_i = \frac{\sqrt{p}}{N} \sum_{j=1}^{N} \xi_i^1 \text{sign}\left(\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu\right) \xi_j^1.$$

The ± 1's represented by ξ_i^1 and ξ_j^1 can be inserted into the sign-function and the term with μ=1 separated from the other patterns to give:

$$h_i S_i = \frac{\sqrt{p}}{N} \sum_{j=1}^{N} \text{sign}\left(1 + \sum_{\mu=2}^{p} \xi_i^1 \xi_i^\mu \xi_j^\mu \xi_j^1\right). \qquad (7.25)$$

The basic reason why the right hand side remains positive despite the clipping is that for finite p we have *self-averaging* again. See e.g.,

Section 4.4.2. In other words, the sum over j in Eq. 7.25 makes the sum over μ inside the sign-function sample all its possible values. But since $\mu \neq 1$ and $i \neq j$ this term takes positive and negative values with equal probability. Yet, there is a bias of 1 added to this term and hence in the sum over j there will be an excess of positive contributions over the negative ones. This is enough to ensure that the state $\{\xi^1\}$ is stable.

As a specific example, consider the case p=3. We have

$$h_i S_i = \frac{\sqrt{3}}{N} \sum_{j=1}^{N} \text{sign}\left(1 + \xi_i^1 \xi_i^2 \xi_j^2 \xi_j^1 + \xi_i^1 \xi_i^3 \xi_j^3 \xi_j^1\right).$$

The sum of the last two terms inside the *sign* can take the values $-2, 2$ and 0. Each of the first two values have a probability of $\frac{1}{4}$. The last one has a probability of $\frac{1}{2}$. The first two values overwhelm the bias of 1, and the corresponding contributions to the sum over j cancel each other. For every site for which these two terms add to 0, the *sign* is $+1$ and this happens for 50% of the sites. Consequently,

$$h_i S_i = \frac{\sqrt{3}}{2},$$

which, while somewhat smaller than the value of 1 for the corresponding quantity in the standard network, provides ample stability for each of the memorized patterns.

7.3 Random vs. Functional Synaptic Asymmetry

7.3.1 Random asymmetry and performance quality

The symmetry of the synaptic connection matrix has been a central pillar of the discussion of the preceding chapters. At the same time, it has been a most blatantly offending constraint from the biologist's point of view. Symmetry has been essential for the existence of a landscape picture for the dynamics of the network and asymmetry excludes such a landscape and risks undermining the basic tenets of the ANN approach. Yet, already in his initial article on neural networks, Hopfield[6] raises the question and foreshadows the full answer:

> Why should stable limit points or regions persist when $T_{ij} \neq T_{ji}$? If the algorithm at some time changes V_i from 0 to 1 or

> vice versa, the change in energy...can be split into two terms, one of which is always negative. The second is identical if T_{ij} is symmetric and is "stochastic" with mean 0 if T_{ij} and T_{ji} are randomly chosen. The algorithm for $T_{ij} \neq T_{ji}$ therefore changes E in a fashion similar to the way E would change in time for a symmetric T_{ij} but with an algorithm corresponding to a finite temperature.

The splitting of the energy change referred to above is nothing but our Eq. 3.60, and the algorithm is what we have been calling the dynamical process, defined by Eq. 2.6.

What is suggested in the above quote is that the zero temperature dynamics of a neural network, in which an anti-symmetric part uncorrelated with the symmetric part is added to the synapses, will be affected by *fast* noise, i.e., noise which introduces a stochastic element at every step of the dynamics. This is an effect akin to temperature. One should therefore expect that if the noise level is not too high, most of the relevant features of the well organized ANN will persist.

To arrive at reliable quantitative estimates of the size of the effect and of the critical level of tolerable asymmetry is a very difficult analytic task. In the absence of an energy function, physics has to put aside most of its arsenal and to start developing new tools, based on dynamical equations alone. Much progress is being made in this new direction[7,8,9,10] and this is one clear domain in which the attention focussed on neural networks will enrich the lore of physics. Some of the theoretical developments will be briefly described below. But first, we discuss some directly pertinent data obtained by simulation.

The dynamics of a network of 1,000 neurons was simulated with the intention of testing the stability of the memorized patterns under *asymmetric dilution* of synaptic contacts. Initially, the synaptic matrix was set up in the standard, symmetric way, Eq. 4.3, storing αN random patterns. A fraction γ of the $\frac{1}{2}N^2$ synapses were set to zero. This was done by selecting ij at random and setting to zero one out of J_{ij} and J_{ji}. Then, the network was run, at T=0, fixed dilution level γ and fixed storage level α, starting from 200 stored patterns as initial states. The statistics of the resulting runs are presented in Figs. 7.8 and 7.9.

Each of the two figures presents the statistics of four different trajectories. The curves marked (a)-(c) all represent trajectories which have terminated in attractors – (a) with mean retrieval quality $m > 0.98$;

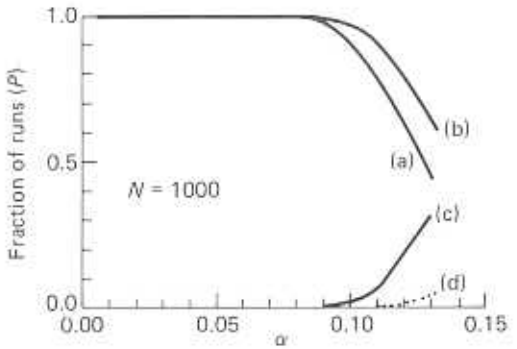


Figure 7.8: Statistics of dynamical trajectories starting from memorized patterns vs storage level α. (a) average retrieval quality $m > 0.98$; (b) $m > 0.9$, (c) low m attractors (spin-glass), (d) no attractor. Dilution level is 50%.

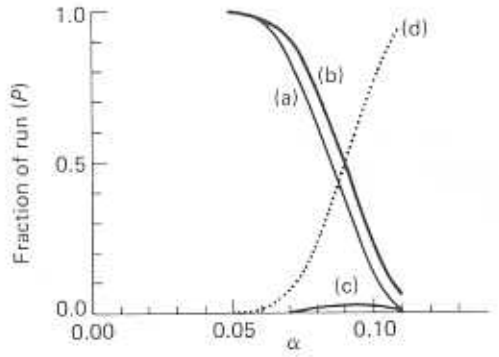


Figure 7.9: Statistics of dynamical trajectories starting from memorized patterns vs storage level α. (a) average retrieval quality $m > 0.98$; (b) $m > 0.9$, (c) low m attractors (spin-glass), (d) no attractor. Dilution level is 100%.

(b) with $m > 0.9$ and (c) attractors with low m, i.e., spin-glass variety. Curves marked (d) count trajectories which have not terminated at a fixed point within 100 cycles. Each point used in generating these curves represents the fraction of 200 runs which have terminated in one of these four fashions. Note that the dilution levels quoted represent the removal of one half of the corresponding fraction of synapses from the total store. Thus 100% dilution in Figure 7.9 implies that one half of the synapses have been removed randomly, leaving every pair of neurons connected in one direction.

To summarize the findings:

- At 50% dilution, retrieval of the memorized patterns with a very small fraction of errors is possible until about α=0.09. This is close to 65% of the storage capacity of the intact network.
- At this level of dilution, the spin-glass states are reached only above the corresponding α_c. Runaway trajectories are observed only well above α_c – for $\alpha > 0.11$ – and in very small numbers.
- At 100% dilution, which is a rather extreme case, effective retrieval is found for $\alpha < 0.06$.
- Spin-glass states show up very weakly and only above α=0.08. This is probably related to the weakening of the spin-glass phase due to the asymmetry[7,11,8,9,10].
- At α_c, wandering trajectories grow rapidly in numbers and soon dominate the scene.

The simulations cannot discriminate between chaotic trajectories and trajectories which take exceedingly long times to reach attractors. Such trajectories are in fact predicted to exist in an asymmetric network[10].

7.3.2 Asymmetry, noise and spin-glass suppression

Hertz et al[7] have suggested that asymmetry has a positive effect of destabilizing the spin-glass states while effecting only mildly the retrieval states of a network. They conjectured that if spin-glass states are destabilized, only retrieval states stand to gain, since their basins of attraction should increase. In particular, since at moderately high storage the spin-glass assimilates **all** spurious states (see e.g., Section 6.3.2), the conclusion might have been extended to read that in the presence of a fair amount of asymmetry, the basins of attraction of the retrieval states span most of the space of states of the network.

At the same time Parisi[11] has proposed that asymmetry may serve another beneficial role. He pointed out that an attempt to implement a process of learning in symmetric ANN's would encounter difficulties, because every stimulus will quickly run into an attractor, either a retrieval state or one of the multitude of attractors of the spin-glass phase which always accompanies the retrieval states (see e.g., Section 6.3.2).

Consequently every stimulus will be perceived as a familiar pattern. This will be the case even if that stimulus has been presented very briefly. The suggestion is that asymmetry may be functional in converting the spin-glass states into *chaotic* trajectories. If this were to be true, then a brief presentation of a stimulus far removed from a memory would lead to a trajectory of temporally uncorrelated states. If the Hebb rule is interpreted as saying that:

- a synapse is modified according by the temporal average of the correlated activity of the two neurons it connects,

then an unfamiliar pattern presented briefly will embark on a chaotic trajectory, produce zero average correlated activity of pairs of neurons and will not be learned. In contrast, a new pattern presented persistently will produce a non-vanishing average temporal activity and hence will be learned. See also Section 9.3.4.

Early studies have revealed, either by simulations[11] or by analysis of closely related models[7], that spin-glass freezing is destroyed. This has been made systematic[8], by extending the analysis to fully time-dependent processes. The conclusions are:

- Asymmetry can be formally expressed as another source of noise.
- For small α, the retrieval states are unaffected even at 100% asymmetric dilution.
- At 100% asymmetric dilution spin-glass effects are totally absent.

Two questions remain open:

- What is the fate of the suggestion that spin-glass states disappear at all levels of asymmetry[7]?
- What becomes of the destabilized spin-glass states?

These questions have been taken up in a further extension of the analytic tools[8,10]. But the context has been shifted to the pure spin-glass. The early results about the disappearance of spin-glass effects at all asymmetry levels[7] have been confirmed. But this is not the whole story. At finite noise levels – finite T – the system is ergodic, as it is at T=0 when the asymmetry is complete, in accordance with simulations[11]. If the asymmetry is intermediate and T=0, the spin-glass effects do indeed disappear but exponentially many attractors –

states stable against single spin-flips – survive. These attractors are of a special type, because the time needed by the network to reach these states increases exponentially with the number of neurons in the network. The persistence of such attractors does not, therefore, preclude a learning scenario like that proposed by Parisi.

7.3.3 Neuronal specificity of synapses - Dale's law

It is widely believed that neurons are either *excitatory* or *inhibitory*. What is implied is that a given type of neuron can release one type of neuro-transmitter and hence can produce a single type of *efferent* synapse, either excitatory or inhibitory, which is referred to as Dale's law[12]. Clearly, a network obeying this law cannot have a symmetric synaptic connectivity matrix – neurons must receive inputs of both synaptic signs and send outputs of a single sign.

This problem has been addressed[13] within the same methodology that has characterized the approach to dilution and clipping. The network is set up as a standard model with synapses storing a set of random patterns according to Eq. 4.3. Then, neuronal specificity is realized by a process of dilution, which is itself partially random. A given excess ratio of excitatory to inhibitory neurons,

$$q \equiv \frac{N_E - N_I}{N}$$

is chosen. Then

$$N_E = \frac{1+q}{2} N$$

neurons are selected at random and the inhibitory synapses issuing out of those neurons are set to zero. Similarly, the excitatory synapses on the remaining $N_I = N - N_E$, inhibitory, neurons are set to zero.

In this process, about one half of the synapses are severed. To compensate for this dilution and preserve the magnitude of the local fields (PSP's), the magnitude of the remaining synapses is doubled. The process can be specified by choosing N numbers $\zeta_j = \pm 1$, which are chosen with the probability distribution:

$$\Pr(\zeta_j = \pm 1) = \frac{1 \pm q}{2}. \tag{7.26}$$

Then, the new synaptic matrix is written as

$$T_{ij} = 2 J_{ij} \theta(\zeta_j J_{ij}). \tag{7.27}$$

The resulting network has been analyzed for low storage levels. In that case, one can perform in a straight-forward way a *signal-to-noise* analysis. One finds[13] that the factor of 2 in the expression for T_{ij} ensures that when the network is in one of the stored patterns the signal term is exactly 1. But the same factor increases the typical magnitude of the noise term by a factor $\sqrt{2}$. The ratio of the noise to the signal is therefore $\sqrt{2p/N}$. The patterns remain robust attractors, even if there may be some decrease in the size of their basins of attraction.

A by-product of this procedure for the neuronal specificity of synapses is the appearance of a ferromagnetic attractor which is interpreted as classifying unrecognized patterns: Suppose that the network has excess excitatory neurons, i.e., $q > 0$. In that case, the two ferromagnetic states, (one with all S_i=1, i.e., all neurons active, and one with all S_i=-1, i.e., all neurons quiescent) are stable states of the network, over and above the stored patterns. This follows from the computation of the local field on any neuron in that state, given the synaptic matrix Eq. 7.27. It is

$$h_i = 2 \sum_{j=1}^{N} J_{ij} \theta(\zeta_j J_{ij}) = \left[2 \sum_{j=1}^{N} J_{ij} \theta(\zeta J_{ij}) \right] \tag{7.28}$$

where the square brackets are an average over the random variable ζ. The value of h_i can be computed explicitly. It is the same type of calculation which has led to the value of m_n in Section 4.5. The result for a network storing p patterns is

$$h_i = \langle\langle |z_p| \rangle\rangle$$

where z_p is defined in Eq. 4.53 and has its own probability distribution given by Eq. 4.54. Whatever this precise value is, one can conclude that h_i is positive, if $q > 0$, and of order one. The fluctuations are negligible if p is kept finite and $N \to \infty$. The ferromagnetic states are therefore stable.

For our present purposes, we find it rewarding that the retrieval performance of the network is largely preserved by the introduction of neuronal specificity of synapses. There is, though, no systematic study, either analytic or numerical, of the quantitative effects of this modification of the synaptic matrix on the performance. In ref. [13] it is suggested that the appearance of 'uniform' attractors may provide

a candidate for a cognitive classification of non-recognition. But much more work is required before the proposal can be evaluated. One would have to know more about:

- The presence (or absence) of additional spurious states.
- The effect of asynchronous dynamics on the results of the rather fragile synchronous dynamics used in ref. [13].
- The effect of fast noise.
- The effect of correlations between patterns at higher storage levels.

7.3.4 Extreme asymmetric dilution

An important recent achievement in the study of less ideal ANN's has been the exact analytic solution of the **dynamics** of a standard ANN whose connections have been randomly diluted to a level where the mean number of remaining synapses per neuron is smaller than $\ln N$[9]. The significance is not that it presents a state of affairs closer to biological reality than the standard model. Quite the contrary, it is probably less so. After all, while cortical connectivity is certainly not symmetrical and even in small columnar regions as those we have chosen to study connectivity is not complete, it remains true that average connectivity is a few thousand per neuron. For a network to have $\ln N > 1,000$, N must be a super-astronomical number. In comparison to such numbers, even the total number of neurons in the cortex is dwarfed.

The significance of the Derrida-Gardner-Zippelius[9] study is twofold:

- It is an almost unique case in which the full dynamical development of the overlaps is unravelled.
- Its conclusions support the growing feeling that the retrieval properties of an ANN, with a standard storage prescription, will sustain any degradation that is not intelligently intended to damage its retrieval properties.

The definition of the framework involves first of all a prescription for the dilution. One starts from a standard synaptic matrix and the dilution is a simple variation on Eq. 7.7, namely

$$T_{ij} = c_{ij} N J_{ij} = c_{ij} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu . \tag{7.29}$$

The dilution matrix is chosen with the distribution

$$\Pr(c_{ij}) = \frac{c}{N}\delta(c_{ij} - 1) + (1 - \frac{c}{N})\delta(c_{ij}). \tag{7.30}$$

This is a random dilution which gives the probability c/N for a synaptic efficacy to remain intact and $(1 - c/N)$ for it be set to zero. Note that c is the mean connectivity per neuron. In contrast to the discussion in Section 7.1.3, here $c/N \to 0$ as $N \to \infty$. In other words, the network is heavily diluted. Specifically, the analysis is carried out under the restriction that

$$c \ll \ln N. \tag{7.31}$$

In this context, it is natural to redefine the storage capacity, as we have done in connection with Figure 7.4, relative to the number of connections per neuron, namely

$$\bar{\alpha} \equiv \frac{p}{c}.$$

What renders the model soluble, for arbitrary loading level p, are the following features:

- The restriction on the connectivity, Eq. 7.31, eliminates **almost** all feedback loops. In other words, the network can be organized into an effective set of trees. See e.g., Figure 7.10.
- To the extent that feedback loops exist, namely cycles of neurons which are connected by synapses, they are typically large as N becomes large and serve effectively as boundary conditions.
- Because of the tree structure, the neurons at every level almost never receive inputs (PSP) from the same neurons, even indirectly. They are, therefore, uncorrelated to a very good approximation.

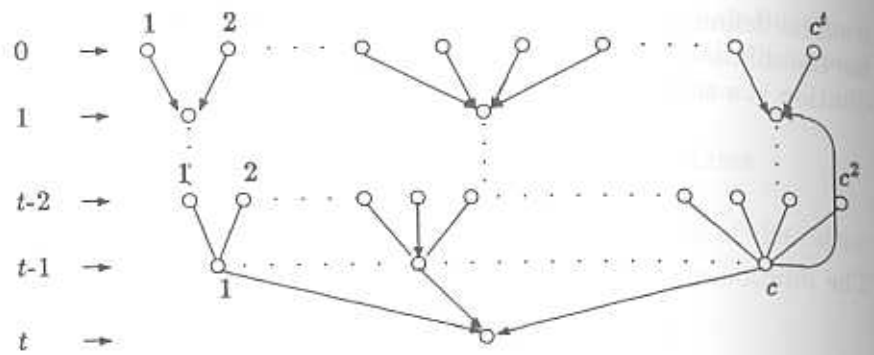


Figure 7.10: Idealized tree structure of an extremely diluted (c=3) standard ANN. Arrows indicate directional coupling (synapses afferent on the neuron receiving the arrow). An example of a feedback loop is included.

- At a time t a neuron will typically receive inputs which have arrived from c^t neurons in the configuration that was present at t=0. See e.g., Figure 7.10.
- The states of two neurons will be uncorrelated at time t if the two trees of afferent neurons connecting them to the initial state at t=0 have no neurons in common.
- Two sets of Q randomly chosen neurons, out of the N available ones, will not have neurons in common if $Q \ll \sqrt{N}$. See e.g., Appendix 7.5.2.
- The condition that neurons will remain uncorrelated after a time t is, therefore,

$$c^t \ll \sqrt{N}.$$

  This, in turn, implies that c must be smaller than any power of N, which is just the condition 7.31.

The absence of correlations between the immediate ancestors (afferents) of a given neuron permits the determination of the probability distribution of the local fields on any neuron and consequently of the resulting new overlaps. What is of central importance is that this assertion holds irrespective of the number of patterns that have been stored in the original synaptic matrix. It depends only on the extent and the randomness of the dilution. The network has been treated

explicitly both in the synchronous and in the asynchronous (Glauber) dynamics, in the presence of fast noise. For retrieval states – a single non-vanishing overlap – it was established that the dynamical equations have the form:

- Synchronous dynamics:

$$m(t+1) = f[m(t)] \tag{7.32}$$

- Asynchronous (Glauber) dynamics:

$$\frac{dm(t)}{dt} = f[m(t)] - m(t). \tag{7.33}$$

with the same function f in both cases. If one takes the thermodynamic limit $N \to \infty$ and then p and c are allowed to become infinitely large, keeping $\bar{\alpha}$ fixed, the form of the function f simplifies to[9]

$$f(m) = \int_{-\infty}^{\infty} \frac{dy}{\sqrt{2\pi}} e^{-y^2/2} \tanh \beta c(\sqrt{\bar{\alpha}}y + m) \tag{7.34}$$

where $\beta = 1/T$.

The equations 7.32 and 7.33 can be analyzed in detail for p finite and for $\bar{\alpha}$ finite. Note first that in both types of dynamics attractors, at which the overlap ceases to change, are determined by the same equation:

$$m = f(m). \tag{7.35}$$

When $f(m)$ of Eq. 7.34 is substituted in this equation for the attractors it becomes almost identical to Eq. 6.32. To see this one simply sets h=0, $\mathbf{m} = (m, 0, \ldots, 0)$. The only difference is that α is substituted by $\bar{\alpha}$ and that the width of the gaussian distribution of the noise is $\sqrt{\bar{\alpha}}$ instead of the more complicated $\sqrt{\alpha r}$. But this is a very significant difference. In particular:

- The transition at finite α from non-retrieval, m=0, to retrieval $m \neq 0$ is continuous at all noise levels.

In other words, as $\bar{\alpha} \to \bar{\alpha}_c(T)$ the retrieval quality decreases continuously to zero.

Given that the transition is of *second order*, one can obtain the relation between the critical storage and the temperature – $\bar{\alpha}_c(T)$ – by

expanding $f(m)$ in powers of m and by comparing the linear terms in Eq. 7.35, as we have done in Section 4.6.2. This gives an $\bar{\alpha}$–T phase diagram corresponding to Figure 6.8, in which the line $\bar{\alpha}_c(T)$ separates a region of retrieval – low $\bar{\alpha}$ low T – from a region of no retrieval m=0 at high $\bar{\alpha}$ and T.

At T=0, the equation for the attractor becomes

$$m = \mathrm{erf}\left(\frac{m}{\sqrt{2\bar{\alpha}}}\right) \tag{7.36}$$

whose fully connected relative is Eq. 6.41. This equation has non-zero solutions for

$$\bar{\alpha} < \bar{\alpha}_c = \frac{2}{\pi}.$$

This is, *prima facie*, a very welcome surprise. The extremely dilute network can not only retrieve the patterns that had been stored in it, its storage capacity, measured per remaining synapse, is significantly higher than that of the fully connected network. But the small values of m near $\bar{\alpha}_c$ are not expected to provide retrieval. Near $\bar{\alpha}_c$ the retrieval quality behaves, typically, as a square root of the displacement, namely

$$m \approx \sqrt{3(\bar{\alpha}_c - \bar{\alpha})}. \tag{7.37}$$

If one were to require retrieval quality of 0.97, as for the saturated symmetric, fully-connected ANN, the maximal $\bar{\alpha}$ would have to be decreased to about 0.3. This value is about twice the α_c for the fully-connected network, an increase which may be accounted for by the fact that the asymmetric network requires twice as many independent synapses as does a symmetric network. On the other hand, if retrieval with low m can be ascribed a useful role, then random dilution – symmetric or asymmetric – may become functional.

- The remarkable fact is that such an extremely randomly (hence asymmetrically) diluted ANN can perform at least as well as the standard, fully connected, symmetric network.

The results of this study should be considered not only as a feat of analytical performance but, equally perhaps, as a climax of the robustness of the standard ANN.

This robustness can be invoked backwards. There are situations where the dynamics has inherently asymmetric synapses, as in networks processing temporal sequences, for example. In such cases, the analysis in terms of a free-energy is powerless. The analysis of the diluted networks is independent of the symmetry of the synapses and can be effectively applied to these cases. The assumption being that the qualitative properties of the more fully-connected asymmetric network will be captured. See e.g., [14].

7.3.5 Functional asymmetry

The last case mentioned in the previous section, that of the asymmetry inherent in networks processing temporal sequences, is typical of the second class of robustness discussed in Section 7.1.1. Asymmetry is *functionally* crucial to the performance of networks storing temporally connected sequences. Without it, there are only fixed points. Such a network cannot be considered as a perturbation on a symmetric network. The situation would be rather similar if asymmetry turned out to be essential for learning, as was mentioned in Section 7.3.2.

These situations are nonetheless discussed in the context of robustness only because, despite the *organized* (non-random) asymmetry that is involved in storing sequences, the basic features of a symmetric ANN are still preserved, albeit on a short time scale. If that time scale is significant cognitively, or computationally, then it is a relevant aspect of robustness. The detailed extent to which the various relevant properties are indeed robust under such functional deviations from the standard ANN are more naturally discussed together with the particular functions.

7.4 Effective Cortical Cycle Times

7.4.1 Slow bursts and relative refractory period

The biological significance of the collective *bursts* of spiking activity in the cortex may still be a matter of controversy. Their significance to the interpretation advocated in this essay is unqualified. Attractors and quasi-attractors, with or without noise, are bursts of spikes. Looking for biological evidence for the presence of bursts in cortical activity one, of course, finds it. There exists data which correlates the appearance of bursts with perception, mainly in the visual and somato-sensory cortex in wakeful animals. This issue was raised in Section 4.3.3. and recordings were illustrated in Figure 4.6. Still, if these bursts are to be the candidates for our attractors, one must cope with the fact that these

bursts are significantly slower than what would have been expected on the basis of the *absolute refractory period* which has determined our discretization interval. As we shall show below, a reduction of the rates in bursts from 500 to 150 per second can be relatively easily accounted for. To account for significantly slower bursts, of 30 or 40 per second, requires more elaborate modifications in structure or in neuronal function. In associative parts of cortex bursts never appear with higher rates.

A straight-forward way of producing such rates has been implemented in simulating the olfactory cortex at Caltech.[3] Structure is introduced via an arrangement of slow communication between excitatory neurons – the associative communication – and fast, local, inhibition via small neurons. The logic of the ANN will quite cleary be robust under such modifications and the rates can be made arbitrarilly slow, depending on the biologically available axonal delays.

The physiological origin of the reduction from 500 to 150 per second may be of a different origin. It may be related to the effects of the *relative refractory period* alongside the absolute one. See e.g., Kuffler et al[15] (p.152). The two can be contrasted as follows:

- Within the absolute refractory period following a spike a neuron cannot emit another spike irrespective of the afferent potential. On the other hand, following that short period, of 1–2 milliseconds, the neuron can fire again but the PSP required is higher than the usual threshold. Effectively, following the absolute refractory period, a neuron has an enhanced threshold. An artificially stimulated neuron, or a neuron connected to a sensory organ, can reach spike rates as high as would be limited by the absolute period, namely as high as 500 per second. In a network, the PSP's are never much higher than the normal threshold and hence the relative refractory period becomes effective.

This explanation is quite satisfactory from the physiological and neuro-chemical point of view, but it poses a real problem for the simplified neural dynamics considered up till now. If, in fact, the PSP's produced within the network are never much higher than the refractory thresholds, how can the network sustain a pattern in which a specified set of neurons fire bursts? If a set of neurons, those that should be active in the memorized pattern, have emitted a spike at a certain instant,

[3] I owe this account Dr. James Bower.

then they will all receive the same PSP's as the ones that have provoked the first set of spikes. Those are just the PSP's that are produced by the particular activity configuration in the pattern. But if following the first emission of spikes the threshold has increased, the neurons cannot fire in the second interval. For, if they did, they would be firing after every absolute refractory period and the burst rate would again be much higher than observed. In the second cycle-time, all neurons will be quiescent and there is nothing in the model to rekindle them.

7.4.2 Neuronal memory and expanded scenario

The missing element is memory in the individual neuron.

- One such element is the state of the neuron's threshold. It records firing events by resetting the neuronal threshold to the higher value. It also records the duration since the emission of the spike, by the gradual approach of the threshold to its refractory, equilibrium, value.

- Another neuronal memory element is the gradual decay of the PSP of a neuron that has not fired, as opposed to the simplified picture in the standard ANN in which neurons lose their PSP every cycle. Such a gradual decay can be accounted for by the resistive leak through the membrane. See e.g., Eq. 2.1.

A candidate scenario that can save the situation may be, schematically, as follows:

1. A set of neurons fire in the initial state which is near a pattern.

2. Neurons that have fired lose their potential and come out with an enhanced threshold.

3. The threshold of the neurons which have fired in the initial pattern begins to relax to its equilibrium value, since in the cycle-time following the spike they will not emit another spike.

4. Neurons that do not fire have their membrane potential decay gradually with time towards a refractory value.

5. In the next time slot, 1–2 milliseconds, each neuron receives afferent potential, as in the standard model. The neurons which

should be active in the pattern arrive at membrane potentials which are higher than the refractory threshold. Yet, they remain quiescent because of their increased threshold.

6. When the decaying potential intersects the relaxing threshold, the neuron can rekindle its action potential.

This is clearly still a caricature of realistic neuronal dynamics. But rather than undertake the system in its full complexity, see e.g., ref. [16], we will continue to add complications one at a time.

The above scenario is merely an indication that when the standard ANN is extended to take into account effects of relative refractory period it is not inconceivable that patterns of active neurons which produced bursts (attractors) on the short time scale may reappear on the longer one. That the scenario can actually proceed in a collective manner and is not washed out by fluctuations has to be verified by analysis and by simulations. A detailed analysis is a rather formidable task, because of the time dependence in the model's parameters. Simulations accompanied by simple heuristic arguments have indeed shown that the extension described here can reproduce pattern retrieval by bursts with maximal frequency which can be significantly lower than the absolute refractory period.

7.4.3 Simplified scenario for relative refractory period

Simulations were carried out on networks[17] with 200 neurons. In order to reduce the effects of correlations between random patterns, the synaptic matrix was constructed as a projector matrix (see e.g., Section 4.2.3). What one finds is a rather rich situation, less ordered than a mere rescaling of the basic bursting rate. In the absence of fluctuations, when $N \to \infty$, all neurons receive the same potential when the network is in a pattern. In that ideal case, the scenario would proceed perfectly, provided the relative decay times of the potential and the threshold have been chosen so that an intersection will take place, despite the fact that the depolarization potential relaxes to a value lower than that to which the threshold decays. What would take place is that a network that enters a memorized pattern would have all the neurons that are supposed to be active in that pattern emit a spike. In the next basic cycle-time (t_{arp}) they will all receive afferent inputs equal to those which have caused the previous spiking activity, yet they

will remain refractory because of the increase in their thresholds. The potentials they have are all above the equilibrium threshold U_T. Now the network awaits the passage of a time interval which corresponds to the intersection time τ_{eff}. After that period, all the neurons active in the pattern fire once more in concert. This behavior will be repeated periodically, with period τ_{eff}/t_{arp}. All that has happened is that the attractor looks just like in the elementary network with slower bursts.

But fluctuations make the time course more complex, even when the network is perfectly aligned with a pattern. Different *active* neurons receive different potentials due to fluctuations. This is less evident with the projector synaptic matrix than it would have been for a standard 'Hebbian' matrix. But fluctuations enter in two ways:

- The subtraction of the self-connecting synapses.
- The transformation to (0-1)-neurons.

Thus neurons which have fired in a pattern will divide into groups, depending on the size of their potential. Those that fire again first will be the ones that had the highest contribution from the initial firing in the pattern. In firing they will contribute to the PSP of other neurons which are *active* in the pattern, but which have had a potential that is still lower than the relaxing, enhanced, threshold. Some of those may now fire, others may still wait. The ones that have not yet crossed the threshold will obtain additional inputs, etc. The result is that a new cycle-time appears that describes the reappearance of the pattern in parts and pieces, not as simultaneous spikes of all the *active* neurons. At different moments in time, different groups of the *active* neurons will fire until all, or most, of them have fired and then a new cycle starts. This is quite similar to the randomization of the basic cycle-time described in Figure 2.2.

The interpretation of retrieval, in a situation in which the pattern never appears in the activity of the network at any given moment, must involve averaging over the activity of the neurons in the longer effective period. This will bring into relief the appearance of an entire pattern. The presence in an attractor must be read as a high average activity over a number of effective cycle-times. If the entire pattern takes n_c short cycle-times to reappear, the mean activity of the *active* neurons will be n_c^{-1}. It must still be much higher than the spontaneous activity due to noise. The introduction of averaging as a step in the

detection of retrieval is, of course, not new. It had to be introduced to allow retrieval in the presence of fast noise, as in Section 2.3.3. But temporal averaging was necessary even in the noiseless situation to establish that the network has actually entered an attractor and is not merely passing a transient. The consequence of a longer effective cycle-time is that averaging must take place on a correspondingly longer time interval. In other words, if n_r cycles were required to establish presence in a noiseless attractor, $n_r n_c$ elementary cycles will be needed for the slower bursts. But this implies simply that one would have to average for a time which is n_r times the effective new burst period. If this is read to be some 7–8ms, then one would expect 25–30ms to be sufficient. This is a rather reasonable order of magnitude as was indicated in Section 2.3.2.

Simulations corroborate the expectations from this scenario. Observing the relative behavior of the membrane potential and the threshold in a functioning network, for different types of neurons, one notices the appearance of early and late groups of rekindled neurons and the corresponding additional contributions to the PSP in the course of the *relative refractory period*. Detailed spike records indicate that the network can retrieve associatively, i.e., it corrects errors. An example is presented in Figure 7.11 in which we present the statistics of networks with 200 neurons, with α=0.05, as a function of the overlap of the initial state with a stored pattern m_0. There is full error correction down to m_0=0.8 for more than 90% of the runs. Recall that with 0.1 of the neurons out of alignment with a pattern in the initial state, the initial overlap is 0.8. From Figure 7.11 one reads that at this level of errors, the mean retrieval is better than 0.9.

The storage capacity of the network is reduced relative to the standard network, but preliminary results indicate that the reduction is not greater than that which would have been implied by the mere transformation to a network of (0,1) neurons[18]. This is presented in Figure 7.12 which indicates that one can store and retrieve up to a level of α=8% and make no more than 2.5% errors.

7.5 Appendix: Technical Details

7.5.1 Digression - the mean-field equations

Since the noise added to the synapses is a gaussian noise, the extension of the theoretical development of Section 6.3.1 is rather natural. In

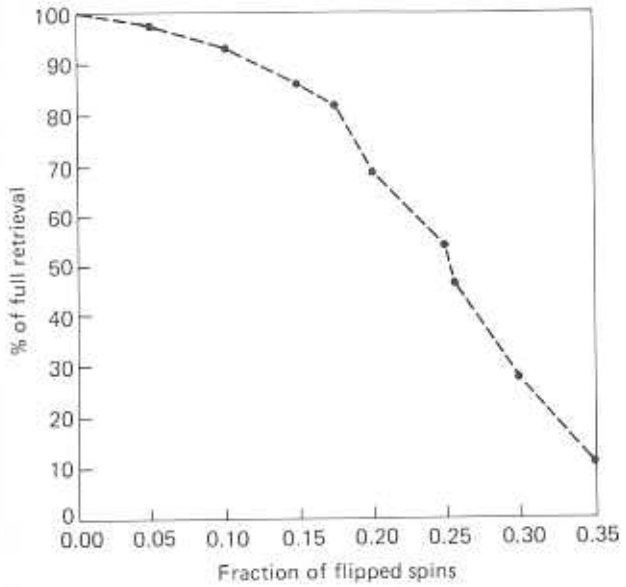


Figure 7.11: Fraction of runs, out of 300, giving retrieval to better than m =0.9 vs fraction of misaligned neurons. N=200, p=10.

this section we will assemble the definitions and the results only. The reader who has mastered the derivations in Chapter 6 will encounter no difficulty in assimilating the extra noise in the procedure.

First, there are the obvious modifications of the order-parameters which have been introduced in Section 6.3.1. They all require averaging over the additional random variables η_{ij}.

- *The overlap* for a retrieval state

$$m = \left[\left\langle \left\langle \frac{1}{N} \sum_{i=1}^{N} \xi_i^1 \langle S_i \rangle \right\rangle \right\rangle \right], \tag{7.38}$$

in which three averages are implied:

- The fast, *thermal*, average – $\langle \cdots \rangle$.
- The *quenched* average over the random patterns – $\langle\langle \cdots \rangle\rangle$.
- The *quenched* average over the gaussian noise – $[\cdots]$.

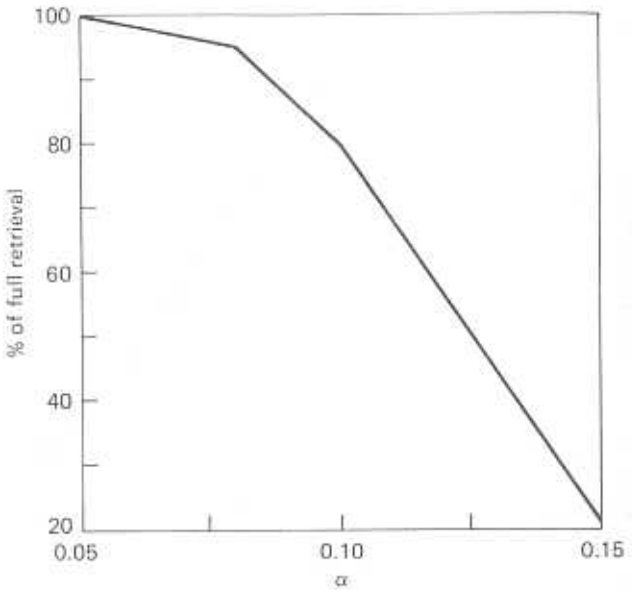


Figure 7.12: Statistics of stability of memorized patterns in a networks of 200 neurons as function of the memory loading α. Each point represents the fraction, out of 300 runs, that have been retrieved with m >0.9. Parameters as in Figure 7.11.

- *The spin-glass order-parameter*

$$q \equiv \left[\left\langle\left\langle \frac{1}{N}\sum_{i=1}^{N}\langle S_i\rangle^2 \right\rangle\right\rangle\right], \tag{7.39}$$

describing the temporal freezing of uncorrelated spins.

- *The effective gaussian noise* parameter

$$r \equiv \left[\left\langle\left\langle \frac{N}{p}\sum_{\mu=s+1}^{p=\alpha N}\langle m^\mu\rangle^2 \right\rangle\right\rangle\right] = \left[\frac{1}{\alpha}\sum_{\mu>s}\left\langle\left\langle\left(\frac{1}{N}\sum_i \xi_i^\mu\langle S_i\rangle\right)^2\right\rangle\right\rangle\right]. \tag{7.40}$$

In terms of these order-parameters, one can write the free-energy of the network, corresponding to Eq. 6.36, as well as the equations for its extrema[1]:

$$\begin{aligned} f &= \tfrac{1}{2}\alpha + \tfrac{1}{2}m^2 + \frac{\alpha}{2\beta}\left(\ln(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}\right) \\ &+ \tfrac{1}{2}\alpha\beta r(1-q) - \frac{1}{4\beta}\eta^2(1-\beta+\beta q)^2 \\ &- \frac{1}{\beta}\int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\ln 2\cosh\beta[\sqrt{\alpha r+\eta^2 q}\,z+m]. \end{aligned}$$

The resulting mean-field equations are:

$$m = \int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\tanh\beta\left[\sqrt{(\alpha r+\eta^2 q)}\,z+m\right] \tag{7.41}$$

$$q = \int_{-\infty}^{\infty}\frac{dz}{\sqrt{2\pi}}e^{-z^2/2}\tanh^2\beta\left[\sqrt{(\alpha r+\eta^2 q)}\,z+m\right] \tag{7.42}$$

$$r = \frac{q}{(1-\beta+\beta q)^2}. \tag{7.43}$$

The main point to notice here is that the synaptic noise adds a gaussian term, of variance $\eta^2 q$, to the gaussian noise coming from the random overlaps of the patterns, which is familiar from the theory of the SK-model[3].

The noiseless limit

The results discussed in the previous sections have dealt with the performance of the network in the absence of fast noise. The above mean-field equations provide the full theory in the simultaneous presence of the three sources of noise, and detailed consequences can be drawn from them by numerical analysis. Here we will proceed to the limit T=0 to fill in the expressions which led to the specific results reported above. The details are very similar to those traced in Section 6.3.3.

The three equations for the retrieval states are:

$$m = \mathrm{erf}\left(\frac{m}{\sqrt{2(\alpha r+\eta^2)}}\right) \tag{7.44}$$

$$r = (1-C)^{-2} \tag{7.45}$$

$$C = \sqrt{\frac{2}{\pi(\alpha r + \eta^2)}} \exp\left(-\frac{m^2}{2(\alpha r + \eta^2)}\right). \quad (7.46)$$

The analysis of the retrieval regime proceeds just like in Section 6.3.3.

- When α or η are too large the only solution is the spin-glass, with m=0 and q=1. This leads to the phase diagram Figure 7.1.
- From Eq. 7.44 it follows that even when α=0, $m < 1$, because of the presence of noise.
- The transition at η_c is continuous, and the value of η_c can be computed to be

$$\eta_c = \sqrt{2/\pi}.$$

Near and below $\eta_c(\alpha{=}0)$, the overlap m is very small. The right hand side of Eq. 7.44 can be expanded, much like the expansion we have performed near T=1 in Section 4.6.2. The critical value of η is obtained by equating the linear term in m on both sides of the expanded equation. The linear term in the expansion of the right hand side is[19]

$$m\eta\sqrt{\frac{\pi}{2}}$$

which leads to η_c.

7.5.2 Dilution requirement

The probability for none of the neurons in the second set of Q randomly selected neurons not to belong to the first set is

$$\Pr = \left(1 - \frac{Q}{N}\right)^Q \approx \exp\left(-\frac{Q^2}{N}\right).$$

If $Q = aN^b$, then

$$\Pr \approx \exp\left(-a^2 N^{2b-1}\right).$$

Consequently, if $b < \frac{1}{2}$, $\Pr \to 1$ when $N \to \infty$, while $\Pr \to 0$ for $b > \frac{1}{2}$.

Bibliography

[1] H. Sompolinsky, Neural networks with non-linear synapses and static noise, *Phys. Rev.*, **A34**, 2571(1986) and The theory of neural networks: The Hebb rule and beyond, in L. van Hemmen and I. Morgenstern eds. *Heidelberg Colloquium on Glassy Dynamics* (Springer-Verlag, Heidelberg, 1987).

[2] G. Toulouse, S. Dehaene and J.P. Changeux, Spin glass model of learning by selection, *Proc. Nat. Acad. Sci. (USA)*, **83**, 1695(1986) and J.P. Nadal, G. Toulouse, J.P. Changeux and S. Dehaene, Networks of formal neurons and memory palimpsests, *Europhys. Lett.*, **1**, 535(1986).

[3] S. Kirkpatrick and D. Sherrington, Infinite-ranged models of spin-glasses, *Phys. Rev.*, **B17**, 4384(1978).

[4] R.J. McEliece, E.C. Posner, E.R. Rodemich and S.S. Venkatesh, The capacity of the Hopfield associative memory, *IEEE Trans. IT*, **33**, 461(1987).

[5] I. Morgenstern, Spin-glasses, optimization and neural networks, L.N. van Hemmen and I. Morgenstern eds. *Heidelberg Colloquium on Glassy Dynamics* (Springer-Verlag, Heidelberg, 1987).

[6] J.J. Hopfield, Neural networks and physical systems with emergent selective computational abilities, *Proc. Natl. Acad. Sci. USA*, **79**, 2554(1982).

[7] J.A. Hertz, G. Grinstein and S.A. Solla, Irreversible spin glasses and neural networks, L.N. van Hemmen and I. Morgenstern eds. *Heidelberg Colloquium on Glassy Dynamics* (Springer-Verlag, Heidelberg, 1987).

[8] M.V. Feigelman and L.B. Ioffe, The augmented models of associative memory: asymmetric interaction and hierarchy of patterns, *Int. Jour. of Mod. Phys.*, **B1**, 51(1987).

[9] B. Derrida, E. Gardner and A. Zippelius, An exactly soluble asymmetric neural network model, *Europhys. Lett.* **4**, 167(1987).

[10] A. Crisanti and H. Sompolinsky, Dynamics of spin systems with randomly asymmetric bonds: I: Langevin dynamics and a spherical model, *Phys. Rev.* **A36**, 4922(1987) and II: Ising spins and Glauber dynamics, *Phys. Rev.*, **A37**, 4865(1988).

[11] G. Parisi, Asymmetric neural networks and the process of learning, *J. Phys.*, **A19**, L675(1986).
[12] J.C. Eccles, *The Physiology of Synapses* (Springer-Verlag, Berlin, 1964).
[13] S. Shinomoto, A cognitive associative memory, *Biol. Cybern.*, **57**, 197(1987).
[14] H. Gutfreund and M. Mezard, Processing temporal sequences in neural networks, *Phys. Rev. Lett.*, **61**, 235(1988).
[15] S.W. Kuffler, J.G. Niccols and A.R. Martin *From Neuron to Brain* (Sinauer, Sunderland, Mass., 1984).
[16] J. Buhmann and K. Schulten, Influence of noise on the function of a "physiological" neural network, *Biol. Cybern.*, **56**, 313(1987).
[17] M. Aharoni, A proposal for a model neural network, MSc. thesis (in Hebrew), Hebrew University (1988).
[18] A.D. Bruce, E.J. Gardner and D.J. Wallace, Dynamics and statistical mechanics of the Hopfield model, *J. Phys.*, **A20**, 2909(1987).
[19] I.S. Gradsteyn and I.M. Ryzhik, *Tables of Integrals Series and Products* (Academic Press, New York, 1965).

8

Memory Data Structures

8.1 Biological and Computational Motivation

8.1.1 Low mean activity level and background-foreground asymmetry

The simplest and most evident requirement for data structuring that goes beyond the random patterns treated so far is the biological finding that the mean spatial level of activity in the cortex is relatively low. It appears that in regions of cortex in which associative functions are detected a representative fraction of neurons that increase their spike activity would be 4–5%.[1] In contrast, the patterns we have been storing, whose elements are chosen with equal probability to be ± 1, represent a situation in which 50% of the neurons are active when the network enters an attractor. Low levels of activity imply that memories have to be selected with a probability which is *biased* toward the –1's. For example, if the probability for choosing the state of a neuron to be $+1$ is

$$\Pr(+1) = \tfrac{1}{2}(1+a), \tag{8.1}$$

then the –1's will be selected with probability $1-\Pr(+1)$. The mean activity in the network will be

$$A = \tfrac{1}{2}(1+a) \tag{8.2}$$

[1] I am indebted to Prof. M. Abeles for enlightening communications on this subject.

and the mean excess of active over passive neurons in every memorized pattern will be

$$N_+ - N_- = Na \quad \left(= \sum_{i=1}^{N} \xi_i^\mu\right). \tag{8.3}$$

Another simple motivation for the construction of ANN's storing and retrieving *biased* patterns comes from the realm of pattern recognition. It would be rather natural to design an artificial ANN for pattern recognition by representing black and white pixels in a two-dimensional image on active and passive 'neurons' in a two dimensional array. It may be the case that certain stages of the visual system operate in this manner. Quite often, visual patterns to be memorized are biased in favor of the background, leading again to an excess of passive to active neurons. This goes sometimes under the heading of *sparse coding*[1].

The reason for considering the subject of low mean activity within the context of data structures is that patterns which are *biased* in the sense discussed above are automatically correlated with each other. From Eq. 8.1, it follows that the overlap of two different patterns ($\mu \neq \nu$) is

$$\langle\langle \xi^\mu \xi^\nu \rangle\rangle = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu = a^2. \tag{8.4}$$

This is, of course, a rather simple form of correlation but it can be used as a building block for more complex data structures; as will be done in Sections 8.3.1 and 8.3.5, below.

8.1.2 Hierarchies for biology and for computation

It has been a tenet of the classical view of forebrain organization that perceptual processing of sensations is performed in hierarchical fashion, via hierarchically organized cortical components. More recent investigations have questioned several of the central theses of the classical view[2]. But the hierarchical aspect was challenged only in regard to its input mode. While the traditional view had it that the hierarchical processing system received sensory input "from unitary 'main-line' projection system," it appears that inputs enter directly at several levels. As we shall see below, a hierarchical structure of ANN's of this type can be formulated and analyzed. See also ref. [3].

Another aspect which suggests hierarchical data structures finds its origin in computational considerations of 'connectionist models'[4].

Function tables – the mapping of sets of $\{operand_1\text{-}operation\text{-}operand_2\}$ onto *result* – have a quite natural description in terms of hierarchical structures. Such structures would also seem appropriate for some aspects of semantic categorization, namely for the analysis of words in terms of categories which consecutively embed a particular word. In fact, some applications of such hierarchical data structures in ANN's have just the flavor expressed in Fodor's words[5]:

> A more likely story is the following: the mental lexicon is a sort of graph, with the lexical items at the nodes and with paths from each item to several others. We can think of accessing an item in the lexicon as, in effect, exciting the corresponding node...Accessing a given lexical item will thus decrease the response time for recognizing token items to which it is connected.

8.2 Local Treatment of Low Activity Patterns

8.2.1 Demise of naive standard model

There is always a ready-made prescription for storing any arbitrary collection of linearly independent patterns, irrespective of any mutual correlations. It is a synaptic matrix – the *pseudo-inverse*, introduced in Section 4.2.3 – which projects the stored patterns in orthogonal directions. If one uses this prescription then patterns which are hierarchically organized are all treated on equal footing and with equal efficiency. Here we follow a different route, constructing a minimal extension of the local standard prescription to allow for the storage and retrieval of *biased* patterns. The motivations are that:

- Locality of synaptic values is an essential ingredient for biological systems.
- There are situations in which the hierarchical data structure, as well as the structure of information processing, is of the essence in terms of the timing and the ordering. Decorrelating prescriptions eliminate these structures.

For patterns generated according to Eq. 8.1, the standard prescription will not do. The reason is that due to the mutual correlations

each neuron experiences a non-zero mean noise contributed by every stored pattern. As a consequence, beyond a very low level of storage, the patterns no longer attract and cannot be retrieved. Specifically, if p patterns are stored with *bias* a using the synaptic matrix of Eq. 4.3, then the local field at neuron i when the network is in pattern number 1 ($S_i = \xi_i^1$), for example, is

$$h_i = \sum_{j\, j\neq i}^{N} J_{ij} S_j = \frac{1}{N} \sum_j \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \xi_j^1. \tag{8.5}$$

Proceeding as in Section 4.3.1, we separate the right hand side into a *signal* term and a *noise* term. The signal originates from the stabilizing effect of the term with μ=1 in J_{ij} and the rest is the noise. The signal is

$$S = \frac{1}{N} \sum_j \xi_i^1 \xi_j^1 \xi_j^1 = \xi_i^1.$$

The noise is

$$R = \frac{1}{N} \sum_j \sum_{\mu=2}^{p} \xi_i^\mu \xi_j^\mu \xi_j^1 = a^2 \sum_{\mu=2}^{p} \xi_i^\mu,$$

where we have used Eq. 8.4. The last factor on the right hand side can take values between $-(p-1)$ and $(p-1)$ and the patterns become destabilized when

$$(p-1)a^2 = 1. \tag{8.6}$$

To appreciate the scale of the disaster consider, for example, the case where the mean level of activity is 5%. The value of the bias a can be deduced from Eq. 8.2 to be

$$a = 2A - 1 \approx -0.9$$

and p must not exceed 2(!) irrespective of the value of N.

It is instructive to construct the modified network in two stages. They will be described in the following two sections.

8.2.2 Modified ANN and a plague of spurious states

Modification of synaptic matrix

By a straightforward modification one eliminates the problem of the non-vanishing mean noise encountered in the naive standard prescription. To keep a symmetric synaptic matrix, one writes

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} (\xi_i^\mu - a)(\xi_j^\mu - a). \tag{8.7}$$

With this prescription one still has an energy function

$$E = -\tfrac{1}{2} \sum_{i\neq j} J_{ij} S_i S_j \tag{8.8}$$

and the local field on neuron i is given by

$$h_i = \sum_{j\, j\neq i}^{N} J_{ij} S_j = \frac{1}{N} \sum_j \sum_{\mu=1}^{p} (\xi_i^\mu - a)(\xi_j^\mu - a) S_j. \tag{8.9}$$

Note that the synaptic matrix in Eq. 8.7 is only quasi-local. The subtraction a in the two factors of the synaptic strength is, of course, a global property of the network. It is, however, a single parameter which comprises this entire non-locality. Eq. 8.7 reinterprets Hebb's rule to read that synaptic modifications are proportional to the correlations of the access activities of the two neurons connected by the particular synapse over the mean activity. See e.g., ref. [6].

Let us now consider the stability of the stored patterns by the simple signal-to-noise estimation. This we do by substituting $S_i = \xi_i^1$ in Eq. 8.9 and separating in the sum the term with μ=1 as the signal from the rest of the terms which act as noise. We find

$$S = (\xi_i^1 - a) \frac{1}{N} \sum_j (\xi_j^1 - a)\xi_j^1 = (\xi_i^1 - a)(1 - a^2), \tag{8.10}$$

keeping the notation of the previous section. The noise is

$$R = \sum_{\mu=2}^{p} (\xi_i^\mu - a) \frac{1}{N} \sum_j (\xi_j^\mu - a)\xi_j^1 = \sum_{\mu=2}^{p} (\xi_i^\mu - a) \left(a^2 - a^2 + O(N^{-\frac{1}{2}}) \right), \tag{8.11}$$

which is essentially zero. The effect of the noise has to be estimated by its variance $\langle\langle R^2\rangle\rangle$, which is (see e.g., Appendix 8.5.1)

$$\rho^2 \equiv \langle\langle R^2\rangle\rangle = \frac{(N-1)(p-1)}{N^2}(1-a^2)^2 \approx \frac{p}{N}(1-a^2)^2. \tag{8.12}$$

Clearly, this network can have a macroscopic number of stable patterns. In fact, the lower limit of the signal is

$$|S| \geq (1-a^2)(1-|a|)$$

and hence

$$\frac{\rho}{S} \leq \sqrt{\frac{p}{N}}\left(\frac{1}{1-|a|}\right). \tag{8.13}$$

The conclusion is that if

$$\alpha = \frac{p}{N} \ll (1-|a|)^2$$

all the memorized patterns are stable in the absence of fast noise. But, the storage capacity is reduced compared to the capacity of the network storing unbiased patterns. Comparing this result with that of Section 6.2.1, namely $\alpha(a=0) \ll 1$, one has

$$\alpha_0(a) = \alpha_0(a=0)(1-|a|)^2,$$

where α_0 denotes the storage capacity provided by the naive signal-to-noise estimate. In fact, this estimate can be made statistically more sophisticated, treating properly the tails of the distribution of the noise, as has been done in Section 6.2.1. Equation 6.15 is then replaced by

$$\alpha_c(a) = \frac{(1-|a|)^2}{2\ln N} = (1-|a|)^2\alpha_c(0). \tag{8.14}$$

As we shall see below, this decrease in the storage capacity with the bias, sketched in Figure 8.2, is not a feature of a proper network storing low activity patterns. Actually, a reasonably functioning network increases its capacity as the level of activity is lowered (as the bias is increased).

Performance difficulties at low loading levels

An indication that the prescription Eq. 8.7 is not satisfactory appears already in networks at low loading. To discuss these, it is convenient to define a modified overlap *order-parameter* as

$$m^\mu = \frac{1}{N}\sum_{i=1}^{N}\langle(\xi_i^\mu - a)S_i\rangle, \tag{8.15}$$

where the angular brackets stand for a temporal (thermal) average, as in Section 4.4.1. This parameter varies in the interval $(0, 1-a^2)$. In terms of this variable, one can write the mean-field equations which determine the meta-stable states of the network, as well as its free-energy. One has

$$m^\mu = \left\langle\left\langle (\xi^\mu - a)\tanh\left(\beta\sum_{\nu=1}^{p} m^\nu(\xi^\nu - a)\right)\right\rangle\right\rangle \tag{8.16}$$

for the set of self-consistency equations, and for the free-energy

$$f = \frac{1}{2}\sum_{\mu=1}^{p}(m^\mu)^2 - \frac{1}{\beta}\left\langle\left\langle \ln 2\left[\cosh\left(\beta\sum_{\nu=1}^{p} m^\nu(\xi^\nu - a)\right)\right]\right\rangle\right\rangle. \tag{8.17}$$

These should appear as rather natural extensions of Eqs. 4.40 and 4.42.

While ξ^μ and ξ^ν are correlated for all μ and ν, one has

$$\langle\langle(\xi^\mu - a)(\xi^\nu - a)\rangle\rangle = (1-a^2)\delta_{\mu\nu},$$

which ensures that retrieval states, with a single non-zero modified overlap, of the form

$$m^\mu = m\delta_{\mu\nu} \tag{8.18}$$

are solutions of the mean-field equations. When Eq. 8.18 is substituted in Eq. 8.16 one finds an equation for m which reads

$$m = \tfrac{1}{2}(1-a^2)[\tanh\beta m(1-a) + \tanh\beta m(1+a)],$$

and in the noiseless limit

$$m = 1 - a^2.$$

One should be aware of the fact that this value of our modified overlap implies full alignment with the pattern ν. Only if $S_i = \xi_i^\nu$ will m^μ=0 for $\mu \neq \nu$ and $m^\nu = 1-a^2$. The real overlaps, which will be temporarily denoted by $\bar{m}$ are then

$$\bar{m}^\nu = 1; \qquad \bar{m}^\mu = a^2 \quad (\mu \neq \nu).$$

In terms of the modified overlaps, one finds that all symmetric mixtures are solutions of the mean-field equations, much as in Section 4.5. The difficulties begin to appear when the stability of the spurious states is examined[7]. First one finds that unlike for the unbiased patterns the even mixtures are usually stable as well as the odd ones. In Section 4.5.2 we have traced the stability of the spurious states to the absence of sites with vanishing local fields. From Eq. 8.16 one can read off the local field at a symmetric mixture with n equal overlaps m_n to be

$$h_i = \sum_{\mu=1}^{p} m^\mu(\xi_i^\mu - a) = m_n(z_n^i - na)$$

where

$$z_n = \sum_{\mu=1}^{n} \xi_i^\mu.$$

For a=0 and n even, h_i=0 is a direct consequence of the fact that z_n^i=0 at

$$\frac{1}{2^n}\binom{n}{\frac{1}{2}n} N$$

of the sites. It is the product of N and the probability that $n/2$ out of n ξ^μ's are +1. But $z_n - na$ will typically not vanish for either odd or even values of n. The number of meta-stable spurious states at least doubles.

Moreover, if the energies E_n of the lowest mixture states are evaluated as a function of a, one finds the peculiar behavior plotted in Figure 8.1 for n=1-5. Note that beyond $|a| = \sqrt{2} - 1$ the energies begin to cross and spurious states become the absolute minima of the energy. The first mixture to become more stable than the retrieval states is the 2-mixture.

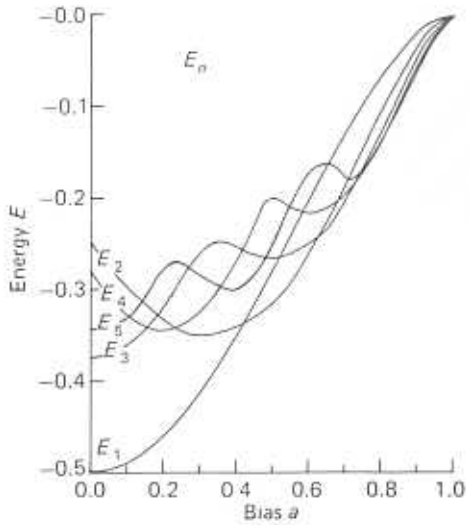


Figure 8.1: Energies E_n of the first five symmetric mixture states vs the bias a in a network at low loading level[7].

Things become even worse when one considers the situation in the presence of noise. One finds for example that for $|a| > \frac{1}{3}$ the retrieval states are unstable near the transition temperature, namely just below

$$T_c(a) = 1 - a^2.$$

As the temperature is decreased, one first encounters stable mixture states. Only at lower temperatures do the retrieval states become stable, at which point some spurious states are already lower in free-energy than the retrieval state.

Near saturation

When $p = \alpha N$ the properties of the network storing biased patterns have to be computed by the techniques of Section 6.3.1 [7]. In Figure 8.2 we reproduce the dependence of the storage capacities for the stability of the retrieval state and a few symmetric mixture states vs the bias a, at T=0. Spurious states predominate here again. Below the curve $\alpha_c(a)$ there are dynamically stable retrieval states, but above $a \approx 0.4$ the 2-mixture is stable at higher storage levels, where the retrieval state

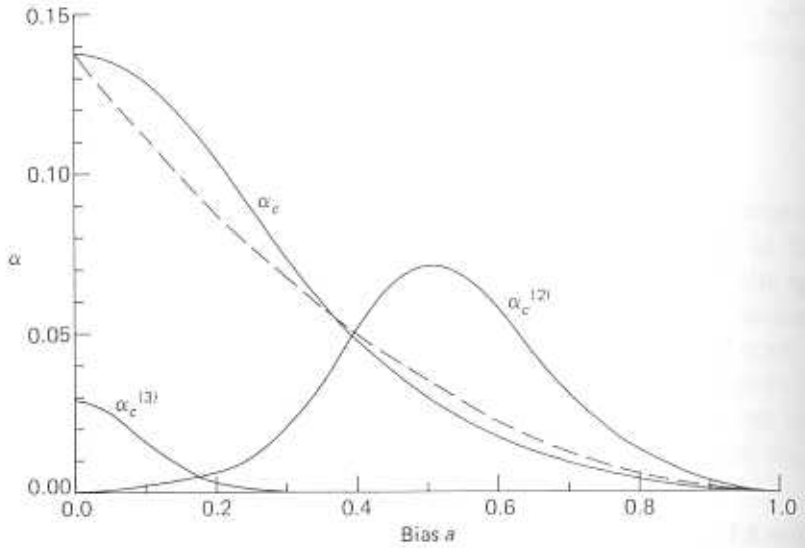


Figure 8.2: Storage capacities of several attractors vs bias a: α_c (retrieval states), $\alpha_c^{(2)}$ (2-mixtures) and $\alpha_c^{(3)}$ (3-mixtures). Dashed curve is $\alpha_c(a = 0)(1 - a^2)$[7].

is still unstable. When the storage level is reduced and the retrieval state becomes stable, the 2-mixture is already much lower in energy. For comparison, Figure 8.2 presents also the reduction of α_c due to signal-to-noise effects, with $\alpha_c(a = 0)$=0.138.

8.2.3 Constrained dynamics – monitoring thresholds

Most of the problematic features of the modified ANN discussed above have been traced to a rather biological origin. If one is to take seriously Hebb's mechanism of synaptic modification based on the actual activity of the network, then the fact that the network is found to have stored *biased* – low activity – patterns must be related to the activity distributions which the network can typically exhibit. It indicates that the dynamics of the network is effectively restricted to a part of the full space of 2^N possible network states, namely the network is restricted to wander only in a subset of states – all of mean activity in the vicinity of the activity in the biased patterns[7]. A natural way of implementing

this idea is to extend the activity constraint of the memorized patterns to the dynamical states of the network, i.e.,

$$\frac{1}{N}\sum_{i=1}^{N} S_i = a. \tag{8.19}$$

One immediate consequence of the constraint is the breaking of the symmetry $S_i \rightarrow -S_i$. This symmetry has led to a doubling of every solution for the attractors, as was discussed in Section 4.5. Two states which have all the corresponding spins reversed are no longer equivalent. In particular, it implies that on storing a set of patterns, their reversed states are no longer attractors unless a=0. Note that these reversed patterns have not become unstable. They have been pushed out of the space of states visited by the dynamics.

Technically, the imposition of the constraint can be affected either in a *hard* or in a *soft* way. The hard constraint imposes that the network move within the strict confines of the subspace in which the constraint holds. In the soft way, the network is allowed to drift outside this subspace, but for a price. It can be implemented by adding to the energy, Eq. 8.8, a cost term of the form

$$E_c = \frac{g}{2N}\left(\sum_{i=1}^{N} S_i - Na\right)^2 \tag{8.20}$$

where the coefficient g measures the strength with which the constraint is imposed. As $g \rightarrow \infty$, the constraint becomes strict again. Adding a term like this lowers the probability of states whose activity deviates either up or down. The membrane potentials (local fields) are modified by

$$\bar{h}_i = -g\left(\frac{1}{N}\sum_{j=1}^{N} S_j - a\right), \tag{8.21}$$

which is a uniform level depending on the momentary activity in the network.

This additional potential can be interpreted in two ways

- As an external control by potential inputs, which ensures that the activity in the network be constantly corrected toward the desired mean.

If it is too high, $\bar{h}_i$ becomes inhibitory (hyper-polarizing), if it is too low all neurons receive an equal excitatory (depolarizing) input. The magnitude of these inputs increases with the magnitude of the deviation from the desired mean activity. It is a servo-mechanism.

- As a uniform inhibitory contribution to the synapses of efficacy $-g/N$ and a constant, uniform excitatory potential ag. This, in turn, can be interpreted as a uniform lowering of the thresholds.

Clearly, the properties of the network should not be too sensitive to the particular value of the constraint parameter g. This is ensured if as g becomes large the properties of the network have a well defined limit, which is the strictly constrained network. Then, if g is large enough, the results become essentially independent of g. Solving the equations, one finds that by the time g=10, the asymptotic behavior is reached for all values of the bias a, as is depicted in Figure 8.3, where the storage capacity vs g is plotted for several values of a. One observes that saturation is achieved latest for intermediate values around a=0.5.

8.2.4 Properties of the constrained biased network

Neither modification of the network dynamics described in Sec. 8.2.3 affects the symmetry of the synaptic matrix. The properties of the network can, therefore, be analyzed by minor extensions of the techniques of Chapters 4 and 6. Here we shall confine ourselves to results and describe the case of the hard constraint only. The translation from the hard to the soft constraint is straightforward[7].

A simple way of visualizing the imposition of the hard constraint is once again in terms of an external servo-control mechanism. The key is the relation between Eqs. 8.2 and 8.3 which expresses the fact that a *bias* is equivalent to a magnetization. Both are directly expressible in terms of the single parameter a. A mean level of activity can, therefore, be ensured by fixing the magnetization, which can be achieved by a *uniform* external magnetic field. It implies that the mean activity level can be monitored by the injection of current into all the neurons. The precise value of the monitoring field must be determined by the requirement that in addition to all other conditions that have to be satisfied, such as mean-field equations for different types of attractors, the magnetization must have the value corresponding to the required mean activity level.

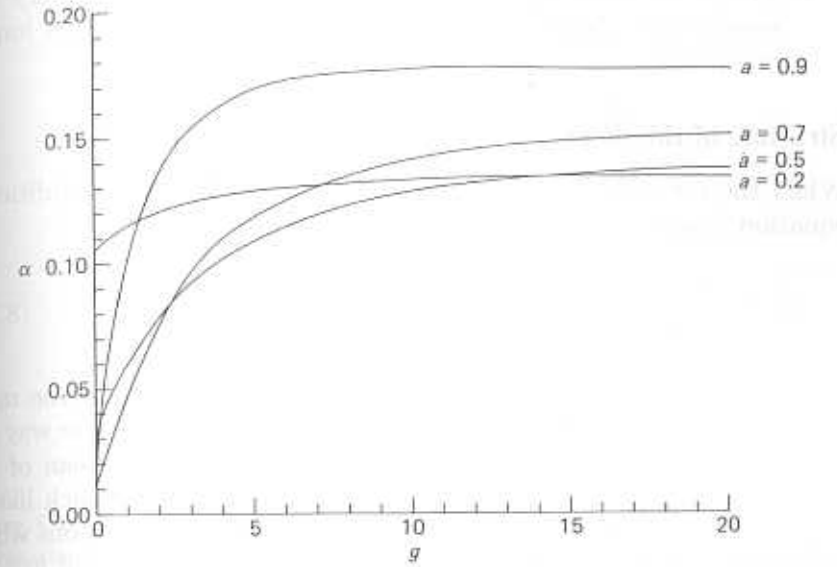


Figure 8.3: Storage capacity α of network storing low activity patterns with softly constrained dynamics vs strength of the constraint g, for several values of the bias a marking the curves[7].

The analysis of the equations reveals the following properties:

- At high noise levels, the system is ergodic, i.e., m^μ=0 for all μ.
- The monitoring field is determined from

$$m_0 = \tanh \beta h_0 = a,$$

  which is just Eq. 8.22 when m^μ=0.
- There is a transition temperature $T_c = (1-a^2)^2$ below which the retrieval states are absolute minima. The suppression of the transition temperature relative to the unbiased network is tantamount to a rescaling of the synaptic efficacies.
- The symmetry $S \to -S$ is broken by the monitoring field. Hence solutions for attractors with a given set of non-zero overlaps do not entail additional solutions in which some overlaps have opposite signs.

- The presence of h_0 suppresses spurious states even at very low noise levels. In particular, the 2- and 3-mixtures are no longer attractors.

Structure of the theory

When the network stores a finite number of patterns, the additional equation is simply

$$m_0 \equiv \frac{1}{N}\sum_{i=1}^{N}\langle S_i\rangle = \left\langle\left\langle \tanh\beta\left(\sum_{\mu=1}^{p} m^{\mu}(\xi_i^{\mu}-a)+h_0\right)\right\rangle\right\rangle \qquad (8.22)$$

where h_0 is the uniform external monitoring field and m_0 is the magnetization per spin. The m^{μ}'s are defined in the appropriate way for biased patterns, Eq. 8.15, and the tanh is the temporal mean of the value of the spin in the presence of the external field h_0, much like in Section 3.4. This equation supplements the mean-field equations which also have to be considered in the presence of the uniform field h_0. For finite-p they are Eqs. 4.40. Despite the fact that the overlaps have been redefined, the combination of modified overlaps and modified storage prescription, Eq. 8.7, preserves the **form** of these equations.

Properties near saturation

The analysis of the network near saturation (when $p = \alpha N$) is again a direct extension of the considerations of Section 6.3.1. Analysis[7] reveals that:

- The storage capacity $\alpha_c(a)$ increases with a for almost the entire interval $0 < a < 1$. In fact, it increases for all $|a| < 0.99$. See e.g., Figure 8.4.
- The storage capacity reaches a maximum at $a \approx 0.925$ where its value is $\alpha_{max} \approx 0.18$, about 30% higher than the unbiased value of 0.138.
- As a approaches unity, the capacity drops to zero. It vanishes according to:

$$\alpha_c \approx \frac{\alpha^*}{|\ln(1-a)|},$$

where α^* is a number of order unity.

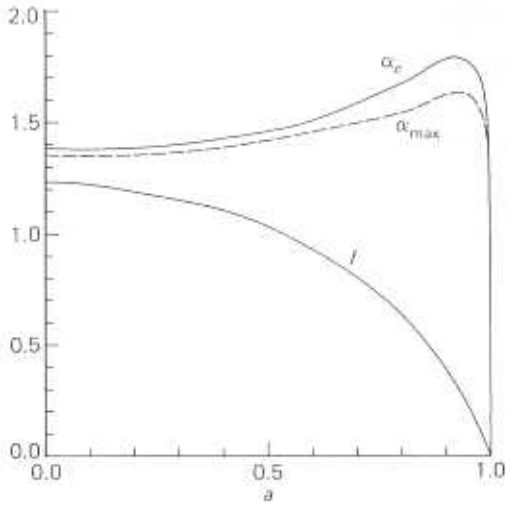


Figure 8.4: The storage capacity α_c, the storage level at which retrievable information is maximized α_{max} and the information content I at saturation vs the bias a. (From ref. [7], by permission.)

This is in sharp contrast to other types of dynamics which allow the storage of an ever increasing number of patterns as the correlations between them increase[9,8,10]. They obtain

$$\alpha_c(a) \approx \frac{1}{(1-|a|)\ln(1-|a|)}.$$

We come back to these recent approaches in Section 8.2.6.

- The information content of the network decreases monotonically as $|a| \to 1$, as is indicated by the curve marked $I(a)$ in Figure 8.4. The estimation of the information content will be taken up in the next section.
- For every value of a there is a storage level α_{max}, at which the *information content* is maximal. One finds that $\alpha_{max}(a) < \alpha_c(a)$, yet it has a similar form. See e.g., Figure 8.4.

Structure of the theory

The equations are just Eqs. 6.31–6.33 with the mere modifications:

- The ξ's have to be replaced by $\xi - a$.
- The overlaps have to be read as in Eq. 8.15.
- The external fields coupled to the patterns can be set to zero and h_0 added to the local field in the arguments of the tanh's.
- The equation for the order-parameter r is changed into

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

These equations are supplemented by the constraint equation for h_0, namely

$$m_0 = a = \left\langle\!\left\langle \tanh\beta\left(\sqrt{r\alpha}z + \sum_{\mu=1}^{s} m^\mu(\xi_i^\mu - a) + h_0\right)\right\rangle\!\right\rangle. \qquad (8.23)$$

Recall that the double angular brackets stand for a double average:

1. Over the discrete s ξ's corresponding to the condensed overlaps;
2. Over the gaussian noise z with zero mean and unit variance, corresponding to the large number of patterns with purely random overlaps. Compare the discussion at the end of Section 6.3.1.

As an example we rewrite the modified equation for the overlaps, Eq. 6.31,

$$m^\nu = \left\langle\!\left\langle (\xi^\nu - a)\tanh\beta\left(\sqrt{\alpha r}\, z + \sum_{\mu=1}^{s} m^\mu(\xi^\mu - a) + h_0\right)\right\rangle\!\right\rangle. \qquad (8.24)$$

The full equations for the retrieval states in the noiseless limit are reproduced in Appendix 8.5.2.

8.2.5 Quantity of information in an ANN with low activity

The increase in storage capacity – number of patterns per neuron – of a network storing and retrieving effectively patterns of low activity raises the question of whether one has also made gains in terms of the *quantity of information* stored per neuron. There are two conflicting trends. On the one hand, the number of patterns increases, on the other each pattern stores less information. As long as the patterns have been completely random, the number of bits in a pattern has been its information content. When the patterns are biased, a significant fraction of their bits can be predicted and hence the amount of information they contain is lower. In the extreme case, when a becomes unity all stored patterns become identical and all their bits become identical. Hence the information could vanish, whatever the number of stored patterns. This intuition will be quantified below.

We have already given away the punch-line in the previous section, namely: in a network that stores biased patterns, has constrained dynamics, and is loaded to saturation, the quantity of information per neuron decreases monotonically with the *bias*. The storage level at which the information content is maximal follows the behavior of the storage capacity, but is always lower in value for the same value of the bias. See e.g., Figure 8.4.

To quantify the information content, it is convenient to proceed with the strictly constrained dynamics. In that case, both stored and retrieved states are in the same subspace of network states. The formalization must capture two essential features:

- The reduced amount of information in the stored patterns.
- The reduction of information due to possible retrieval errors.

The information in any given pattern is just

- minus the logarithm of the probability that it be picked at random from the ensemble of states of which it is a member.

This is nothing but the logarithm of the total number of states in our subspace, which in turn is the logarithm of the total number of states of N neurons $\frac{1}{2}(1+a)N$ of which are active, namely are $+1$. This information, computed per neuron can be written as

$$S(a) = \frac{1}{N}\ln\left(\begin{array}{c} N \\ \frac{1}{2}(1+a)N \end{array}\right). \qquad (8.25)$$

For large N, one can apply Stirling's formula[11]:

$$\ln K! \approx (K + \tfrac{1}{2}) \ln(K + 1) - K$$

to arrive at the information per spin stored in each pattern

$$S(a) = -\tfrac{1}{2}(1 + a) \ln[\tfrac{1}{2}(1 + a)] - \tfrac{1}{2}(1 - a) \ln[\tfrac{1}{2}(1 - a)]. \quad (8.26)$$

Note that this expression has the appropriate limits:

- As $|a| \to 1$ it tends to zero.
- As $a \to 0$, the patterns are uncorrelated, the information per bit is the full $\ln 2$.

To estimate the informational cost of errors we argue as follows: Suppose that retrieval quality (overlap) is $\bar{m}(= \langle\langle \xi^\mu S \rangle\rangle)$. The information that is missing in the retrieved configuration is the *entropy* of the subspace of states which have the same overlap with the pattern number μ. This subspace would, of course, shrink to a single state were there no errors, i.e., if $\bar{m}$=1. This entropy is the logarithm of the total number of states in our restricted space, i.e., with $\frac{1}{2}(1 + a)N$ active neurons, which have an overlap $\bar{m}$ with a given pattern. That number is evaluated in Appendix 8.5.3. The result for the entropy is

$$\begin{aligned} S(\bar{m}, a) \;=\; & -\frac{1}{4}(1 + 2a + \bar{m}) \ln(1 + 2a + \bar{m}) \\ & -\frac{1}{4}(1 - 2a + \bar{m}) \ln(1 - 2a + \bar{m}) - \frac{1}{2}(1 - \bar{m}) \ln(1 - \bar{m}) \\ & +\frac{1}{2}(1 + a) \ln(1 + a) + \frac{1}{2}(1 - a) \ln(1 - a) + \ln 2. \end{aligned} \quad (8.27)$$

The *information content* that can actually be retrieved from this network, storing p biased patterns, is the difference of 8.26 and 8.27 multiplied by the number of patterns p, namely

$$I(\alpha, a) = \frac{pN}{\ln 2}[S(a) - S(\bar{m}, a)] = \frac{\alpha N^2}{\ln 2}[S(a) - S(\bar{m}, a)]. \quad (8.28)$$

Note that the units of information have been normalized to be unity per bit, rather than the canonical $\ln 2$.

A few features of Eq. 8.28 are worth pointing out,

- I has been written as a function of α and a alone because $\bar{m}$ at T=0 is determined, in the mean-field theory, by α and a.
- Any random state in the constrained space has an overlap of $\bar{m} = a^2$ with every pattern. Eq. 8.28 then gives $I = 0$, as it should.
- The retrievable information at saturation,

$$I(a) \equiv \frac{1}{N^2} I(\alpha_c(a), a) \quad (8.29)$$

is plotted in Figure 8.4.

- The retrievable information in the unbiased network, divided by N^2, is obtained by setting a=0, namely

$$I(0) \equiv \frac{1}{N^2} I(\alpha, 0) = \tfrac{1}{2}\alpha[(1 + \bar{m}) \ln(1 + \bar{m}) + (1 - \bar{m}) \ln(1 - \bar{m})]. \quad (8.30)$$

Computed at $\bar{m} = \bar{m}(\alpha_c)$, it is where the curve $I(a)$ intersects the vertical axis in Figure 8.4.

- The retrievable information is not maximized, for a given value of a, where the storage reaches saturation, but at a lower α. The values of α at which it is maximized, $\alpha_{max}(a)$, are shown in Figure 8.4.

8.2.6 More effective storage of low activity (sparse) patterns

The detail with which we have discussed the modified dynamics of an ANN storing low-activity (sparsely coded) patterns provides tools and criteria which are useful, even when assumptions about storage prescriptions and dynamics vary. Much more effective modifications have been recently designed[8,9,10]. We shall concentrate here on the approach of refs. [8,9], rather than on that of Gardner[10], because it is an explicit modification of the Hopfield prescription, whose properties can be unravelled in full detail.[2] It may turn out, though, that Gardner's

[2] It is an amusing aside, worthy of the attention of the historian of science, that the identical approach has been stumbled upon, simultaneously and independently, by

approach, which is more closely related to learning (to be discussed in the next chapter) will turn out to be more natural or to possess other advantages.

The biological way of describing this approach is to go back to the (0,1) description of neural states rather than (–1,+1), as representing passive and active states, respectively. This is, of course, a mere formal transformation, see e.g., Section 1.4.1 or 2.1.2, which by itself could bring about no improvement in performance. The lever is that the transformation between the two sets of variables was accompanied by a shift in the local fields (PSP's). It was assumed, Section 2.1.2, that the local thresholds cancel the local terms which are generated by the transformation, i.e., that

$$T_i = \sum_j J_{ij}. \tag{8.31}$$

If instead T_i is uniform, performance changes rather dramatically:

- It worsens for unbiased patterns[12,13] – the capacity drops by 50%.
- It greatly improves for low levels of activity – the storage capacity **diverges** as $|a| \to 1$.

To provide insight into this situation we shall limit ourselves mainly to considerations of signal-to-noise. The studies have been carried out to much higher sophistication, making full use of the fact that interactions remain symmetric and the thermodynamic apparatus is available. We shall describe this approach in its own notation, providing a running translation (in square brackets) to the notation of the previous discussion.

One stores p patterns of zeros and ones.

The pN bits of the patterns are chosen independently, at random according to the biased distribution

$$\Pr(\eta = 1) = b \ \ [= \tfrac{1}{2}(1+a)]; \qquad \Pr(\eta = 0) = 1 - b \ \ [= \tfrac{1}{2}(1-a)],$$
$$[\eta = \tfrac{1}{2}(\xi + 1)].$$

two groups – one in Moscow[8] and the other in Münich[9]. It has been motivated by two different research projects followed by the two groups. While the Moscow group has been investigating hierarchical data structures, discussed below, the Münich group has been concerned with temporal sequences of sparsely coded patterns.

The unbiased network has $b = \frac{1}{2}$. Low activity is represented by low values of b $[a \to -1]$.

The mean activity in a pattern is

$$A = \frac{1}{N}\sum_{i=1}^{N} \eta_i^\mu = b \ \ [= \tfrac{1}{2}(1+a)]. \tag{8.32}$$

The overlap between any two patterns is

$$\frac{1}{N}\sum_{i=1}^{N} \eta_i^\mu \eta_i^\nu = b^2. \tag{8.33}$$

The synaptic matrix, for $i \neq j$, is taken as

$$T_{ij} = \frac{1}{b(1-b)N}\sum_{\mu=1}^{p} (\eta_i^\mu - b)(\eta_j^\mu - b), \tag{8.34}$$
$$[\eta_i^\mu - b = \tfrac{1}{2}(\xi_i^\mu - a)].$$

For $b = \frac{1}{2}$, this is just the standard synaptic matrix with a=0. Apart from a factor $4b(1-b)$ $[= 1 - a^2]$ this is just the biased synaptic matrix Eq. 8.7. The states of the network, σ_i, are also N-bit words of 0 and 1, $[\sigma = \frac{1}{2}(S+1)]$.

Let us now investigate the stability of the patterns, $S_i = \eta_i^1$ by evaluating the PSP arriving at neuron i, after subtracting a possible uniform threshold, U.

$$\begin{aligned} h_i &= \textstyle\sum_{jj\neq i}^{N} T_{ij}\eta_j^1 - U \\ &= [b(1-b)N]^{-1}\textstyle\sum_{j\neq i}^{N}(\eta_i^1 - b)(\eta_j^1 - b)\eta_j^1 \\ &+ [b(1-b)N]^{-1}\textstyle\sum_{jj\neq i}^{N}\sum_{\mu=2}^{p}(\eta_i^\mu - b)(\eta_j^\mu - b)\eta_j^1 - U. \end{aligned}$$

The three terms are the signal S, the noise R and the threshold. The signal can be easily evaluated in terms of Eq. 8.32 and 8.33. It is

$$S = \eta_i^1 - b \ \ [= \tfrac{1}{2}(\xi_i^1 - a)]. \tag{8.35}$$

Compare Eq. 8.10. The magnitude of the signal is $1 - b$ for η=1 and $-b$ for η=0. Since the stability of the state depends on the stability of

the weaker neurons, it is expedient to choose the threshold to equalize the signals. This choice is

$$U = \tfrac{1}{2} - b$$

and then, for both neural states,

$$|S - U| = \tfrac{1}{2}.$$

Turning to the noise, one notes that it includes a factor

$$\frac{1}{b(1-b)N} \sum_{j}^{N} \sum_{j \neq i} (\eta_j^\mu - b)\eta_j^1$$

which is usually zero for $\mu \neq 1$. But, when p increases with N, the fluctuations, of order $N^{-1/2}$, may add up to overcome the signal. The mean-square noise is computed in Appendix 8.5.4 to be,

$$\rho^2 \equiv \langle (R^2) \rangle \approx b\frac{p}{N} = b\alpha. \tag{8.36}$$

The signal-to-noise ratio is therefore

$$\frac{S}{\rho} \approx \frac{1}{\sqrt{4b\alpha}}.$$

This result is very interesting.

- For unbiased patterns, $b = \frac{1}{2}$, it gives

$$\frac{S}{\rho} \approx \frac{1}{\sqrt{2\alpha}}.$$

This implies that the patterns will be destabilized already at $\alpha = \frac{1}{2}$, one half the capacity of the standard model. Compare Section 6.2.1.

- As the activity goes down, the relative magnitude of the noise decreases and the storage capacity grows as

$$\alpha_c \approx \frac{1}{4b}. \tag{8.37}$$

Mean-field theory along the lines of Chapter 6 has led to the result that[9,8]

$$\alpha_c \approx \frac{1}{b \ln b}.$$

Basically, the effect is that by choosing 0,1 as the neural states one adds a random potential to each neuron relative to the potential it would receive in the standard model, were the cancellation Eq. 8.31 to take place. This random potential acts in two ways:

- It reduces the signal by one half.
- It is correlated with the usual noise, and as the patterns become increasingly biased, this correlation leads to significant cancellations.

The discussion in Section 8.2.2 has indicated that it is not sufficient to mend the signal-to-noise ratio when the patterns are biased. The behavior of spurious states may still cause problems. This has been appreciated[9] and the solution proposed is an additional uniform inhibitory synaptic coupling to all connections in the network, over and above the uniform potential, or threshold, U. This is a soft way of imposing a constraint on the dynamics, as was shown in Section 8.2.3. There is also a fringe benefit. The state of total inactivity becomes an attractor and has been proposed[9] as a cognitive identifier of non-recognition, much like in Shinomoto[14], the expectation being that stimuli which are too far from the memorized patterns flow to this unique, special attractor.

8.3 Hierarchical Data Structures in a Single Network

8.3.1 Early proposals

The idea that hierarchical organization of memorized data is a desirable feature has appeared relatively early in the present phase of interest in neural networks. The fact that the low-lying states in the energy landscape of a spin-glass with long-range interactions – the SK-model – form a hierarchical structure[15] has been very tempting. These states are organized in an *ultrametric* tree, which implies that a *distance* between states can be defined such that any set of three states are the sides of an isosceles triangle with two equal long sides. This distance is 1 minus the overlap between the two states. A set of states with this property can be viewed as the 'leafs' of a tree – the distance between two states being the distance to the nearest common branching point.

Such a structure is an attractive candidate for a scheme classifying categories. Moreover, if the branching points in the tree could also be associated with states, these could be candidates for generalizations.

The standard storage prescription has been interpreted as implying an *instructionist* learning process building memory on a *tabula rasa*. See e.g., Section 4.1.2. In contrast, it was proposed[16] that the initial state may be a state of random synaptic connectivity and as such should behave as a spin-glass. Learning is then a *selectionist* pruning process which eliminates connections based on experience, much in line with the findings on the development of the neuro-muscular region, (see e.g., ref. [17]). If learning is a process by which low-lying attractors are deepened by strengthening certain synapses and eliminating others, the outcome may be a hierarchically organized memory set.

The same objective was also pursued in a more constructive way[18]. The increased familiarity with the relation between the spin-glass ground states and the particular realization of the randomly chosen coupling constants led to a prescription for a hierarchy of attractor states realized by a synaptic matrix constructed in a specific stochastic process.

8.3.2 Explicit construction of hierarchy in a single ANN

We shall start by describing the storage and retrieval of a very simple tree of states, composed of a single parent and one generation of descendants[22]. The parent is an unbiased pattern ξ of ± 1's, hence

$$\langle\langle \xi \rangle\rangle = 0.$$

We then generate the offspring stochastically as a set of p patterns $\{\xi^\mu\}$, each bit of each pattern is chosen, independently, to be either equal or minus the corresponding bit in the parent, according to a given probability distribution[18]. This can be expressed as

$$\xi_i^\mu = \xi_i \eta_i^\mu, \tag{8.38}$$

with

$$\Pr(\eta = \pm 1) = \tfrac{1}{2}(1 \pm a). \tag{8.39}$$

As a consequence, the overlap of two different descendants is

$$\langle \mu | \nu \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu = \frac{1}{N} \sum_{i=1}^{N} \eta_i^\mu \eta_i^\nu = a^2 \tag{8.40}$$

i.e., they are all equi-distant.

To store parent and descendants, the synaptic matrix could be

$$T_{ij} = \frac{1}{N} \xi_i \xi_j \left(1 + \frac{1}{\Delta} \sum_{\mu=1}^{p} (\eta_i^\mu - a)(\eta_j^\mu - a) \right), \tag{8.41}$$

which should be compared to Eq. 8.7. The main differences are that a 'ferromagnetic' term $-\xi_i \xi_j$ - has been added to the matrix storing the p patterns, and a coefficient $1/\Delta$ has been introduced. The stability of the memorized patterns is verified by a calculation of the local fields produced in the states ξ and ξ^1, for the choice $\Delta = 1 - a^2$. The first is

$$h_i(\xi) = \xi_i \left(1 + \frac{1}{(1-a^2)N} \sum_{\mu=1}^{p} \sum_{j\, j \neq i}^{N} (\eta_i^\mu - a)(\eta_j^\mu - a) \right). \tag{8.42}$$

The signal is 1. The noise vanishes and has an RMS of α, as is easily evaluated following the examples in Appendix 8.5.4. Hence, the parent is stable if $\alpha \ll 1$. For the descendants,

$$h_i(\xi^1) = \xi_i \left(a + \frac{1}{(1-a^2)N} \sum_{\mu=1}^{p} \sum_{j\, j \neq i}^{N} (\eta_i^\mu - a)(\eta_j^\mu - a)\eta_j^1 \right). \tag{8.43}$$

In the second term on the right hand side we separate the $\mu = 1$ contribution from the rest, which is noise with vanishing mean and RMS of α. The signal is just

$$S = \xi_i \left(a + \frac{1}{(1-a^2)N} \sum_{j\, j \neq i}^{N} (\eta_i^1 - a)(\eta_j^1 - a)\eta_j^1 \right) = \xi_i \eta_i^1 = \xi_i^1$$

which shows that the descendants are also stable for small α.

Two further conclusions follow

- With the particular choice of $\Delta = 1 - a^2$, parent and descendants have a signal of the same magnitude, and hence the same energy.
- The calculation of the noise indicates that the storage capacity for the retrieval of all $p+1$ memories is the same as for the usual, unbiased patterns. This is confirmed by detailed analysis[22].

The degeneracy between the parent and the descendants is lifted when the value of Δ is changed. In fact, this parameter can vary in the interval

$$1 - a^2 \leq \Delta < \frac{(1-a^2)(1+a)}{a} \tag{8.44}$$

and still have all the stored patterns stable.

- The parent is the absolute energy minimum and the descendants are p satellite meta-stable states[22].

8.3.3 Properties of hierarchy in a single network

The detailed mean-field analysis of the network described in the previous section reveals the following properties:

- The degenerate network – $\Delta = 1 - a^2$ – is equivalent to the unbiased network with $p+1$ stored patterns. It has the same storage capacity and the same *phase-diagram*.
- When the degeneracy is lifted, there are two values of the storage capacity of the noiseless network, i.e., $\alpha_c^{(1)}(\Delta)$ and $\alpha_c^{(2)}(\Delta)$. Above the first there are no retrieval states; between the two only the parent is an attractor and below $\alpha_c^{(2)}$ both parent and descendants are retrievable. For $\Delta = 1 - a^2$ the two values coincide. The two storage capacities are plotted in Figure 8.5 as a function of Δ.
- When the degeneracy is lifted there are two separate critical noise levels for retrieval of parent and descendants – $T_M^{(1)}(\Delta, a)$, $T_M^{(2)}(\Delta, a)$. The fact that the energy of the satellites is higher implies that

$$T_M^{(1)}(\Delta, a) > T_M^{(2)}(\Delta, a).$$

8.3.4 Prosopagnosia and learning class properties[3]

The increase in the level of noise leads, therefore, to a situation in which memory which has stored specific objects together with their class representatives can recall only the 'generic class' to which the objects belong. Such is the situation between the lines $T_M^{(1)}$ and $T_M^{(2)}$, for

[3] I owe the ideas of this section to Prof. M.A. Virasoro.

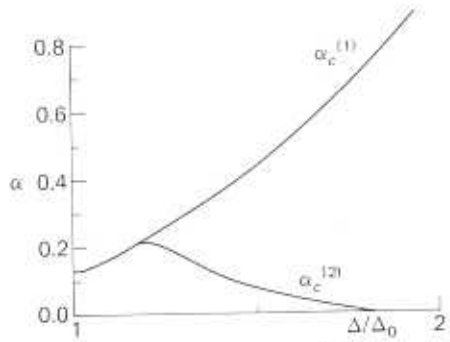


Figure 8.5: The storage capacities of parent ($\alpha_c^{(1)}$) and descendants ($\alpha_c^{(2)}$) vs the relative synaptic strength parameter Δ. (From ref. [22], by permission).

the fast noise, and between $\alpha_c^{(1)}$ and $\alpha_c^{(2)}$ in Figure 8.5, for overloading. Synaptic deterioration can be described as in Section 7.1.2, where we have seen that it acts very much like fast noise. This sounds just like *prosopagnosia*[19]. Without entering into a detailed account of this fascinating mental disturbance we try to communicate its flavor with two quotes:

> The fifth Marquess of Salisbury found it hard to recognize the faces of his fellow men, even his relations, if he met them in unexpected circumstances. Once, standing behind the throne at a Court ceremony, he noticed a young man smiling at him. 'Who is my young friend?' he whispered to a neighbor. 'Your oldest son,' the neighbor replied.
>
> Lord David Cecil[19]

> Patients with prosopagnosia only, know that a face is a face and name it as such... Yet they are unable to recognize a given familiar face; i.e., they do not know to whom it belongs and consequently are unable to name it. Recognition of the generic class to which the stimulus belongs presents no difficulty, but recognition of an individual member of that class, whose identity has previously been learned, is impaired.[19]

Thus the symptoms of random synaptic deterioration in an ANN would appear like those of *prosopagnosia*.

But what is perhaps even more surprising is that an inverse phenomenon is also to be expected. If a network learns a set of correlated patterns, by any of the algorithms to be described in the next section, it would tend to learn their *generic classes* as well[20]. This is a consequence of Gardner's work on optimal storage capacity[21], discussed in Section 6.2.3. In fact, one of the consequences of searching in the space of coupling constants for a set of biased patterns is that typically the ancestor is also an attractor, and it is more stable than the daughter states. This is just as in the model discussed in Section 8.3.2 with parameters satisfying Eq. 8.44, and hence learning classes this way organizes a network that is susceptible to *prosopagnosia*. In a sense, this goes against the grain of our position on spurious states. The class representatives are, strictly speaking, spurious. Yet here their appearance seems desirable. But, of course, something rather special would have to happen for the network to know in the learning process, which groups of individuals to classify together and which belong to different 'ancestors'.

8.3.5 Multy-ancestry with many generations

The hierarchy described in the previous section is rather simple. In essence it is a mere set of biased patterns superposed on top of a ferromagnetic interaction.[4] It becomes less trivial when the number of ancestors is increased or with an increasing number of levels in the trees. Feigelman and Ioffe[22] have constructed and analyzed this extension. It is naturally done in two steps:

Multi-ancestors and two levels

Instead of the single ancestor ξ of the previous section, we now take p patterns $\{\xi^\mu\}$ with $\mu = 1, \ldots, p$. Each of these will have q descendants $\{\xi^{\mu,\nu}\}$. In general, the *branching ratio* can vary from parent to parent, i.e., $q = q_\mu$. Moreover, the stochastic process generating the branching at each ancestor can vary with the ancestor, namely the probability (bias a) for transmitting the structure of the ancestor to its descendant may depend on μ, i.e., $a = a_\mu$. See e.g., Figure 8.6(a).

[4]The reader can formally verify this statement by performing a Mattis transformation Eq. 4.31 using the parent.

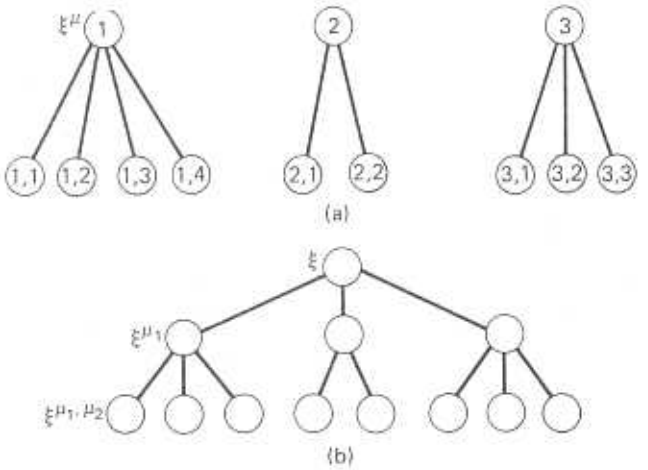


Figure 8.6: Ancestors and descendants. (a) A two level tree with different branching ratios. (b) Progenitor and two generations of descendants. Items at the bottom level are sisters or cousins.

The generalization of the stochastic construction, Eqs. 8.38 and 8.39 is just

$$\xi_i^{\mu,\nu} = \xi_i^\mu \eta_i^{\mu,\nu}, \tag{8.45}$$

with

$$\Pr(\eta^{\mu,\nu} = \pm 1) = \tfrac{1}{2}(1 \pm a_\mu).$$

Now the overlap of two different descendants can have two values: if they belong to the same bunch (same parent μ), then

$$\langle \mu, \nu | \mu, \lambda \rangle \equiv \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu,\nu} \xi_i^{\mu,\lambda} = \frac{1}{N} \sum_{i=1}^{N} \eta_i^{\mu,\nu} \eta_i^{\mu,\lambda} = a_\mu^2, \tag{8.46}$$

i.e., they are all equi-distant. If they belong to two different groups and the parents are uncorrelated, then

$$\langle \mu, \nu | \rho, \lambda \rangle = 0.$$

At this stage we still have ultrametricity. If we choose any three states which are either in one bunch or in three separate bunches, then they are equi-distant. If two are in one bunch and the third is in a separate one, then two distances are maximal, =1, and one distance is less than 1.

The synaptic matrix is now

$$T_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu\left(1+\frac{1}{\Delta}\sum_{\nu=1}^{q_\mu}(\eta_i^{\mu,\nu}-a_\mu)(\eta_j^{\mu,\nu}-a_\mu)\right). \qquad (8.47)$$

The properties of this network can be directly deduced from those of the simpler one described in the previous sections[22].

- When $\Delta = 1-a^2$, the degeneracy between parents and descendants sets in. The storage capacity is the familiar $\alpha \approx 0.14$, where α refers to the total number of memorized patterns, i.e.,

$$\alpha N = p + \sum_{\mu=1}^{p} q_\mu.$$

  All these states become attractors, at the same temperature.

- If $\Delta > 1-a^2$ the degeneracy is lifted and the parents become lower in energy than the descendants.

- The total storage capacity remains the same, but the ancestors appear first, at higher loading levels, and then the detailed descendants become retrieval states at lower loading levels.

- The process of generalization becomes much richer as each bunch of detailed patterns obtains its own generalizing state.

Multi-level hierarchy – structure only

For clarity of presentation, we shall restrict ourselves to a single progenitor pattern ξ with two generations of descendants. The generalization to several independent progenitors as well as to a higher number of successive generations should be self-evident. The one stage hierarchy of Section 8.3.2 will be denoted by

$$(\xi;\ \{\xi^{\mu_1}\};\ \{\eta^{\mu_1}\};\ a;\ p),$$

which are, respectively, the ancestor; the descendant patterns; the transforming biased patterns; the bias in the stochastic generation; the number of descendants.

To construct the second generation, we need the set

$$(\{\eta^{\mu_1,\mu_2}\};\ a_{\mu_1};\ p_{\mu_1}),$$

with p_{μ_1} N-bit words generated for each first generation state μ_1 according to:

$$\Pr(\eta^{\mu_1,\mu_2}) = \tfrac{1}{2}(1+a_{\mu_1}). \qquad (8.48)$$

The patterns of second generation are

$$\xi_i^{\mu_1,\mu_2} = \xi^{\mu_1}\eta^{\mu_1,\mu_2}. \qquad (8.49)$$

Their total number is

$$p_{total} = \sum_{\mu_1=1}^{p} p_{\mu_1}.$$

A three-levelled tree is presented in Figure 8.6(b).

Consider the correlations between states in the second generation. There are two possible situations, either the two patterns are sisters (have a common first ancestor), or they are cousins (with second common ancestor). In the first case, the overlap is

$$\langle\mu_1,\mu_2|\mu_1,\lambda_2\rangle = a_{\mu_1}^2, \qquad \mu_2 \neq \lambda_2.$$

For two cousins one has

$$\langle\mu_1,\mu_2|\lambda_1,\lambda_2\rangle = a_{\mu_1}a_{\mu_2}a^2, \qquad \lambda_1 \neq \mu_1.$$

If the bias used for generating the second generation is uniform, i.e., $a_{\mu_1} = b$, then the tree will be ultra-metric. Any set of three patterns will be:

- 3 sisters, with equal mutual distances of $1-b^2$;

- 3 cousins, with equal mutual distances of $1-a^2b^2$;

- 2 sisters and a common cousin. The distance between sisters is $1-b^2$ and between the two sisters and their cousin is $1-a^2b^2$. Since $|b|<1$, the two equal distances are longer than the third one.

This type of tree structure is much in the spirit of Virasoro's proposal[18].

Finally, the synaptic structure which stabilizes this set of patterns is

$$T_{ij} = \frac{1}{N}\left(\xi_i\xi_j + \frac{1}{\Delta_1}\sum_{\mu_1=1}^{p}\delta\xi_i^{\mu_1}\delta\xi_j^{\mu_1} + \frac{1}{\Delta_2}\sum_{\mu_1=1}^{p}\sum_{\mu_2=1}^{p_{\mu_1}}\delta\xi_i^{\mu_1,\mu_2}\delta\xi_j^{\mu_1,\mu_2}\right).$$

supplemented by

$$\delta\xi_i^{\mu_1} = \xi_i^{\mu_1} - a\xi \tag{8.50}$$

$$\delta\xi_i^{\mu_1,\mu_2} = \xi_i^{\mu_1,\mu_2} - a_{\mu_1}\xi_i^{\mu_1}. \tag{8.51}$$

One can now embark on a classification of the various breakings of the full degeneracy, $\Delta_1 = 1 - a^2$ and $\Delta_2 = 1 - a_1^2$, at which patterns of all three levels have the same energy. Here we only indicate the richness of the possibilities and leave further investigation to the reader who may imagine an application worthy of the required effort.

8.4 Hierarchies in Multi-ANN: Generalization First

8.4.1 Organization of the data and the networks

We now turn to a construction which does allow for generalities first[23]. The data structure is exactly the same as described in Section 8.3.5, except that now the patterns of each level in the hierarchy are stored in a separate network. The top network stores a set of progenitor patterns $\{\xi^\mu\}$, which may be chosen to have a common bias a_1. The next network stores the progeny of first generation $\{\xi^{\mu,\nu}\}$, produced from their ancestors by the process Eqs. 8.49 and 8.48. Each bunch of sisters in the second network, with a fixed value of μ, is generated from ξ^μ with its own bias a_μ and comprises p_μ patterns. This is depicted schematically under STORAGE in Figure 8.7. From each of the members of the second generation, one can proceed to generate a bunch of daughter states, by iterating the above procedure, to form the third generation. Its states will be stored in a third ANN, etc.

The various correlations in the two-level hierarchy consist of the following items

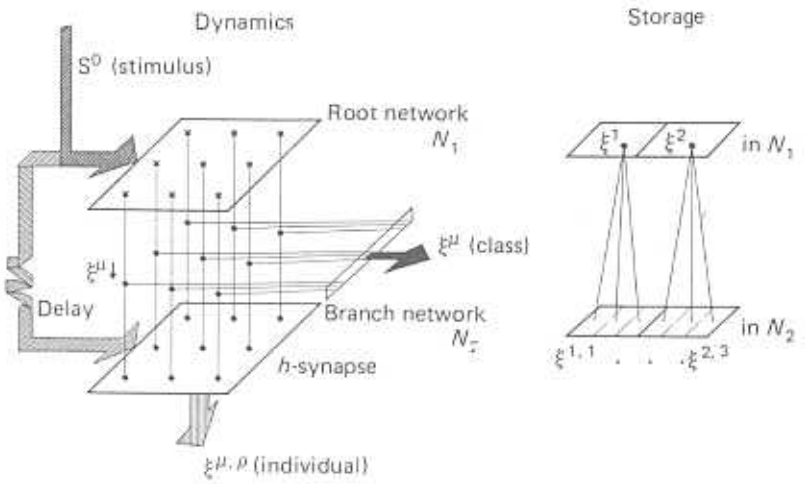


Figure 8.7: Schematic representation of multi-ANN for the storage and retrieval of a two-level hierarchy

- Mean activities:
  - In the top network:
    $$A_1 = \tfrac{1}{2}(1 + a_1).$$
  - In the lower network:
    $$A_\mu = \tfrac{1}{2}(1 + \langle\langle \xi^{\mu,\nu}\rangle\rangle) = \tfrac{1}{2}(1 + a_1 a_\mu).$$
- Intra-level correlations:
  - In top network:
    $$\langle\langle \xi^\mu \xi^\nu\rangle\rangle = a_1^2 \qquad \text{for } \mu \neq \nu.$$
  - In lower network:
    $$\langle\langle \xi^{\mu,\nu}\xi^{\mu,\rho}\rangle\rangle = a_\mu^2; \qquad \langle\langle \xi^{\mu,\nu}\xi^{\lambda,\rho}\rangle\rangle = a_1^2 a_\mu^2. \tag{8.52}$$
    The left expression is the correlations between sisters and the right one is between cousins.

- Inter-level correlations:

$$\langle\langle \xi^{\mu,\nu}\xi^{\mu}\rangle\rangle = a_\mu; \qquad \langle\langle \xi^{\mu,\nu}\xi^{\rho}\rangle\rangle = a_1^2 a_\mu. \tag{8.53}$$

  On the left is the correlation of parent-daughter and on the right is the correlation of uncle-niece.

The storage of the patterns in the two networks is designed to ensure the stability of the biased patterns, much as in Section 8.2.2, namely in the top network the synapses are

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}(\xi_i^\mu - a_1)(\xi_j^\mu - a_1)$$

and in the lower one

$$J_{ij} = \frac{1}{N}\sum_{\mu,\nu}(\xi_i^{\mu,\nu} - a_\mu\xi_i^\mu)(\xi_j^{\mu,\nu} - a_\mu\xi_i^\mu). \tag{8.54}$$

The two networks are interconnected by synapses which connect each neuron in the top network to one in the lower one, afferently. These are unidirectional synapses. They are represented by the down-going dots in Figure 8.7, under DYNAMICS. Their role is to translate the network state in the top network to a set of afferent PSP's in the lower network.

8.4.2 Hierarchical dynamics

Finally we come to the description of the dynamics of the network. The input stimulus, which is associated with the detailed pattern – it is a noisy $\xi^{\mu,\nu}$ – enters both networks in the hierarchy. This is represented by the stream S^0 on the left of Figure 8.7. It is not unlike the kind of paths from the thalamus to different areas of cortex which are supposed to be organized hierarchically, as traced empirically by Merzenich, Kaas, Van Essen and others[2,24,25] and theoretically by Ballard[3]. See e.g., Figure 8.8. When it enters the top network, with the class representatives ξ^μ, it will retrieve just that pattern since according to Eq. 8.53 the correlation with it is a_μ while with any other ancestor it is $a_1^2 a_\mu$, which is much smaller. If there are correlations in the top network, i.e., $a_1 > 0$, there will have to be an imposed constraint, as in Section 8.2.3, to avoid spurious states.

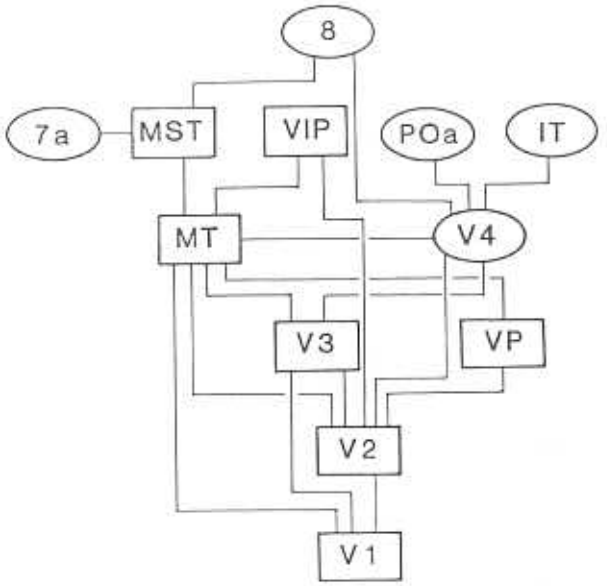


Figure 8.8: Organization of macaque monkey cortical areas into a functional hierarchy. Arrows denote direction of hierarchy. (From ref. [25](b), by permission.)

When retrieval has taken place in the top network, N_1 in Figure 8.7, the state ξ^μ is projected down to the lower network N_2 and on the way it can be read for action on the level of the generalizations. This is represented by the arrows moving to the right in the figure. The h-synapses, connecting the two networks, convert the retrieved state into a set on afferent PSP's in N_2, of the form

$$h_i^{(2)} = h\xi_i^\mu.$$

It has been shown[23] that there is a range of values of the synaptic strengths h for which the effect of these PSP's is to constrain the dynamics of the lower network to states which obey

$$\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu S_i = a_\mu.$$

This constraint, when operating in conjunction with the synaptic matrix Eq. 8.54, ensures that the stimulus that has arrived in the second network by a parallel route will retrieve the associated detailed pattern.

8.4.3 Hierarchy for image vector quantization

A practical application of multi-ANN hierarchical memory[26] is motivated by the following consideration: It is often the case that one is required to store and retrieve rapidly a large quantity of patterns, a quantity that is significantly larger than the number of bits in each of the words to be stored. In most hardware implementations to date, the number of neurons that could be engineered into a fully connected network has been rather limited – 50 to 100. This may be ample as a number of bits, but, it could not possibly store and retrieve more than twice that many words, which is low. The attraction of insisting on an implementation in neural networks is the speed of retrieval, 1-2 microseconds, which may allow rather massive tasks to be performed in real time by very simple devices. See e.g., Chapter 10.

Suppose that we have a set of p progenitor states of N bits each and from each of those we construct an equal number q of descendants. Each bunch of sisters we now store in a separate network of N bits. Thus, we are storing pq patterns of N bits in $p+1$ networks. Since both p and q can be of order N, we can store $O(N^2)$ N-bit patterns this way. As an example, consider ANN's with 20 neurons, in which one can store 3 patterns *á la* Hopfield, or as many as 40 patterns *á la* Gardner. One can store as many as 400 patterns of 20 bits each in 11 networks of 20 neurons.

Retrieval is performed by starting with a stimulus which is sent simultaneously to all $p+1$ networks. Some pattern is retrieved in each one of them, if the synapses are symmetric. This stage takes the usual, electronic, minuscule time of 1-2 μsec. The retrieval in the top network serves as an indicator for selecting the appropriate answer, out of the p retrievals on the lower level.

The same idea can be used in a more general context. The special hierarchy borrowed from the previous sections has been used only because of familiarity. We can start with an arbitrary set of patterns from which a hierarchy is constructed in the following way: The patterns are grouped by Hamming distance into a number of groups that is about the square root of the total number of patterns. The grouping is reshuffled until they reach a situation in which the distances inside each group are smaller than the inter-group distances. Then, a representative is chosen from each group to serve in the upper network.

The context can be further enlarged by considering data compression in image transmission. Due to the regularity of normal visual images, the detailed information in a projected image is highly redundant and a compressed version provides a satisfactory reproduction. An image is divided into small regions, each a 4×4 pixel 'card', and each pixel carries, usually, 256 levels of gray. All together, one would need $16 \times 8 = 128$ bits to communicate all the possible information in a 'card'. It turns out that a code-book of 8,000 out of the 2^{128} possibilities suffices for a good reproduction of an image. One must, therefore, find the code word corresponding to every 'card' data. This is a task of associative memory in which one must be able to retrieve one of 8,000 patterns. For television broadcasting, the time must be of the order of 1μsec. To store that number of patterns in a single network will be a waste, because one does not need anywhere near that many bits in each word. Moreover, to implement a hardware network storing so many patterns is extremely difficult. All of this makes it a natural task for a hierarchical neural network.

8.5 Appendix: Technical Details for Biased Patterns

8.5.1 Noise estimates for biased patterns

Starting from Eq. 8.11, we proceed to calculate the mean square of the noise when the network, storing biased patterns, is in a pattern.

$$\langle\langle R^2\rangle\rangle = \left\langle\left\langle \frac{1}{N^2}\sum_{\mu,\nu=2}^{p}(\xi_i^\mu - a)(\xi_i^\nu - a)\sum_{j,k}(\xi_j^\mu - a)\xi_j^1(\xi_k^\nu - a)\xi_k^1\right\rangle\right\rangle .$$

Note first that the average on the right hand side vanishes if $\mu \neq \nu$ because of the first two factors in the sum, and it vanishes unless $j = k$ because of the third and fifth factors. We therefore have

$$\langle\langle R^2\rangle\rangle = \langle\langle \frac{1}{N^2}\sum_{\mu=2}^{p}\sum_{j\neq i}^{N}(\xi_i^\mu - a)^2(\xi_j^\mu - a)^2\rangle\rangle .$$

Because $j \neq i$, the averages factor and we can write:

$$\langle\langle R^2\rangle\rangle = \frac{1}{N^2}\sum_{\mu=2}^{p}\sum_{j\neq i}^{N}\left\langle\left\langle(\xi_i^\mu - a)^2\right\rangle\right\rangle\langle\langle(\xi_j^\mu - a)^2\rangle\rangle .$$

But,

$$\langle\langle(\xi_i^\mu - a)^2\rangle\rangle = \tfrac{1}{2}(1+a)(1-a)^2 + \tfrac{1}{2}(1-a)(1+a)^2 = 1-a^2$$

and hence,

$$\langle\langle R^2\rangle\rangle = \frac{1}{N^2}\sum_{\mu=2}^{p}\sum_{j\neq i}^{N}(1-a^2)^2$$

which is just Eq. 8.12.

8.5.2 Mean-field equations in noiseless biased network

The same technical steps which have led us to Eq. 6.38 imply for the overlap in the retrieval states:

$$m = \tfrac{1}{2}(1-a^2)\left[\mathrm{erf}\left(\frac{(1-a)m+h_0}{\sqrt{2\alpha r}}\right) + \mathrm{erf}\left(\frac{(1+a)m-h_0}{\sqrt{2\alpha r}}\right)\right]. \quad (8.55)$$

Again, as in Section 6.3.3, in this limit $q \to 1$ and this variable is replaced by $C = \lim_{T\to 0}(1-q)/T$. The parameter C is then expressed in terms of r and the two equations become one, corresponding to Eq. 6.43. Here it reads:

$$\begin{aligned} C &= \frac{r^{-\frac{1}{2}}-1}{1-a^2} \\ &= \frac{1}{\sqrt{2\pi\alpha r}}\left[(1+a)\exp\left(-\frac{[(1-a)m+h_0]^2}{2\alpha r}\right)\right. \\ &+ \left.(1-a)\exp\left(-\frac{[(1+a)m-h_0]^2}{2\alpha r}\right)\right]. \end{aligned} \quad (8.56)$$

And finally, the equation of the constraint reads

$$a = \tfrac{1}{2}(1+a)\mathrm{erf}\left(\frac{(1-a)m+h_0}{\sqrt{2\alpha r}}\right) - \tfrac{1}{2}(1-a)\mathrm{erf}\left(\frac{(1+a)m-h_0}{\sqrt{2\alpha r}}\right). \quad (8.57)$$

8.5.3 Retrieval entropy in biased network

For a state to have the overlap $\bar{m}$ with a given pattern, it must be the same as that pattern except that $\frac{1}{2}(1-\bar{m})N$ spins are reversed. But

in order to comply with the constraint on the dynamics, the flipped spins must total to zero. In other words, one half of the spins flipped away from the pattern must go from +1 to −1 and the other half from −1 to +1. We must, therefore, evaluate the number of different ways of selecting (for flipping) $\frac{1}{4}(1-\bar{m})N$ of the $\frac{1}{2}(1+a)N$ +1's, and at the same time of selecting the same number of spins out of the $\frac{1}{2}(1-a)N$ −1's, and then calculate its logarithm to obtain the missing information due to the errors. This number is

$$S(\bar{m},a) = \frac{1}{N}\ln\left[\begin{pmatrix}\frac{1}{2}(1+a)N\\ \frac{1}{4}(1-\bar{m})N\end{pmatrix}\begin{pmatrix}\frac{1}{2}(1-a)N\\ \frac{1}{4}(1-\bar{m})N\end{pmatrix}\right]. \quad (8.58)$$

Using Stirling's expansion again, one finds[7] Eq. 8.27.

8.5.4 Mean-square noise in low activity network

The noise term is

$$R = \frac{1}{b(1-b)N}\sum_{jj\neq i}^{N}\sum_{\mu=2}^{p}(\eta_i^\mu - b)(\eta_j^\mu - b)\eta_j^1.$$

Because $\mu \neq 1$ and $j \neq i$ the average of its square can be factorized as follows:

$$\begin{aligned}\langle\langle R^2\rangle\rangle = \frac{1}{[b(1-b)N]^2}\sum_{j,k\neq i}^{N}\sum_{\mu,\nu=2}^{p}&\langle\langle(\eta_i^\mu - b)(\eta_i^\nu - b)\rangle\rangle \\ &\langle\langle(\eta_j^\mu - b)(\eta_k^\nu - b)\rangle\rangle\langle\langle\eta_j^1\eta_k^1\rangle\rangle. \end{aligned} \quad (8.59)$$

Using Eqs. 8.39 and 8.40, one finds

$$\langle\langle(\eta_i^\mu - b)(\eta_i^\nu - b)\rangle\rangle = b(1-b)\delta_{\mu\nu}.$$

This converts the second factor into

$$\langle\langle(\eta_j^\mu - b)(\eta_k^\mu - b)\rangle\rangle = b(1-b)\delta_{jk}$$

and the third factor is simply

$$\langle\langle\eta_j^1\eta_j^1\rangle\rangle = \langle\langle\eta_j^1\rangle\rangle = b.$$

When N and p are large and these expressions are substituted in Eq. 8.59, they lead to Eq. 8.36.

Bibliography

[1] D.J. Willshaw, O.P. Buneman and H.C. Longuet-Higgins, Non-holographic associative memory, *Nature*, **222**, 960(1969).

[2] M.M. Merzenich and J.H. Kaas, Principles of organization of the sensory-perceptual systems in mammals, *Progress in Psychobiology and Physiological Psychology*, **9**, 1(1980).

[3] D.H. Ballard, Cortical connections and parallel processing: structure and function, *The Behavioral and Brain Sciences*, **9**, 67(1986).

[4] J.A. Feldman and D.H. Ballard, Connectionist models and their properties, *Cognitive Science*, **6**, 205(1982).

[5] J.A. Fodor, *The Modularity of Mind* (MIT Press, Cambridge, Mass., 1983).

[6] J. Buhmann and K. Schulten, Influence of noise on the function of a "physiological" neural network, *Biol. Cybern.*, **56**, 313(1987).

[7] D.J. Amit, H. Gutfreund and H. Sompolinsky, Information storage in neural networks at low levels of activity, *Phys. Rev.*, **35A**, 2293(1987).

[8] M.V. Tsodyks and M.V. Feigel'man, The enhanced storage capacity in neural networks with low activity level, *Landau Institute, Moscow, preprint* (1988).

[9] J. Buhmann, R. Divko and K. Schulten, Associative memory with high information content, *Technische Universität, Münich, preprint* (1988).

[10] E. Gardner, The space of interactions in neural network models, J. Phys., **21A**, 257(1988).

[11] E.T. Whittaker and G.N. Watson, *A Course of Modern Analysis* (Cambridge University Press, London, 1927), Section 12.33.

[12] G. Weisbuch and F. Fogelman-Soulié, Scaling laws for the attractors of Hopfield networks, *J. Physique Lett.*, **2**, 337(1985).

[13] A.D. Bruce, E.J. Gardner and D.J. Wallace, Dynamics and statistical mechanics of the Hopfield model, *J. Phys.*, **A20**, 2909(1987).

[14] S. Shinomoto, A cognitive associative memory, *Biol. Cybern.*, **57**, 197(1987).

[15] M. Mezard, G. Parisi and M. Virasoro, *The Replica Method and Beyond* (World Scientific, Singapore, 1987).

[16] G. Toulouse, S. Dehaene and J.P. Changeux, Spin glass theory of learning by selection, *Proc. Natl. Acad. Sci. (USA)*, **83**, 1695(1986).

[17] J.-P. Changeux, *Neuronal Man* (Oxford University Press, NY, 1986)

[18] M. Virasoro, Ultrametricity, Hopfield model and all that, in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman Soulié and G. Weisbuch eds. (Springer-Verlag, Berlin, 1986) and N. Parga and M. Virasoro, The ultrametric organization of memories in a neural network, *J. Physique*, **47**, 1857(1986).

[19] A.R. Damasio, H. Damasio amd G.W. Van Hoesen, Prosopagnosia: Anatomic basis and behavioral mechanisms, *Neurology*, **32**, 331(1982).

[20] M.A. Virasoro, The effect of synapses destruction on categorization in neural networs, *Europhys. Lett.*, **7**, 293(1988).

[21] E. Gardner, Maximum storage capacity in neural networks, *J. Phys.*, **A19**, L1047(1986).

[22] M.V. Feigelman and L.B. Ioffe, The augmented models of associative memory: asymmetric interaction and hierarchy of patterns, *Int. Jour. of Mod. Phys.*, **B1**, 51(1987).

[23] H. Gutfreund, Neural networks with hierarchically correlated patterns, *Phys. Rev.*, **A37**, 570(1988).

[24] R.E. Weller and J.H. Kaas, Cortical and subcortical connections of the visual cortex in primates, in *Multiple Visual Areas*, Vol. 2, C.N. Woolsey ed. (Humana Press, Atlantic Hillside NJ, 1981).

[25] (a)D.C. Van Essen and J.H.R. Maunsell, Hierarchical organization and functional streams in the visual cortex, *Trends in Neurosciences*, **6**, 370(1983); (b) D.C. Van Essen, Functional organization of primate visual cortex, in *The Cerebral Cortex* (Plenum Press, NY, 1985).

[26] L.D. Jackel, R.E. Howard, J.S. Denker, W. Hubbard and S.A. Solla, Building a hierarchy with neural networks: An example – image vector quantization, AT&T Technical Memorandum, April, 1987.

9

Learning

9.1 The Context of Learning

9.1.1 General comments and a limited scope

It would have been wonderful if some distinguished thinker had written: 'Thinking about learning is a real headache' and we could have used this phrase as a quote. A glimpse at the type of difficulties that present themselves when a new experience should lead to learning is expressed by Minsky[1] as follows:

> Which agents could be wise enough to guess what changes should then be made? The high level agents can't know such things; they scarcely know which lower-level processes exist. Nor can lower-level agents know which of their actions helped us to reach our high-level goals; they scarcely know that higher-level goals exist. The agencies that move our legs aren't concerned with whether we are walking toward home or toward work – nor do the agents involved with such destinations know anything of controlling individual muscle units...[p.75]

Hebb[2] adds that 'the more we learn about the nature of learning, the farther we seem to get from being able to give firm answers.'

The scope is truly awesome and very soon it overlaps the deep controversy between *innateness* and *behaviorism*. There are several

courses of action which avoid, at this stage, the full dimension of the issue:

- To consider artificial devices which can serve as associative memories with increasingly complex performances and which are decreed by the designer, and to restrict the question of learning to that of training such machines to their final, known performing organization.
- To accept a biologically constrained point of view, namely that '...learning does not define a single kind of business; among the various kinds can already be distinguished habituation, sensitization, imprinting, one-trial learning, classical conditioning, instrumental conditioning, and place learning...'[3] [p. 368].
- To search for ways in which coefficients (synaptic efficacies) can be adjusted, based on 'behavioral' maneuvers, keeping in mind 'that significant learning at a significant rate presupposes some significant prior structure. Simple learning schemes based on adjusting coefficients can indeed be practical and valuable when the partial functions are reasonably matched to the task...'[4].

Here we will deal with a combination of the first and the third approaches. This is quite natural since any progress on the first can be adopted as a hypothesis for learning in specialized areas of cortex and vice versa. We will describe some of the ideas and mechanisms which are being proposed to allow for the modification of the synaptic matrix – the repository of memory in the paradigm under discussion. The synaptic efficacies prescribed in Section 4.2.2 have been motivated by reference to Hebb's learning by synaptic plasticity due to local coincident activity of pre- and post-synaptic neurons. But there was no learning, despite the fact that the prescription for the synaptic matrix has been often referred to in recent literature as the 'learning rule'. The least we can say about learning is that:

- It is a dynamical process in which the *organization* of the network (the structure of its connectivity) is modified on the basis of the interaction of the network with external stimuli.

Note that in the entire discussion leading to this point, the only ingredient of *organization* in the ANN, as such, is the connectivity of

the network. A network comprising a dynamical element which allows for the modification of its connectivity is sometimes referred to as a self-organizing network.

One of the most important things we will avoid is the learning, or the prior design, of the communication of the network with the world. This is certainly a major difficulty for any behavioristic scheme. After all, to paraphrase Minsky, to be rewarded for some response, the system must have learned to respond in a rewardable way in the first place. This does not mean, of course, that our learning networks do not receive external stimuli. What it does mean is that

- the input and output channels are not modified in the learning process;
- what is presented to the network is to be learned, as a memory, and there is no independent input which rewards or punishes to influence alternative reactions to different learned inputs.

This may appear severely limited, but it is not completely unreasonable when learning is viewed as a process in which various parts of memory are loaded to serve as cognitive recognizers.

To conclude this section, it must be pointed out that the learning to be discussed below is mostly of the Hebbian type. An exception is Section 9.3.3. The modification of a synapse will be related to correlations of its pre-synaptic and post-synaptic activities. There are other organizational principles and not all synaptic modification aiming to adopt a neural system to proper processing of stimuli is necessarily of this type. I would tend to believe that the Hebbian mechanism is a particularly favored candidate when the relevant operation of the network is strongly dependent on feedback, as is the case for ANN's[2]. In other situations, one may be organizing networks which directly feed information forward. Such is the communication of retina to lateral geniculate nucleus, or from there to the primary visual cortex. In such circumstance there could be, and probably are, more effective organizing principles. These will not concern us here.

9.1.2 Modes, time scales and other constraints

The scope of learning to be discussed in this chapter can be described as follows:

- The network is presented, for a relatively long duration, with stimuli. It should then modify its synaptic structure so as to assimilate the impinging stimulus as a *memory*. That is to say that the future behavior of the network in reaction to a similar stimulus, presented for a short duration, will provoke in the network a dynamical process which has been defined as retrieval of the representation of the stimulus that had been learned.

Once it is accepted that learning is a process of modification of synaptic connections due to the reaction of the network to arriving stimuli, we will distinguish learning proposals by two main criteria:

- The existence or absence of a learning mode, namely whether learning (synaptic modification) takes place while the usual network dynamics is suspended, or whether it is a natural companion to normal network operation.
- The relative time scales of the dynamics of the neurons and of the synapses.

Our general background premise here has been that synapses change much more slowly than neurons. In other words, synapses change very little on neuronal cycle-times. This provides a rationale for the treatment of the dynamics of neural states with synaptic values kept fixed.[1]

There is an alternative view, developed by von der Malsburg[5], according to which synapses change rapidly and their values participate actively in retrieval. We will not elaborate on this view, and restrict ourselves to two classes of learning, in the restricted sense announced above: one, with a learning mode; the other, more natural, operating two coupled dynamics, side by side with two different time scales – neurons fast, synapses slow. In fact, we will tentatively adopt the position that effects of very short term memory, on the scale of tens of milliseconds, are manifestations of temporal sequences. Further distinctions between short-term (seconds, minutes) and long-term (years) will not be addressed.

Different biological constraints may or may not be respected by different learning dynamics. Let us briefly mention some constraints of this type:

[1] This attitude should remind the physicist of the celebrated Heitler-London approach to molecular systems.

- Synapses should not change sign upon plastic modification.
- Efferent synapses on a neuron should be of one type – Dale's law.
- The magnitude of synaptic efficacies is probably bounded.
- The *analog depth* – internal resolution between different values – is limited.

Some of the scenarios to be described respect some of these constraints, but this will not serve as a basis for judgment.

9.1.3 The need for learning modes

Why does one need learning modes? The reason is that a network that has learned a number of patterns has thereby developed attractors with large basins of attraction. Any future incoming stimulus will be attracted to those attractors and the dynamics of learning will be distorted. One could imagine that stimuli to be learned arrive with sufficiently large amplitude that the existence of attractors is irrelevant, but that is another way of distinguishing between stimuli that are arriving to be learned from those arriving to be processed. Another problem affecting learning has been indicated in Section 7.3.2. When the loading level of the memory increases, there is a vast number of spin-glass (spurious) attractors, and incoming stimuli are likely to find an attractor in their vicinity, as if it had been previously learned. One way of dealing with this problem may be asymmetry, another is the introduction of a learning mode. In both cases the problem is that of pre-existing attractors.

What learning modes are at our disposal? The simplest, though not very plausible biologically, is a prior identification of certain inputs as candidates for learning. The PSP's of such stimuli must be greatly amplified internally, so as to be able to induce the network into a state corresponding to the stimulus, irrespective of where it had been before. The stimulus must then be removed, to allow the synapses to test themselves. In other words, since the objective is that the network state corresponding to the stimulus should become an attractor, the stimulus should be removed so that it can be checked whether the existing synaptic values are able to sustain it. If not, some corrections of synaptic efficacies must intervene and the stimulus must be tested again. The task of analysis has been to propose and evaluate various

synaptic correction schemes. Clearly, a mode of this kind can be easily implemented in an artificial device.

Another possibility, no less artificial, is to suppose that the stimulus to be learned is imposed with large amplitude for a long duration. The stimulus itself determines the state of the network at consecutive time intervals, rendering the existing synaptic structure temporarily irrelevant. The synaptic modifications can then be introduced as if the network were in an attractor, which is imposed by the amplified input, irrespective of the existing synaptic strengths.

A more sophisticated version of such learning modes is the *master/slave* 'self-optimizing' network[6]. Here, a *master* network of N^2 neurons accompanies and controls a subservient network of N neurons. The set of memories to be learned serves to form synaptic efficacies[2] in the master network. The dynamics of the N^2 master neurons is designed to optimize a function which has its minima at neural values which, if they were the synaptic efficacies of the slave network of N neurons, would cause the memories to be fixed points. Thus the scenario in such a learning mode is that the *master* network is allowed to relax and the neural states at its attractors are coded into the synapses of the *slave* network.

9.1.4 Results for learning in learning modes

Most of the technical results for the effectiveness of learning algorithms have so far been derived for learning in *learning modes*. There must be some message in the fact that the strongest results are very explicit descendants of the *perceptron* learning algorithm and its concomitant celebrated theorem[4]. Some bold steps beyond perceptron learning have indeed been taken, especially in the context of the PDP program. These include such methods as the *Boltzmann machine*[7][Vol. I, ch. 7] and *back propagation op. cit.* [p. 328]. and e.g., ref [8]. The latter should also be classified as learning in a learning mode. But the increase in boldness and promise has so far been accompanied by a significant reduction in analytic tractability and control. These comments would also apply to the very promising approach of investigating the constraints imposed by the network's 'innate' structure on the learning of rules from examples[9]. We now proceed to list some of the results

[2] In what appears to be an infinite regress.

derived in the context of ANN's and then to a description of some technical details.

One can, of course, augment all synapses by the 'outer product', the product of the activities of the two neurons in the last pattern that is presented. If patterns are presented with equal frequency, one would arrive at the standard model with its good news and its bad news. The equal intensity is at the origin of the blackout catastrophe. If old synaptic values are assumed to decay while new memories are acquired in the above brutal way, then one arrives at a *palimpsest* short term memory, which was described in Section 6.6.1. As far as these two 'learning' procedures are concerned not much can be added beyond the analysis of Chapters 4 and 6, since they are simple impositions of the synaptic prescription.

Perceptron algorithms[10,11] introduce the idea of local modification of the synaptic efficacies. Only synapses which are connected to errors in the stabilization of a pattern are changed. The main consequence is that the storage capacity that can be achieved is as high as is theoretically possible. This follows from the perceptron learning theorem which teaches that the procedure will find a set of coefficients which performs almost as well as the best set that is known to exist. This implies, in particular, that one should be able to reach $\alpha = 2$[12]. Such algorithms can be applied in a way which imposes lower limits on some measure of the size of the basins of attraction of the memorized patterns, see also refs. [13,19]. The storage capacity decreases as the size of these basins increases. A network can learn correlated patterns, in the sense of the biased patterns of Section 8.1.1, by this algorithm. It can then achieve unlimited capacity as the bias tends to one.

One can modify the perceptron algorithm[11] so as to construct iteratively a projector synaptic matrix of the Kohonen[16] type (see e.g., Section 4.2.3). The convergence rate, however, is not as well controlled.

9.2 Learning in Modes

9.2.1 Perceptron learning

The only justifications for paraphrasing here the classic exposition[4] of the perceptron learning theory are on the one hand completeness and on the other the long absence of the original text. We shall briefly

describe the learning algorithm of the perceptron and then go on to extend it to learning in a neural network. This extension is not a mere application. Combined with Gardner's computation of the optimal storage capacity as a function of the the size of the basin of attraction K, see e.g., Section 6.1.4, it attains a new dimension.

- The learning time for any set of random patterns whose number, per neuron, is lower than $\alpha_c(K)$, can be shown to be at most polynomial in the number of neurons.

The perceptron, Section 1.3.2, is a linear threshold function ψ $(= 0, 1)$ operating on N variables ϕ_i $(= 0, 1)$, which are the truth functions of some elementary predicates computed in parallel.

$$\psi = 1 \quad \text{if} \quad \sum_{i=1}^{N} A_i \phi_i > T$$
$$\psi = 0 \quad \text{if} \quad \sum_{i=1}^{N} A_i \phi_i < T, \tag{9.1}$$

where the parameters A_i define the perceptron function and are the repository of learning.

- Learning is an algorithm for arriving at a set of A_i which will separate two sets of $\{\phi\}$'s upon the presentation of examples and the indication of errors. If ϕ_i^μ belong to set F^+ $(\mu = 1, \ldots, p_+)$ and ϕ_i^ν belong to set F^- $(\nu = 1, \ldots, p_-)$, learning has been achieved if a set of A_i's has been found such that for all ϕ's in F^+, $\psi = 1$ and for all ϕ's in F^-, $\psi = 0$.

Code for learning by example

The perceptron learning algorithm is:

START: Choose any set $\mathbf{A} = \{A_i\}$.

TEST: Choose a $\Phi = \{\phi_i\}$ from either F^+ or F^-.
If Φ belongs to F^+ and $\psi = 1$ then goto TEST
If Φ belongs to F^+ and $\psi = 0$ then goto ADD
If Φ belongs to F^- and $\psi = 0$ then goto TEST
If Φ belongs to F^- and $\psi = 1$ then goto SUB

ADD: $\mathbf{A} \to \mathbf{A} + \Phi$
Goto TEST
SUB: $\mathbf{A} \to \mathbf{A} - \Phi$
Goto TEST.

Perceptron learning theorem

Theorem 9.1 *If there exists* $\mathbf{A}^*$ *which separates* F^+ *from* F^- *with*

$$\begin{aligned} \mathbf{A}^* \cdot \Phi > \delta > 0 \quad &\text{for} \quad \Phi \in F^+ \\ \mathbf{A}^* \cdot \Phi < -\delta < 0 \quad &\text{for} \quad \Phi \in F^- , \end{aligned} \tag{9.2}$$

then the code will arrive at a separating solution $\mathbf{A}$ *by passing through ADD and SUB a finite number of times – independent of* N.

A few comments:

- Finding an $\mathbf{A}$ is finding N coefficients which can satisfy the appropriate $p_+ + p_-$ inequalities implied by the separation.
- The solution found by the code is not necessarily the one whose existence had been assumed.
- The proof of the theorem provides an upper bound on the number of 'correcting presentations' and not on the total number of presentations.
- Note that while separation is, strictly speaking, Eq. 9.2 with $\delta = 0$, the assumed solution $\mathbf{A}^*$ is required to have $\delta > 0$. In fact, as we shall see below, the convergence time grows when the maximal δ decreases.

Proof of learning theorem

To make the proof more streamlined it is convenient to introduce a few modifications.

1. Allow ϕ_i to take any real value, in particular 0 and 1.
2. The Φ's in F^- are transformed to $-\Phi$. This implies simply that the inequalities for these patterns must be reversed.
3. The Φ's are normalized, i.e.,

$$\Phi \to \frac{\Phi}{|\Phi|}.$$

   Implying that δ has to be normalized according to the norm of the Φ with the largest norm.
4. Normalize $\mathbf{A}^*$ to $|A^*| = 1$, which is yet another rescaling of δ.

Following these modifications one has a single set F of normalized patterns, comprising both F^+ and F^-. The theorem can be restated relative to the new and simpler code:

START: Choose any $\mathbf{A}$.

TEST: Choose a Φ from F.
If $\psi = 1$ then goto TEST else goto ADD

ADD: $\mathbf{A} \to \mathbf{A} + \Phi$: goto TEST.

Note that the $\mathbf{A}^*$ assumed by the theorem refers now to a single positive threshold δ.

The proof consists of showing that there is a 'Lyapunov function' for the coefficients $\mathbf{A}$, namely a function that increases monotonically at every correcting step of the algorithm – when the code goes through ADD – and is bounded from above (see e.g., Section 3.3). This bound would be reached in a number of steps that can be computed in terms of the parameter δ. It is therefore also a bound on the number of correcting steps for arrival at a separating solution. The function is:

$$G(\mathbf{A}) = \frac{\mathbf{A}^* \cdot \mathbf{A}}{|A|} \tag{9.3}$$

Since $\mathbf{A}^*$ is normalized $G(A)$ is a cosine of an angle and the the upper bound mentioned above is

$$G(\mathbf{A}) \leq 1.$$

At every correcting step, both numerator and denominator increase but the numerator increases faster.

Let us denote by $\mathbf{A}_i$ the coefficients following i updatings. Then, on passing through ADD, we have

NUMERATOR	DENOMINATOR
$\mathbf{A}^* \cdot \mathbf{A}_{i+1}$	$\lvert\mathbf{A}_{i+1}\rvert^2$
$= \mathbf{A}^* \cdot (\mathbf{A}_i + \Phi)$	$= (\mathbf{A}_i + \Phi) \cdot (\mathbf{A}_i + \Phi)$
$= \mathbf{A}^* \cdot \mathbf{A}_i + \mathbf{A}^* \cdot \Phi$	$= \lvert\mathbf{A}_i\rvert^2 + 2\mathbf{A}_i \cdot \Phi + \lvert\Phi\rvert^2$
$\geq \mathbf{A}^* \cdot \mathbf{A}_i + \delta$	$< \lvert\mathbf{A}_i\rvert^2 + 1$

Hence

$$\mathbf{A}^* \cdot \mathbf{A}_n \geq n\delta \qquad |\mathbf{A}_n|^2 < n,$$

for large enough values of n. As a consequence,

$$G(\mathbf{A_n}) = \frac{\mathbf{A}^* \cdot \mathbf{A_n}}{|\mathbf{A_n}|} > \sqrt{n}\delta.$$

The bound on G would be violated if a solution is not found in $n \leq 1/\delta^2$ correcting passes through the code. This completes the proof.

It should be pointed out that this is not unmitigated magic. In fact, one of the main examples in *Perceptrons*[4], i.e., the *parity* perceptron [Sec. 11.5], turns out to learn in an exponentially long time. The exponential explosion can be traced to the magnitude of δ for which a normalized solution can be supposed to exist. What is particularly remarkable about this algorithm is that it does not stumble into false solutions on the way. In some sense the procedure is an optimization scheme, yet there are no local stationary points. This cannot be said about any of the more elaborate learning algorithms.

9.2.2 ANN learning by perceptron algorithm

The stabilization of a set of p patterns in a network of N neurons is equivalent to a set of N perceptrons each learning to separate the p patterns in accordance with one given bit of all the patterns. Formally, it implies that at the post-synaptic neuron number i, stability of the p patterns is expressed by the condition

$$h_i \xi_i^\mu = \sum_{j\, j\neq i} J_{ij} \xi_j^\mu \xi_i^\mu > K||J||_i, \tag{9.4}$$

for $\mu = 1, \ldots, p$. The parameter K has been introduced to provide larger basins of attraction and

$$||J||_i \equiv \sqrt{\sum_{j\, j\neq i}^{N} J_{ij}^2} \tag{9.5}$$

is required when $K \neq 0$ to eliminate a possible effect of the overall norm of the coupling matrix.[3]

For i fixed, this can be mapped onto a perceptron by the identification:

$$A_j = J_{ij}, \qquad \phi_j = \xi_i^\mu \xi_j^\mu, \qquad T = K.$$

As i goes from 1 to N, one has N perceptrons. The assumption on the existence of a solution, corresponding to Eq. 9.2, is here

$$\sum_{j\, j\neq i} J_{ij}^* \xi_j^\mu \xi_i^\mu > (K + \delta)||J^*||_i. \tag{9.6}$$

Note the introduction of δ over and above the threshold K.

The question about the existence of such a solution, which is a source of much concern in general perceptron theory, is now completely controlled as a result of the modesty of the goals. In the present context, it has been answered by Gardner[12]. As long as

$$K + \delta < K_c(\alpha),$$

given in Figure 6.2, Eq. 9.6 possesses solutions with just the proper normalization.

Next we have to prescribe the correction procedure of the synaptic matrix. If the stability condition, Eq. 9.4, is violated for some neuron i and some pattern ν, then all connections leading **to** it are modified according to

$$\Delta J_{ij} = \xi_i^\mu \xi_j^\mu. \tag{9.7}$$

Synapses which lead to neurons which found this pattern acceptable rest unmodified.

The proof follows the logic of the previous section, but the result goes beyond it. Instead of the single cosine to be violated we will have N cosines, defined as

$$G_i^n = \frac{\sum_j J_{ij}^n J_{ij}^*}{||J^n||_i ||J^*||_i}, \tag{9.8}$$

[3] I am grateful to Dr. E. Gardner for a helpful discussion of this point.

where J_{ij}^n is the synaptic matrix after n correcting passes through all perceptrons. The final result, skipping some of the detail, is

$$1 \geq G_i^n > \frac{K+\delta}{K+\left(\frac{\ln n}{n}\right)\left(\frac{N}{2(K+\delta)}\right)}. \tag{9.9}$$

This is a rather interesting result which deserves a few comments:

- The limit corresponding to our proof for the perceptron of the previous section is $K \to 0$. In this case one finds:

$$1 > \frac{2\delta^2 n}{N \ln n}.$$

This leads to the bound on the convergence time:

$$\frac{n}{\ln n} < \frac{N}{2\delta^2}.$$

This bound is somewhat weaker than the one obtained before. This is a price paid for the more general bound that Gardner has derived for the case with $K > 0$. Note also the appearance of the factor of N on the right hand side, which was absent in the result for the perceptron. This is due to the different normalization of the learned patterns. Here each component is of magnitude 1, while it was $N^{-1/2}$ in Section 9.2.1. Given that the number of patterns for which a solution exists is computed with the Gardner normalization, this becomes a non-trivial difference. In fact, Gardner's result implies that solutions exist for the perceptron only if $\delta \approx N^{-1/2}$, which means that the convergence time for perceptron learning is also proportional to N.

- For $K > 0$, the upper bound on the time n is:

$$\frac{n}{\ln n} < \frac{N}{2\delta(K+\delta)}, \tag{9.10}$$

reached when the right hand side of Eq. 9.9, which is an increasing function of n, reaches 1.

- If $\delta \to 0$, then Eq. 9.9 provides no bound on the number of required learning cycles.

- The fact that there are N perceptrons remains mute because learning takes place at different neurons independently and hence can take place in parallel.

- Learning time is therefore linear in N.

- The resulting matrix is usually not symmetric.

Finally, there is an additional lever of control in learning a set of random patterns. Recall that the bound we have found in Eq. 9.10 is an upper bound and it may overestimate n. To assume that the process of learning will terminate when $G = 1$ is tantamount to assuming that there is but a single solution to the stability of the p patterns. Or, more generally, that all solutions lie very close to each other. On the other hand, if the volume of solutions for a given number of patterns is large, the search will typically terminate at a much lower value of G, hence much faster. But in the same paper[12], we find the relevant measure of the typical angle cosine as a function of α and K. This is just q of Eq. 6.26 of Section 6.2.3.

As a consequence, one would **typically** have, for $K = 0$ for example, dropping logarithmic terms,

$$n < \frac{q^2(\alpha)}{\delta^2(\alpha)},$$

where $\delta(\alpha)$ is the largest value of δ for the given storage level. It can be read from Figure 6.2. Thus, as α decreases, q decreases with the increasing volume of possible solutions, and $\delta(\alpha)$ increases as it becomes easier to stabilize the patterns. Both functions are explicitly known.

9.2.3 Local learning of the Kohonen synaptic matrix

The previous learning algorithm can be extended to the iterative generation of some specific forms of the synaptic matrix. Alternatively, it has been sometimes modified in the expectation that it could embed the patterns in a shorter time. In none of the extensions does one possess nearly as much theoretical control as in the historic perceptron algorithm. One line of extensions starts by writing the updating condition for the synaptic efficacies, Eqs. 9.4 and 9.7, as

$$\Delta J_{ij} = D(h_i)\xi_i^\mu \xi_j^\mu, \tag{9.11}$$

where D is the function which expresses the stability condition Eq. 9.4, namely

$$D(h_i) = \Theta\left(K||J||_i - h_i\xi_i^\mu\right). \tag{9.12}$$

To extend this scheme one can generalize the function D, which modulates the 'Hebbian' modification of the synapses.

An interesting case emerges if D is taken to be

$$D(x) = x. \tag{9.13}$$

With this choice all patterns modify synapses, irrespective of whether they have provoked errors. The process stops when all patterns have been imprinted with equal intensity, namely when

$$h_i\xi_i^\mu = K||J||_i. \tag{9.14}$$

Moreover, if the right hand side is taken to be unity, the algorithm becomes

$$\Delta J_{ij} = (1 - h_i^\mu\xi_i^\mu)\xi_i^\mu\xi_j^\mu, \tag{9.15}$$

where h_i^μ is the local field on neuron i when the network is in pattern μ. Then, for a set of linearly independent patterns, as the number of updates tends to **infinity**, J_{ij} becomes Kohonen's *pseudo-inverse* or *projector* matrix,[4] discussed in Section 4.2.3[11]. A proof of this statement is given in Appendix 9.4.1. Here we shall conclude with a few comments.

On the good side, one observes that the process is essentially local. A synapse is modified by the activity of the two neurons it connects and by the PSP of the post-synaptic neuron. Previous attempts to construct this specific connectivity matrix have employed non-local prescriptions[17,18]. That this may not be a very biological mechanism should not be taken too seriously. The whole of the learning context under discussion suffers from the same criticism. The pseudo-inverse arrived at by this algorithm is a network which can store more than the standard model. Yet, since the result has a non-zero diagonal, the storage capacity is only $\alpha_c = 0.5$[20]. Moreover, it requires an infinite number of steps. Hence, with the discovery that the Gardner limit, which improves with the level of the correlations – for arbitrary values of K – is attainable by the strongly convergent perceptron algorithm this procedure has lost much of its attractiveness.

[4] This is essentially the Gauss-Seidel for the computation of an inverse of a matrix.

The function D can be further generalized to include cases which are intermediate between the perceptron step-function and the linear function leading to the pseudo-inverse. Such extensions have been proposed by Peretto[14,19]. A particularly suggestive form is the sigmoid function, which appears natural as the desired interpolation. We shall not elaborate on these extensions.

9.3 Natural Learning - Double Dynamics

9.3.1 General features

There have been a number of concrete proposals for models of the double dynamics of neurons and synapses (See e.g., [21,22,23,24,25]). The Boltzmann Machine[7] should also be included in this class, but the question of the relative time scales has not found a very explicit statement in this framework. These models have tried to capture some of the features that emerge when neurons follow their dynamical course and in the process modify the synapses which in turn determine the course of the neural states within the network. The results produced so far by all these efforts are at best tentative, but given the scope of the problem this is no minor achievement. One remains with the impression that the subject is still awaiting a clarifying breakthrough of a magnitude comparable to that ushered in by Hopfield. The projects to be described below should therefore not be considered as accounts of solid accomplishment but rather as a provocation for much further study.

Here we briefly describe three of the attempts: (1) a network of 'physiological neurons'[24]; (2) a network 'learning associations'[23]; (3) and a network that 'maintains memory'[25].

- The first comes closest to describing physiological complexity.
- The second tries to capture a specific cognitive aspect in the process of learning.
- The third is closest in spirit to the philosophy of ANN's, in two senses
  - The learning process is intended to create attractors based on exposure to external stimuli.
  - The learning process itself is a *gradient flow*.

Inverting the order of things, I will start by mentioning what appear to be the major shortcomings in the three proposals to be described.

- The network of physiological neurons does not specify the relevant cognitive event.

The synaptic modifications which take place when the network is exposed to a training stimulus do indeed allow for a synchronized regeneration of the training pattern upon presentation of a partial pattern. Yet, the reconstructed pattern has nothing to distinguish it from any other, spontaneous, insignificant network activity. In other words, the concept of attractors, or some equivalent concept, has not yet been built into the mechanism. Moreover, the memory seems to be of a very short duration.

- The second model – 'learning associations' – operates under the influence of persistent strong external fields, which represent either the stimulus to be learned or to be recognized.

The presence of these fields has a strong influence upon the dynamics and prevents the simple, rapid type of association by error correction, typical of ANN's. That is why the network has to learn associations the hard way.

Furthermore, such prolonged presence of external stimuli in an internal cortical unit is rather atypical. Even prolonged presentations of stimuli are blocked out after short communication. It is just then that an ANN would perform its spontaneous retrieval.

- The third variety is particularly attractive within an ANN program. But for the time being it consists of little more than a promise. Even simulations have not yet been presented.

9.3.2 Learning in a network of physiological neurons

In the present context it is appropriate to recall that:

> Two different dynamical field variables enter the network dynamics, the cell potentials U_i and the synaptic strengths J_{ik}. The dynamics of these variables proceeds on two very different time scales, the potentials U_i being the fast variables, the synaptic strengths J_{ij} the slow variables[24].

In this model the fast variables are neural membrane potentials, while the neurons communicate by means of spikes. See e.g., Section 2.1.1.

Here we briefly describe the dynamics of the slow variables – the synapses. It consists of two main components:

- A relaxation of synaptic efficacies to some initial values with a time constant $T_S \approx 1$–2 seconds.
- A plastic modification of synapses based on:
  1. The activity of the two neurons connected by the synapse.
  2. The current value of the synaptic strength.

The first part can be written simply as

$$\frac{dJ_{ij}}{dt} = -\frac{J_{ij}(t) - J_{ij}(0)}{T_S}. \tag{9.16}$$

The plastic modification does not take place if the synaptic efficacy tries to cross either a lower or an upper bound in magnitude. That is, unless

$$J_u \geq |J_{ij}(t)| \geq J_s,$$

only relaxation is effective. This constraint implies, in particular, that synapses cannot change sign in the course of its plastic modification. The modifications depend on the coincidence of pre- and post-synaptic spikes. In the noiseless network, if the pre-synaptic neuron j has emitted a spike and, within the following period of 2–3ms (the duration of the first spike) the post-synaptic neuron i has also emitted one, then there is a positive increment to the rate of change of the synapse connecting these two neurons. An excitatory synapse is strengthened and an inhibitory one is weakened. If a spike in the first neuron is not followed within spike duration by a spike in the second one, or if a spike in the second one appears outside the spike duration of the first, then there is a negative contribution to the rate of change of the synapse – the excitatory is weakened and the inhibitory is strengthened – to reduce the likelihood of such uncoordinated firing. Otherwise the synapse is unchanged.

Formally, the complete equation for the dynamics of the synapses, when plastic modification can take place, is

$$\frac{dJ_{ij}}{dt} = -\frac{J_{ij}(t) - J_{ij}(0)}{T_S} + \Omega G_j\left(\frac{\Delta t_j}{T_M}\right)\kappa(G_i, G_j).$$

The additional term on the right hand side is composed of:

- A time scale, $\Omega^{-1} \approx 300\,ms$, which measures the rate of learning.
- G_j, which is an exponentially decaying factor expressing the fact that the time rate of change diminishes with Δt_j, the elapsed time after the pre-synaptic spike, on a time scale T_M.
- A decision element κ which realizes the various conditions of spike coincidence and determines the sign of the modification as explained above.

When the network is noisy, the factor κ has to be extended to prevent spurious learning effects due to spontaneous spikes[24]. These are most problematic when they appear in an uncorrelated fashion on the pre- and post-synaptic neurons, causing uncalled for reduction in synaptic efficacy. This is corrected by demanding that κ vanish unless the average firing rates of the two neurons are both significantly above or below the mean spontaneous firing rate.

Learning performance

The typical experiment exhibiting the performance of the network proceeds roughly as follows:

- The connections in the network are chosen to be initially random.
- The connections of the neurons in the network to the receptors are fixed and structured in a mapping of a two-dimensional retina on a two-dimensional network.
- A two-dimensional pattern is presented synchronously, at high frequency and for a long duration to the receptors. It is presented in time intervals of 1ms, for 300ms. This is, in our language, a long term imposition of strong, local, external fields.
- During this period, the synapses are significantly modified in an environment of high neural excitability.
- The inputs are removed for a period of a few tens of milli-seconds. This is long enough for the fast neural dynamics to subside.

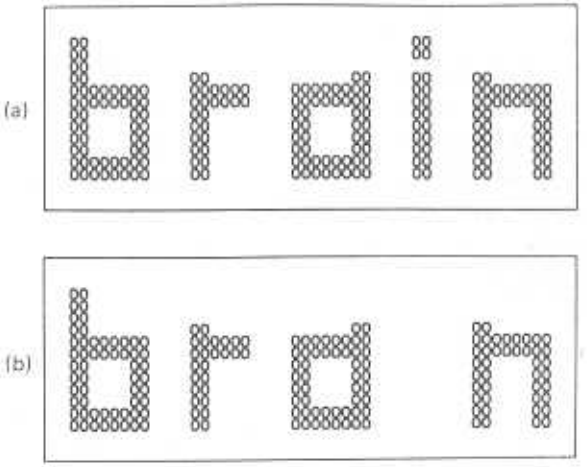


Figure 9.1: The (a) training and (b) the practice patterns for the physiological network. (From ref. [24], by permission.)

- Then the receptors are activated again for 40ms, with a synchronous input of about 85% of the original training pattern. The favorite seems to be the form **brain** for training and **bra n** for the association task. See e.g., Figure 9.1.
- Within about 5ms after the persistent presentation of the test stimulus, all neurons of the network, belonging to the pattern, fire a spike.

An additional variation on the synaptic modification function near the saturation values of the synaptic strengths allows for training of the network with several patterns.

9.3.3 Learning to form associations

The second experiment in learning takes place in a much simpler context. The neurons are our familiar binary variables $S(= \pm 1)$. So the complications connected with the internal dynamics of the neurons are absent. Any successful mechanism found in this context will have to stand the test of robustness against the reintroduction of the biological features. The dynamics of the synapses is the following:

1. If a required modification attempts to drive a synaptic efficacy out of the range $(-1, 1)$, then that synaptic strength is set at the extreme value in the direction it is being modified. The result will be either -1 or 1. This constraint ensures a *palimpsest* feature, namely the erasure of old memories to make room for new ones.

2. If in a neural cycle-time, in which all neurons have been updated, a certain neuron does not preserve its neural state, i.e., $S_j(t+1) = -S_j(t)$, then the synaptic connections efferent from this neuron are weakened multiplicatively. That is,

$$J_{ij}(t+1) = WJ_{ij}(t)$$

with $W < 1$, but very close to 1. This represents a significant departure from Hebbian learning. It is purely pre-synaptic. To rationalize it[23], one appeals to 'police logic', by which 'the best one can do' about a 'dangerous (unstable) element' is to 'weaken its influence on other units'.

3. If two neurons are 'stable', i.e., the updating sweep leaves their states unchanged, both are in the same state, and they are connected by an excitatory synapse, then the synapse connecting them is reinforced by a factor $R > 1$, but close to 1.

4. If two neurons are stable, but are in opposite states and are connected by an inhibitory synapse, then this inhibitory synapse is also reinforced by the factor R.

5. In all other cases synapses remain unchanged.

The difference of the synaptic and neuronal time-scales is expressed by the fact that both W and R are close to unity and thus the modification of the synapses on every sweep through the neuronal states modifies the synaptic efficacies relatively little, while a typical change in the neural state is twice as large as the value of the variable that describes it. The relevant ratio of synaptic to neuronal time scales is approximately $(1-W)^{-1}$ or $(R-1)^{-1}$.

Learning performance

Neural dynamics is identical to the standard model. The local field on a given neuron is compared with a threshold and the new neural state is chosen to be +1 or −1 according to whether that field is higher or lower than the threshold, respectively.

The training procedure starts with a network that has random connections. Specifically, the synaptic efficacies are uniformly distributed between -1 and +1. It then consists of a persistent imposition of an external stimulus, in the form of a strong external field injected into a subset of the neurons in the network. The stimulus fields on different neurons take on the values $+E, -E, 0$. In other words, a subset of the neurons in the network is essentially constrained for the duration of the training period into either active or inactive states. The rest of the neurons are left to their natural threshold dynamics. The network is set in a random configuration and allowed to vary according to usual neural dynamics, with the fields of the stimulus added to afferent network PSP's. The synapses follow the neural dynamics, according to the rules 1–5 set above. This is continued until the **neural** system enters an attractor – a fixed point. Then it is stopped, pending a new experiment. No spontaneous activity is allowed, so the synapses do not change. The pattern of neural activity in the network at the fixed point is considered to be an *internal representation* of the stimulus, much in the spirit of PDP[7].

Recognition is a rapid convergence into an attractor when the **same** stimulus is presented at a future occasion. This is considered a significant event because upon the new presentation the network is started again in a random *network* state. The state of the synapses is preserved from the learning session. It has been found that the network can be trained to memorize, in this sense, several patterns. But in the recognition there is no error correction, since any error is stabilized by a variation of synaptic values. *Association* is, therefore, introduced in a structured way. An association is defined in the following manner:

- Two stimuli are of different *types* if no neuron receives non-zero input from both.

Two such input patterns can be simply combined into a composite pattern. It will have non-zero inputs to any neuron that had an input from either of the two elementary stimuli. An example is shown below.

$$(E, -E, E, E, 0, \ldots, 0) \oplus (0, 0, 0, 0, -E, -E, E, E, 0, \ldots, 0)$$

$$\rightarrow (E, -E, E, E, -E, -E, E, E, 0, \ldots, 0).$$

Association is then the effect of training the network well with the combined stimulus to find that the fixed points, 'the reactions' to the two elementary stimuli become gradually nearer to each other.

9.3.4 Memory generation and maintenance

Two important features are desirable in a learning mechanism.

- The first is that the growth of a given synaptic coefficient is related to the correlated activities of the two neurons connected by that synapse.
- The second is that overloading does not lead to total loss of memory, but at worst to a displacement of old memories by new ones – the palimpsest property. See e.g., Section 6.6.1.

A naive way of implementing the first desideratum is the one implied in Section 4.2.2, and mentioned also in Section 9.1.4. A more sophisticated implementation of this attractive idea (see e.g., Section 7.3.2) is to postulate that the change in the synaptic efficacy is continuously affected by the temporal mean of the correlation of the two connected neurons[26], i.e.,

$$\langle S_i S_j \rangle,$$

which is also the ensemble average. Questions of ergodicity connected with this equivalence will be discussed below.

The second feature has been implemented either by setting a bound on synaptic efficacies[27,28], or by attributing an exponential decay of synaptic strengths with age[29]. See e.g., Section 6.7.3. It would imply that

$$\Delta J_{ij} = -\gamma J_{ij}.$$

The main observation of Shinomoto[25] is that when one puts these two features together, one finds a double dynamical process which is a gradient flow on a surface which is bounded from below. To be specific, suppose that the synaptic dynamics is

$$\Delta J_{ij} = -\gamma J_{ij} + \langle S_i S_j \rangle_J \tag{9.17}$$

where the subscript J indicates that the correlation is computed with a fixed synaptic matrix J_{ij}. This involves the assumption that the synapses vary very slowly on the neural time scale. If this equation has fixed points then J_{ij} does not vary. The average correlation is then

$$\langle S_i S_j \rangle_J = \gamma J_{ij}, \tag{9.18}$$

which is computed with the constant J_{ij}. This is, therefore, simultaneously an attractor for the neural dynamics. One should note that the correlation function is averaged over that part of the space of network states which is accessible ergodically. This important issue will keep coming up. It is advisable at this point to refresh the discussion in Section 3.3.4.

What is the 'Lyapunov' function for the process Eq. 9.17? Consider the function

$$G\{S, J\} = F + \gamma U \tag{9.19}$$

with

$$U\{J\} = \frac{1}{2} \sum_{i,j\, i \neq j} J_{ij}^2 \tag{9.20}$$

and F is the free-energy of a network with the momentary set of synaptic efficacies, given by Eqs. 3.46 and 3.44, i.e.,

$$F\{m, J\} = -\frac{1}{\beta} \ln \left[Tr_S (\exp -\beta E\{S\}) \right].$$

Thus G or F depend on a set of order-parameter variables m, such as overlaps etc., and on $N(N-1)/2$ variables J_{ij} of the symmetric synaptic matrix.

Next, one observes that when a state of a neuron changes following the Glauber dynamics, Eq. 3.23, which is the natural neural dynamics, U does not change. On the other hand, this dynamical neuronal process drives F, at fixed synaptic matrix, toward its minimum, in the sense explained in Sections 4.4.3 and 3.4.2. Neural dynamics, therefore, decreases the function G.

When the synaptic efficacies are varied by ΔJ_{ij}, the change in G is

$$\Delta G = \left[\gamma J_{ij} + \frac{\partial F}{\partial J_{ij}} \right] \Delta J_{ij}.$$

But a simple exercise in statistical mechanics, see e.g., Appendix 9.4.2, gives

$$\frac{\partial F}{\partial J_{ij}} = -\langle S_i S_j \rangle_J, \tag{9.21}$$

with the notation of Eq. 9.17. Then,

$$\Delta G = \left[\gamma J_{ij} - \langle S_i S_j \rangle_J \right] \Delta J_{ij}.$$

Now, when Eq. 9.17 for ΔJ_{ij} is substituted in the last equation, one finds:

$$\Delta G < 0.$$

The conclusion is that the Glauber neural process combined with the synaptic variation of the desired form Eq. 9.17 reduce the function G at every step. Both parts of G are bounded from below. The process must terminate at minima of that function. At such a stationary point, the synaptic matrix will have to satisfy Eq. 9.18.

This is all well and nice except that if the dynamics were ergodic the correlation function would be very small and no learning would take place. In fact, there is not much to be learned. For the mechanism to make sense the network must receive external stimuli. These can serve three roles

- To establish the initial synaptic matrix.
- To extract the system from a specific attractor, so that the attractors corresponding to other patterns can be refreshed.
- To retrieve patterns from memory.

The maintenance of memory

To learn a given set of patterns on a background of previously stored memory, or alternatively on a *tabula rasa*, one would have to present the network with frequent repetitions of each stimulus, so as to dictate the temporal average of the correlations for a while. This is how attractors are dug. Once an attractor is there, the dynamics is no longer ergodic and the system is capable of retrieval. In this way, one can develop a synaptic matrix which stores a given set of patterns. How will such memory fare?

The general behavior of a system under the dynamical process Eq. 9.17 is extremely complicated and adequate analytical tools have not yet been developed. But what seems to be the case is that:

- If a standard associative synaptic matrix, of the Hopfield type, is imprinted with sufficient strength, then it will be maintained with a strength which is determined by a fixed point of the combined dynamical process[25].
- If that matrix is initially imprinted too weakly it will be erased.

The argument proceeds as follows: Suppose that at some moment the synaptic matrix is

$$J_{ij} = B\frac{1}{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu. \tag{9.22}$$

Consider the variation of J induced by changes of the amplitude B. Eq. 9.17 becomes

$$\Delta B = -\gamma B + \frac{\langle S_iS_j\rangle_J}{C_{ij}}, \tag{9.23}$$

where C_{ij} is the matrix multiplying B in Eq. 9.22, and the subscript J indicates that the correlation function is evaluated with the full synaptic matrix J.

Note that the scale factor B is equivalent to a rescaling of the network's temperature. This observation allows one to draw on the structure of the phase diagram of an ANN, Figure 6.8, as explained in Section 6.4.1. The various regions in the phase diagram determine the behavior of the correlation function entering the right hand side of Eq. 9.23. Absorbing β in B one has that

- If
$$B < \beta_g = \frac{1}{T_g} = \frac{1}{1+\sqrt{\alpha}},$$
at 'high temperature', the system is paramagnetic.

In the paramagnetic phase, the network is ergodic and whatever brief, low-frequency, uncorrelated stimuli that will arrive, ergodicity will imply

$$\langle S_iS_j\rangle = 0.$$

Consequently, the synaptic strengths will decay to zero.

- The second phase is the spin-glass phase
$$\beta_g < B < \beta_M = \frac{1}{T_M}.$$
In this phase there is an exponential number of attractors, uncorrelated with the stored patterns.

When in this region of parameters the network receives low-frequency external stimuli it will relax to some spin-glass attractor. But, since these are uncorrelated, when the correlation function is averaged over a time long compared to the time between stimuli, it will again be very small. It is expected to decay exponentially with time as is B[25].

- Then there is the retrieval phase,

$$B > \beta_M. \tag{9.24}$$

In this phase, most stimuli converge rapidly to one of the attractors ξ^μ. If the refreshing stimuli are not selected with some particular bias, then one expects

$$\langle S_i S_j \rangle = \frac{1}{p} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu = \frac{N}{p} C_{ij} = \frac{1}{\alpha} C_{ij}.$$

Substituting in Eq. 9.23 one finds that there is a fixed point value for B, namely

$$B_f = \frac{1}{\alpha\gamma}.$$

But this equality must be satisfied together with the condition that the network be in the retrieval phase, i.e., $B > B_M$, Eq. 9.24. This puts a condition on γ,

$$\gamma < \frac{1}{B_M \alpha}.$$

One expects, therefore, that if $B < B_M$ memory decays to zero. If $B_M < B < \infty$, then as time goes on the amplitude of the synaptic imprinting flows to B_f and remains in its neighborhood, as is schematically represented in Figure 9.2.

This is a rather attractive scheme, but at this stage it leaves many questions unanswered. When the process has non-trivial solutions it surely has very many local stationary states. Very little is known about how these influence the performance and the outcome of the scheme. One may approach such questions by developing analytic tools, but to date one does not even have the phenomenological insight provided by simulations.

To summarize, one may observe that a number of technical projects have been sketched to deal with synaptic modification. Each one of

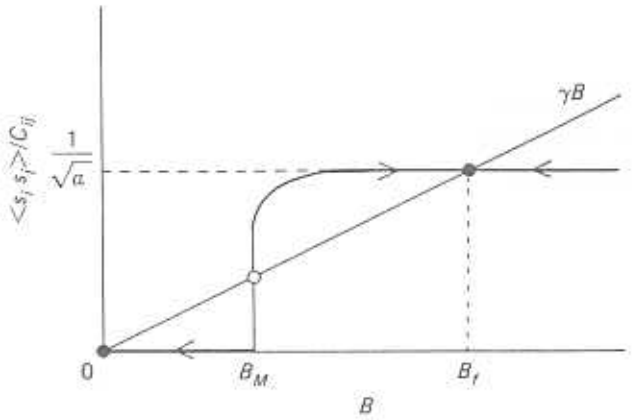


Figure 9.2: Schematic representation of the correlation function $\langle S_i S_j \rangle$ vs the storage strength B. The arrows indicate the time development for a given initial condition B. B_M is the separator between flows to zero memory and flows to the fixed point, stationary synaptic matrix. (After ref. [25], by permission.)

them can be developed and deepened. Yet, what seems, to date, to be lacking mostly is an accompanying biological or psychological interpretation to go with the mechanism. What makes this task so difficult is the need to interpret the interaction of the network with the external world.

9.4 Technical Details in Learning Models

9.4.1 Local Iterative Construction of Projector Matrix

We prove here, following Ref.[11], that if one starts from $J_{ij} = 0$, then the iteration of

$$\Delta J_{ij} = \frac{1}{N}(1 - h_i^\mu)\xi_i^\mu \xi_j^\mu, \tag{9.25}$$

leads to the projector, pseudo-inverse, synaptic matrix:

$$J_{ij} = \frac{1}{N} \sum_{\mu,\nu} (C^{-1})^{\mu\nu} \xi_i^\mu \xi_j^\nu, \tag{9.26}$$

for a set of p linearly independent patterns, where

$$C_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu, \tag{9.27}$$

is the matrix of inter-pattern correlations.

Following l iterations, a pattern μ will be embedded with intensity x_l^μ, which implies that

$$J_{ij}^l = \frac{1}{N}\sum_{\mu=1}^{p} x_l^\mu \xi_i^\mu \xi_j^\mu. \tag{9.28}$$

The dynamics of the iteration is therefore the dynamics of the x^μ's, each of which is the sum over the iterations of

$$(1 - h_i^{\mu,r}),$$

where $h_i^{\mu,r}$ is the local field produced by pattern μ after r iterations. Note that the x^μ's depend on i as well, but each post-synaptic neuron can be treated independently and so this index is sometimes omitted. The change in the x's can be written as

$$\begin{aligned} x_{l+1}^\mu - x_l^\mu &= 1 - \sum_{k=1}^{N} J_{ik}\xi_i^\mu \xi_k^\mu \\ &= 1 - \frac{1}{N}\sum_{\nu=1}^{p}\sum_{k=1}^{N} x^\nu \xi_i^\nu \xi_k^\nu \xi_i^\mu \xi_k^\mu \end{aligned} \tag{9.29}$$

where Eq. 9.28 has been substituted for J_{ij}.

In updating x^μ in the $(l+1)$th iteration, some patterns have already been processed in this iteration and some have not yet. The first ones we denote by $\nu < \mu$, the others by $\nu > \mu$. Equation 9.29 can be rewritten as

$$x_{l+1}^\mu = 1 - \sum_{\nu<\mu} B^{\mu\nu} x_{l+1}^\nu - \sum_{\nu>\mu} B^{\mu\nu} x_l^\nu, \tag{9.30}$$

where

$$B^{\mu\nu} = \frac{1}{N}\sum_{k=1}^{N} \xi_i^\nu \xi_k^\nu \xi_i^\mu \xi_k^\mu = \xi_i^\nu \xi_i^\mu C^{\mu\nu}, \tag{9.31}$$

with C given by Eq. 9.27.

Assuming that the limit

$$x^\mu = \lim_{l\to\infty} x_l^\mu$$

exists, then, recalling that $B^{\mu\mu} = 1$, Eq. 9.30 implies that this limit satisfies:

$$\sum_{\nu=1}^{p} B^{\mu\nu} x^\nu = 1. \tag{9.32}$$

A sufficient condition for the limit to exist is that the matrix B be positive definite, which for linearly independent patterns is ensured due to the identity

$$\sum_{\mu,\nu} B^{\mu\nu} y^\mu y^\nu = \frac{1}{N}\sum_{k=1}^{N}\left(\sum_{\nu=1}^{p} \xi_i^\nu \xi_k^\nu y^\nu\right)^2.$$

In the limit Eq. 9.28 reads

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} x^\mu \xi_i^\mu \xi_j^\mu. \tag{9.33}$$

Thus solving 9.32 for x^μ and substituting in this expression will give the limiting form for the synaptic matrix. The solution for the x's is simply

$$x^\mu = \sum_{\nu=1}^{p} (B^{-1})^{\mu\nu}.$$

But,

$$(B^{-1})^{\mu\nu} = \xi_i^\nu \xi_i^\mu (C^{-1})^{\mu\nu}$$

and when the last two equations are substituted in Eq. 9.33 one finds,

$$J_{ij} = \frac{1}{N}\sum_{\mu,\nu} \xi_i^\nu \xi_i^\mu (C^{-1})^{\mu\nu} \xi_i^\mu \xi_j^\mu, \tag{9.34}$$

which is just Eq. 9.26.

9.4.2 The free energy and the correlation function

The usual free-energy is defined,[5] Eq. 3.46,

$$F = -\frac{1}{\beta} \ln \left(Tr_S \exp \left[\beta \sum_{i,j\, i \neq j} J_{ij} S_i S_j \right] \right).$$

Taking a derivative with respect to J_{ij} leads to

$$\frac{\partial F}{\partial J_{ij}} = -\frac{Tr_S S_i S_j \exp[\beta \sum_{i,j\, i \neq j} J_{ij} S_i S_j]}{Tr_S \exp[\beta \sum_{i,j\, i \neq j} J_{ij} S_i S_j]} = -\langle S_i S_j \rangle.$$

This looks like the desired result, Eq. 9.21. It is not quite. The free energy which is used as a Lyapunov function for the neural dynamics is a slight variation on F. It is its Legendre transform. See e.g., Section 3.4.2. But, when this transform is not performed with respect to the variable involved in the derivative, the result holds.

[5] Here we have not divided by N.

Bibliography

[1] M. Minsky, *The Society of Mind* (Heinemann, London, 1985).

[2] D.O. Hebb, Physiological learning theory, *Journal of Abnormal Child Psychology*, **4**, 309(1976).

[3] P. Churchland, *The Neurophilosophy of Mind* (MIT Press, Cambridge Mass., 1986).

[4] M. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, Mass., 1969).

[5] C. von der Malsburg and E. Bienenstock, Statistical coding and short-term synaptic plasticity: a scheme for knowledge representation in the brain, in *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman Soulié and G. Weisbuch eds. (Springer-Verlag, Berlin, 1986).

[6] A. Lapedes and R. Farber, A self-optimizing, nonsymmetrical neural net for content addressable memory and pattern recognition, *Physica*, **22D**, 247(1986).

[7] D.E. Rumelhart and J.L. McClelland eds. *Parallel Distributive Processing: Explorations in the Microstructure of Cognition* Vols. I and II (MIT Press, Cambridge Mass., 1986).

[8] E. Domani, Iterated learning in a layered feed-forward neural network,*Phys. Rev.*, **A37**, 2660(1988).

[9] P. Carnevali and S. Paternello, Exhaustive thermodynamic analysis of Boolean learning networks, *Europhys. Lett.*, **4**, 1199(1987); J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel and J. J. Hopfield, Large automatic learning, rule extraction, and generalization, *Complex Systems*, **1**, 877(1987).

[10] D.J. Wallace, Memory and learning in a class of neural network models, in *Lattice Gauge Theory – A Challenge in Large Scale Computing*, B. Bunk and K.H. Mutter eds., (Plenum, NY, 1986).

[11] S. Diederich and M. Opper, Learning of correlated patterns in spin glass networks by local learning rules, *Phys. Rev. Lett.*, **58**, 949(1987).

[12] E. Gardner, The phase space of interactions in neural network models, **21A**, 257(1988).

[13] W. Krauth, J.-P. Nadal and M. Mezard, The roles of stability and symmetry in the dynamics of neural networks, *J. Phys.*, **A21**, 2995(1988).

[14] P. Peretto, On the dynamics of memorization processes, *Neural Networks*, **1**, 309(1988).

[15] T.B. Kepler and L.F. Abbott, Domains of attraction in neural networks, *J. Physique*, **49**, 1657(1988).

[16] T. Kohonen and M. Rouhonen, Representation of associated data by matrix operators, *IEEE Trans. Comput.*, **22**, 701(1973).

[17] T. Kohonen, E. Reuhkala, K. Mäkisara and L. Vainio, Associative recall of images, *Biol. Cybernetics*, **22**, 159(1976).

[18] L. Personnaz, I. Guyon, A. Johannet, G. Dreyfus and G. Toulouse, A simple selectionist learning rule for neural networks, in *Proceedings of the Snowbird Conference on Neural Networks for Computing*, April(1986).

[19] L.F. Abbott and T.B. Kepler, Optimal learning in neural network memories, Brandeis University preprint, 1988.

[20] I. Kanter and H. Sompolinsky, Associative recall of memory without errors, *Phys. Rev.*, **A35**, 380(1986).

[21] S. Amari, Learning patterns and pattern sequences by self-organizing nets of threshold elements, *IEEE Trans. Comput.*, **21**, 1197(1972).

[22] P. Peretto and J.J. Niez, Collective properties of neural networks, in E. Bienenstock, F. Fogelman-Soulié and G. Wiesbuch eds. *Disordered Systems and Biological Organization*, (Springer-Verlag, Berlin, 1986)

[23] D. Lehmann, Memory and the formation of associations in neural nets, Computer Science, Hebrew University, preprint.

[24] J. Buhmann and K. Schulten, Associative recognition and storage in a model network of physiological neurons, *Biol. Cybern.*, **54**, 319(1986) and Influence of noise on the function of a "physiological" neural network, *Biol. Cybern.*, **56**, 313(1987).

[25] S. Shinomoto, Memory-maintenance in neural networks, *J. Phys.*, **A20**, L1305(1987).

[26] G. Parisi, Asymmetric neural networks and the process of learning, *J. Phys.*, **A19**, L675(1986).

[27] J.P. Nadal, G. Toulouse, J.P. Changeux and S. Dehaene, Networks of formal neurons and memory palimpsests, *Europhys. Lett.*, **1**, 535(1986).

[28] G. Parisi, A memory which forgets, *J. Phys.*, **A19**, L617(1986).

[29] M. Mezard, J.-P. Nadal and G. Toulouse, Solvable models of working memories, *J. Physique*, **47**, 1457(1986).

10

Hardware Implementations of Neural Networks

10.1 Situating Artificial Neural Networks

10.1.1 The role of hardware implementations

One obvious attraction of artificial neural networks is their potential technological applications, for which they serve as early feasibility studies. This is an issue that is better left at this stage to popular journalism. See e.g., [2,3,4]. Inasmuch as an individual neuron is interpreted as a computing device, artificial neural networks may provide answers to some of the outstanding questions of parallel computing – the coherent coordination of a multitude of processors. This motivation will also not be discussed here. Instead, such networks will be described below for several other reasons.

- To provide a physical environment in which any set of simplifying assumptions about neural networks can be literally implemented. This possibility was raised in Section 1.1.3 in the context of the methodological discussion about verifiabilty of the theoretical results. Some of it can, of course, be investigated by computer simulations.

- In addition, there are various uncontrollable variables which are naturally present in a real system, such as various random

delays, inhomogeneities of components, etc. In this sense such real networks are one step removed from computer simulation.

- In some cases the networks can be constructed out of analog 'neurons', which are amplifiers with a controllable gain. In these networks, one can study the transition from analog to discrete (digital) behavior as a function of the amplifiers' gain (see e.g., Section 2.1). Such networks come a long way toward an inanimate realization of a biological neural networks. In the spirit of this manuscript the main feature which is missing is the communication via spikes. A neurobiologist might add other missing features, such as:
  - Non-uniformity of neuronal structures.
  - The mechanisms of absolute and relative refractivity.
  - The structure of information processing within the arborized dendritic tree[5].

10.1.2 Motivations for different designs

In the following, three different approaches to the implementation of ANN's will be described.

1. The amplifier-neuron resistor-synapse silicon circuit.
2. The electro-optical network.
3. The semi-parallel shift-register large scale network.

The search for ever newer designs, apart from reflecting the diversity of techniques at different laboratories, reflects also various constraints. As a result, different aspects are enhanced or sacrificed in the different approaches. Let us briefly discuss a few such cons and pros.

The first type of network[6,7] is perhaps the closest in spirit to the system one is trying to mimic – a system of neurons. An amplifier with a capacity and a resistive leak is a reasonable schematization of a neuronal *soma*. The synapses are resistors and since those come with only one sign, axonal lines must be doubled. See e.g., Section 10.2, below. Such a network operates in a truly unsynchronized, parallel way. Hence, despite the fact that there are no spikes and the range of

resistor values is narrow, such a network appears closest to what one is trying to capture.

On the other hand, it suffers from very severe restrictions.

- Mass-produced resistors have a limited range of values and are very hard to modify. Thus prospects of implementing learning on such networks are rather remote. From knowledge in the public domain, one tends to conclude that the only synaptic modifications are switching on and off and changing the signs of a set of fixed resistors.
- The number of wires that have to be included in a network of N such neurons increases linearly with N, and faster than $2N$. Moreover, if the synapses are to be modified by external decision elements, then each synapse must be approached by yet another wire. This constrains the design, which is inherently two dimensional, to relatively low numbers of neurons. Biology has a clear advantage in terms of neural growth dynamics as well as in insulating material, which permits disordered three-dimensional densely packed structures.

The electro-optical scheme[8], see e.g., Section 10.3, is an attempt to convert wires into light rays in order to avoid at least two problems: wire routing at high density and power dissipation in resistors. It has an additional appeal which is related to the fact that the synaptic matrix, which is an optically active screen of some sort, is more easily accessible. This may allow 'learning' to be affected by optical means. The electro-optical implementation is very attractive, but at present it seems to be even more restricted than the electronic network for several reasons:

- In the early designs, the optical neurons, which are implemented in *light emitting diodes* (LED), are restricted to a one dimensional organization. The optics involved is very sensitive to the spatial dimension of the array which puts a severe limit on the number of neurons. To date, only a network of 32 optical 'neurons' has been implemented. In addition though some ideas have been presented which promise electro-optical networks with two-dimensional arrays of LED's[8], those require much more sophisticated setups and are still rather theoretical.

- Miniaturization of the electro-optical network, which includes both optical and electrical components, seems a major technological challenge.

Neither wiring nor optics limit the size of a shift-register implementation[9]. The 'neurons' meet their 'synapses' at the 'surface' of the synaptic matrix, which moves around to meet its neurons. See e.g., Section 10.4. There are, therefore, essentially no connections to speak of. The size is limited only by the available high speed shift-registers, and at present is 2,000. This should be more than sufficient since the parallel processing of more than 1,000 bits can lead to little of significance. Beyond such scales one must invoke an operation which utilizes distributed processing in separate networks whose labors are synthesized serially. Moreover, the synaptic matrix is easily accessible and can be varied essentially continuously, which is an aspect which is very partially solved in the electro-optical device.

Yet not all is rosy. This network moves significantly further away from the biological features one has set out to capture.

- It is not fully parallel. The neurons receive their inputs from one neuron at a time.
- It is highly synchronous, in that the rolling of the synaptic matrix has to be well coordinated, which is ensured by a clock.
- It is essentially digital and is conceptually closer to a parallel processing computer simulating neural networks.
- It is relatively slow. A network of 1,000 neurons with state of the art components can perform a cycle in about 0.1ms, which is still two orders of magnitude slower than an electronic implementation.
- And, what is worse, the processing time scales linearly with the size of the network, while in a veritable neural network it is independent of it.

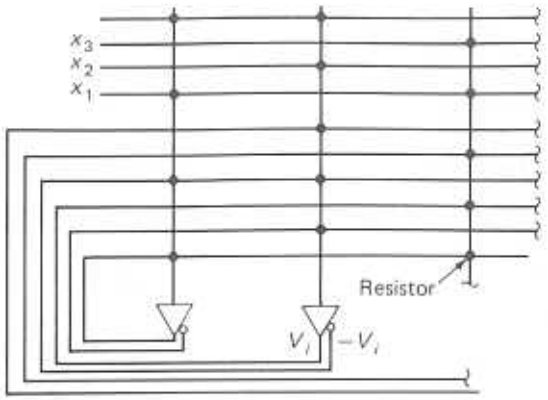


Figure 10.1: A schematic representation of an electronic neural network. The open triangles are amplifiers (neurons). Each has two outputs one, V_i, for the direct output and one for an inverted output $-V_i$, simulating excitatory and inhibitory synapses. Full circles are resistors. The x_i on the left are the inputs fed back into the neurons. (From ref.[1], by permission.)

10.2 The VLSI Neural Network

10.2.1 High density high speed integrated chip

The basic structure is sketched in Figure 10.1. Here we shall comment on some of the physical properties of such networks and on some of the characteristics of their performance. In the VLSI approach the tendency from the very beginning has been to produce high density networks on an integrated chip[6,7]. Then we shall go on to describe a relapse into less sophisticated electronics where the level of control over the parameters is significantly higher and a wide variety of network performances can be observed.

As has already been mentioned the main constraining factor in the fabrication of a neural network is the connectivity matrix, whose size increases with the square of the number of dynamic elements – neurons. The first realizations have involved a small number of neurons, but embarked directly on a technique of high density. It is worth listening to the words of some of the pioneers[7]:

> The array had 484 resistors arranged as a 22×22 matrix, interconnecting the inputs and outputs of 22 CMOS inverters used as high-gain amplifiers. The array was made by

depositing a 0.1μm tungsten film on an oxidized Si wafer and patterning it into a series of parallel 2μm wide lines and spaces using optical lithography and reactive ion etching. These lines were covered with 0.25μm of polyimide, 25nm of evaporated germanium, and 0.25μm of PMMA electron resist.

What this is leading to is a matrix of programmed holes in an otherwise insulating film. On the two sides of the film are two crossed sets of wires, which are the inputs (dendrites) and outputs (axons) of the amplifiers (neurons). Into the holes which correspond to existing connections in the programmed memory an amorphous silicon resistor is injected. These are the synapses. To allow for positive and negative synapses each neuron is doubled, as has already been explained. A picture of such an arrangement, taken by a scanning electron microscope, is presented in Figure 10.2.

The resistances, are each the inverse of a synaptic efficacy. It is determined by the cross-section of the holes and those are uniform. The resistances used in the network described above were rather high, about 300,000 Ohms. These high values are dictated by the parallel dynamics of the network which may lead to excessively large currents. Because the resistors have essentially the same values, determined by the standard size of the holes in the matrix, the available synaptic values are three: +J, -J and 0. As we have seen in Section 7.2.2, this does not imply a severe reduction in the effectiveness of the network as a *content addressable memory*. Where this limitation is more severely felt is in the ability to train the network.

The technique described above allows for densities of up to a few million resistors per square centimeter. The 22×22 resistor matrix, for example, is only 88×88 microns in area. Since the ingredients are just wires and resistors, the density increases with the refinement in lithographic precision. On the other hand, it should be recalled that a matrix once created remains fixed. Moreover, the limitation on the magnitude of the allowed currents leads to another type of scaling: The magnitude of the resistors must increase with the number of resistors which operate in parallel in the network, i.e., with the number of neurons.

This last consideration has led to a special microfabrication process in the production of a chip with 256 neuron-amplifiers[10]. In this

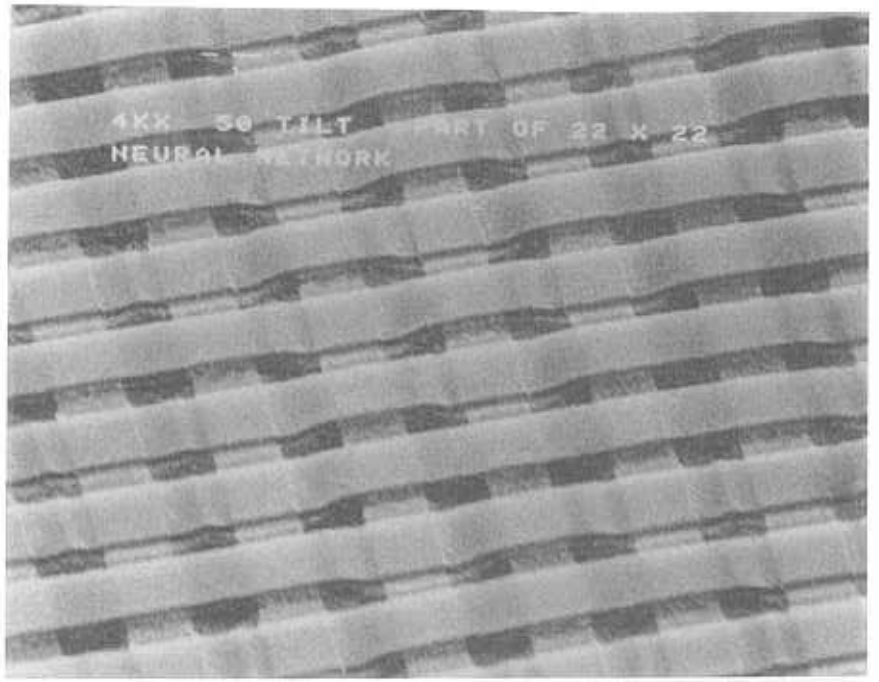


Figure 10.2: A scanning electron microscope photograph of integrated network. One observes the two crossed layers of wires. They are separated by an insulating layer with a matrix of holes filled with silicon resistor synapses. (From ref.[7], by permission.)

network, there are over one hundred thousand resistors and consequently their values must be in the millions of Ohms. In the specific network, each resistor is of 2MΩ. All these are packed into an area of a square of 5.7mm on a side. The resistors in a given chip are fixed, and so are, therefore, the stored memories. Another difficulty which surfaces at this scale is the input-output mechanics. The initial state of 256 neurons has to be communicated to the amplifiers. This is done as follows:

The data are brought in over a 16-bit wide bus and stored in a buffer. When the buffer is filled with the new input data a signal is given that enables the gates, initializing the circuit. All the amplifiers are turned off and their input lines are charged to one of three voltage levels. Once the circuit is

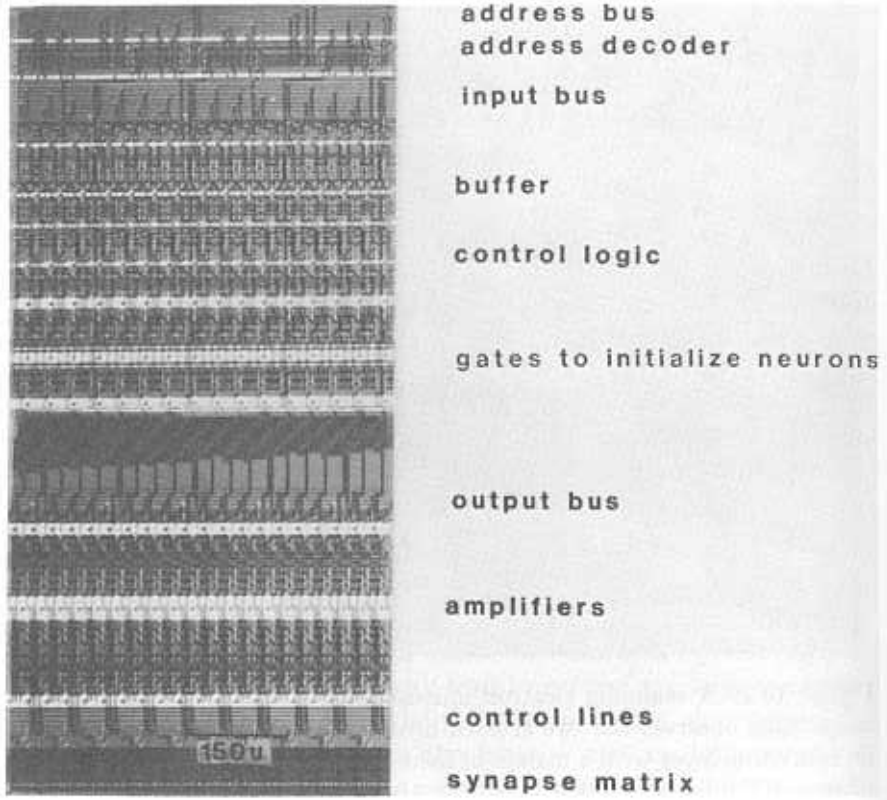


Figure 10.3: A magnified view of a section of the CMOS circuit. Seen are 20 amplifiers and 17 input/output units. (From ref. [10], by permission.)

> initialized, the amplifiers are turned on and the whole circuit settles down freely to a stable state. In the network of 256 neurons this takes about 400ns. Speed is limited though by the input data loading time.[10]

A similar process is required to read the final state – the attractor of the retrieved memory, and all this has to be imprinted into the chip. A view of the final CMOS circuit is reproduced in Figure 10.3.

10.2.2 Smaller, more flexible electronic ANN's

It is mentioned in Ref.[10] that a smaller network, of 54 neurons, has been constructed whose synaptic connections can be turned on and off by software control. A similar approach has been taken in constructing an electronic ANN of 24 'silicon neurons' at the Racah Institute[11]. We will describe this network in some detail not because it is any better or more important than any other, but rather because being home-made we are more familiar with it. Moreover, it has been constructed with a somewhat different attitude in mind. Most other networks are conceived as pilots in an effort to harness ANN's to the commercial-technological promise that has been advertised. This is no mean task, but hardware networks can serve a role in the realization of an increasing number of theoretically predicted features. *Content addressability* is extremely well understood and is a natural candidate for applications. Other features of ANN's are not quite as well understood and their investigation can be much stimulated by observing the behavior of physical implementations. Such features include,

- The relation of synaptic noise to analog neuronal behavior.
- The functioning of networks with different types of *temporal sequences* stored by means of delayed synapses.
- The behavior of networks with synaptic matrices which are significantly different from the standard model – such as extremely asymmetric connections, neurons obeying Dale's law etc.
- Networks with simple synaptic variability, as a preliminary study of learning mechanisms.

Features of this type have been built into our ANN. The basic element becomes, of course, more cumbersome. A 'neuron' is sketched in Figure 10.4. It comprises four amplifiers. The first, numbered (1) in the figure, is the main one. It performs the nonlinear processing of the current arriving afferently from the other neurons in the network. It is parametrized by the cellular leak resistance R, capacitance C and gain. Its basic time constant is (RC), which is about a microsecond. Its input is the sum of currents from four potential sources.

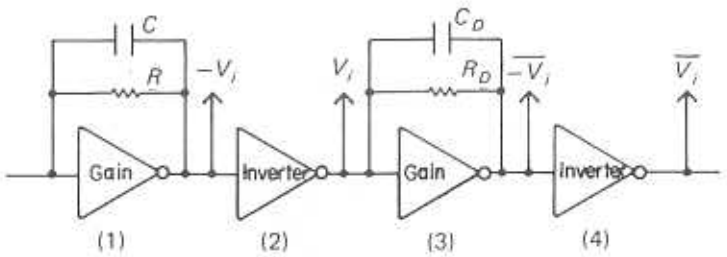


Figure 10.4: The elementary neuron composed of four operational amplifiers, (1) Current amplifier – the cell soma, (2) Inverter, for negative synapses, (3) Delay element – simulating delaying synapses, (4) Inverter of delayed output.

- Synaptically weighted fast currents from other neurons

$$\sum_{j \neq i} J_{ij} V_j,$$

  where V_j is the output voltage of neuron j converted into current through the 'synaptic resistance', or rather conductance, and then summed.

- Delayed currents weighted by their own synapses.

- External input currents I, which are used either to establish the initial state of the network or to provide input during operation.

- Noise, which comes along the same lines as the external currents, but is generated by a noisy source.

The amplifier transduces the current into a voltage via a non-linear gain function $g(I)$. The dynamics at stage (1) is governed by

$$C\frac{dV_i}{dt} + g^{-1}(V_i) = I_{tot}, \tag{10.1}$$

with I_{tot} standing for the sum of all input currents. The time constant is RC, where the capacitance appears explicitly in the equation while the resistence is hidden in the linear term of the amplifier. The signs of the outputs are related to the fact that all amplifiers are inverters. At (2), we have an inverter amplifier with unit gain. It changes the sign of the output potential and provides the input from this neuron

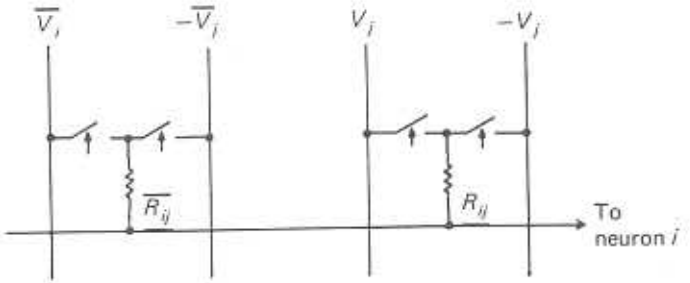


Figure 10.5: Synapse governing input from neuron j into i. Lines marked by $\pm\bar{V}_j$ are delayed inputs, the other two are the fast inputs. Each pair has a single resistance representing the synaptic strengths of the two types of inputs. The arrows point to switches.

to a post-synaptic neuron connected by an inhibitory synapse. At (3), there is a mechanism for simulating a delaying synapse. It comprises an amplifier with an RC feed-back having $R = R_D$ and $C = C_D$ and a gain λ. This circuit operates on the output of the neuron, stage (1), producing an output potential $\bar{V}_i$ according to

$$\tau\frac{d\bar{V}_i}{dt} + \bar{V}_i = \lambda V_i,$$

where V_i is the output of the neuron. This feature integrates the output of neuron i with an exponential decay weight which corresponds to the delay function $w(t)$ in Section 5.6.1, with delay time $\tau = R_D C_D$. The delayed output is then inverted at (4), to provide inhibitory delayed afferents.

The 'synapse' is drawn in Figure 10.5. It connects four input lines into a neuron, one pair for excitatory and inhibitory delayed inputs and one for the fast ones. Each such pair is connected via a single resistance. It is by the choice of switches that the synapses are made excitatory or inhibitory. The setting of the switches is computer controlled. By leaving both switches off the synapse can be disconnected.

Each neuron has an incoming line of external inputs as explained above. The structure of this line is given in Figure 10.6. The top two lines are used to set the initial conditions and are switched off to let the system retrieve. The system can operate in the presence of external persistent input $\pm V_B$ and/or noise.

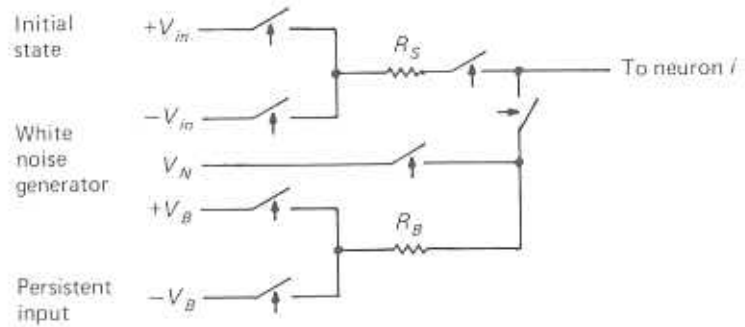


Figure 10.6: Neuronal input mechanism: $\pm V_{in}$ are for setting the initial state; $\pm V_B$ are external inputs during operation; V_N is input from a noise source.

Finally all these elements are connected to make a fully connected network, presented in Figure 10.7. The insets 1,2,3 correspond, respectively, to Figures 10.4, 10.5, 10.6. One additional element which appears in the integrated network are the network switches on all the lines which connect neuronal outputs to inputs. These are instrumental in the imposition of the initial conditions on the network. All these switches are kept open while the amplifiers are brought to their saturated state in the direction desired in the initial state. This is the last step in initializing the operation of the network, as is described below.

The network is operated under micro-computer control. The steps are as follows:

- Patterns to be memorized are described to the micro-computer (PC) and the corresponding standard synaptic matrix is computed. It is then clipped by setting to zero synapses with values below a prescribed threshold and to ± 1 all the others, according to their sign. This allows a symmetric dilution of the matrix, as described in Section 7.2.1.

- The ordering of patterns in a temporal sequence is provided and the corresponding matrix of delayed synapses is constructed.

- These two matrices of 1's, –1's and 0's are then used to set the switches of all synapses in the network. A 0 implies that both switches on a synapse are left open.

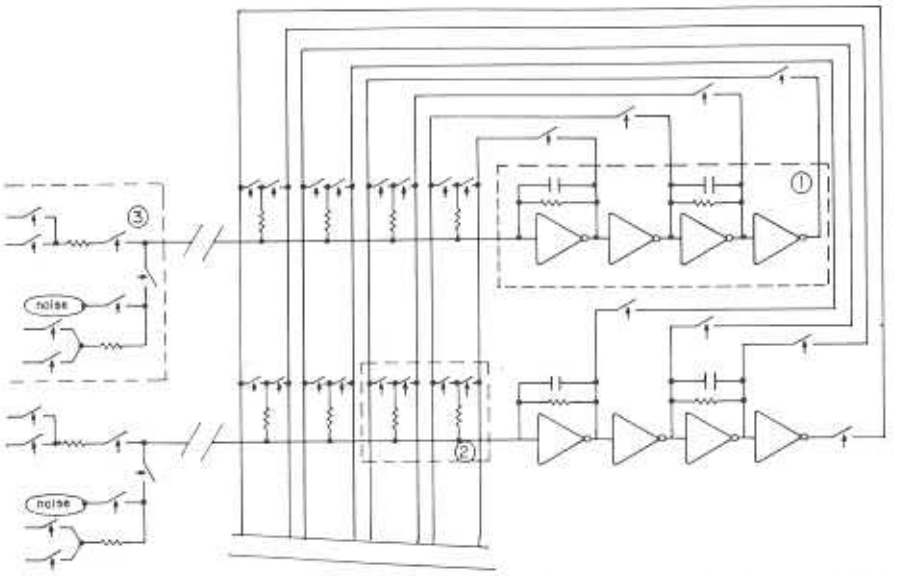


Figure 10.7: The integration of the elements into a network. Insets 1,2,3, correspond to Figures 10.4, 10.5, 10.6.

- A similar procedure is used to help the PC fix the switches on the input channels associated with the noise and the persistent bias.

- The initial state is given and the network switches, mentioned above, are opened to allow the neurons to faithfully enter their states.

- The switches on the input lines of the initial state are then opened and the network switches are closed. The network is left to operate under its own dynamics, including the delays, the noise and the external biases.

- The analog states of all neurons are monitored in real time and stored at a rate of 15 mega-cycles.

- Network states from the high-speed storage are then displayed in human rates (of seconds per state) under PC control.

10.3 The Electro-Optical ANN

The basic idea underlying the electro-optical devices[8] is that the intensity of a light ray can be easily multiplied by ones and zeros when transmitted through or reflected from a medium. A checkerboard of transparent and opaque squares will multiply selectively by zero rays which fall upon opaque squares and by ones those that fall upon transparent squares. This is one step toward realizing the *analog* part of the action of a clipped synaptic action on an arriving set of inputs. Three problems remain:

- If the axon of a firing neuron is to be represented by a light source, then its synaptic contributions to other neurons, its 'axonal branching', must involve a geometric partitioning of the light coming from this neuron onto the squares representing the synaptic values connecting it to the post-synaptic neurons.

- If the network is to perform in an interesting way, the synapses must be of both signs.

- Once the rays have been appropriately multiplied, they must be reassembled to feed every post-synaptic neuron with the weighted sum of the inputs.

The solutions to all three problems, which are all optical, are schematically shown in Figure 10.8. The synaptic matrix in part (a), marked T_{ij}, is an $N \times N$ black and white checkered transparency. Parallel to it is an array of N LED lights – the (0,1) neurons. The synapses emanating from a neuron are the column of squares in the transparency perpendicular to the array of lights, at the horizontal position of the particular neuron-light. The task is, therefore, to focus the light of the LED horizontally and to spread it vertically uniformly over the column of N 'synapses'. This can be done by a cylindrical lens, see e.g., Figure 10.9. Its functioning depends on the performance of ANN's in totally asynchronous conditions. It resolves the first question mentioned above.

It also indicates a possible solution to the third one. The light emitted on the far side of the transparency should be summed over rows to provide post-synaptic input to a neuron. This can be performed by

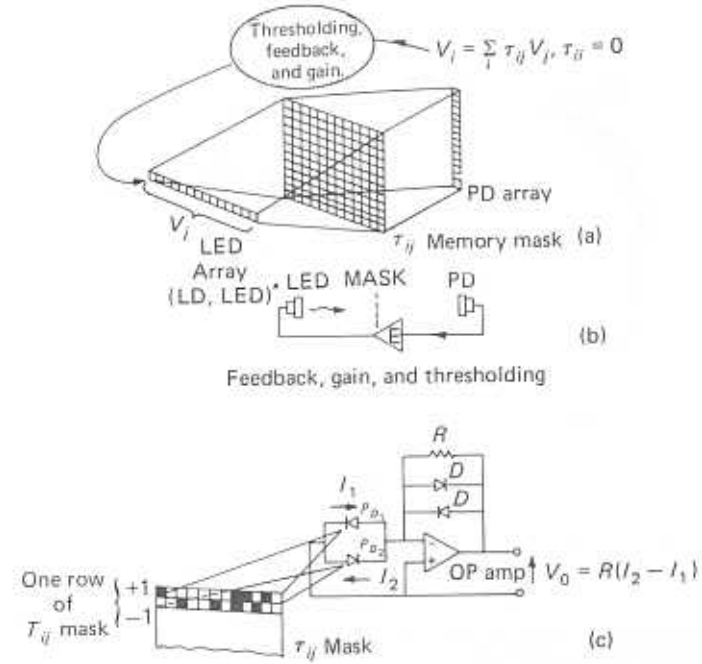


Figure 10.8: Schematic representation of the electro-optical ANN. (a) The synaptic-matrix neural-vector multiplication. (b) The non-linear feedback, thresholding electronic component. (c) The realization of positive and negative synapses. (After ref. [8], by permission.)

a cylindrical lens again, but this one perpendicular to the first one, to focus the values of

$$h_i = \sum_{j=1}^{N} J_{ij} V_j$$

on N vertical sites, denoted by the column to the right of the transparency. Note that the condition $J_{ii} = 0$ is implemented simply by blackening all diagonal squares in the matrix. This column, which represents the neuronal dendrites of the network, consists of photo-detectors, which in turn convert their inputs into electronic signals. A less schematic picture of this part can be seen in Figure 10.9. The current produced by the photo-detectors is summed directly on the back side of the transparency. This eliminates the need for the second cylindrical lens. The summed currents are collected on top of the

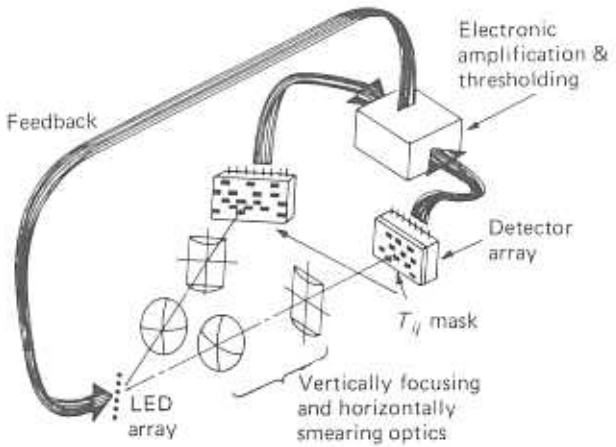


Figure 10.9: An electro-optic circuit diagram of the ANN. Here the positive and negative synapses are on two separate transparencies. The dendritic summation is performed electronically by multichannel photo-detectors. (From ref. [8], by permission.)

transparency and channelled into the non-linear thresholding device. Each element in the column of photodetectors on the right in Figure 10.8(a) is fed into the corresponding LED on the left, via an amplifier and threshold device. This is the *soma* or the 'body of the neural cell'. It is depicted in Figure 10.8(b).

We come back now to the second of the three problems listed above. The optical matrix elements are *unipolar* – the black or white squares imply that for incoherent light the synaptic efficacies are either 0 or 1 rather than of both signs. This is handled[8] as follows: Synapses are restricted to be +1, −1 or 0. Each row is then doubled and one subrow codes the positive clipped synapses belonging to that row and the other the negative ones. The two subrows are either placed right under each other, as in Figure 10.8, or the positive and negative rows are segregated into two separate transparencies, as in Figure 10.9. The output of the two subrows are collected separately on two different photo-detectors, which produce currents of opposite sign. The number of these elements are therefore double the actual number of optical neurons, much as in the electronic network of the previous section.

In practice, a network of 32 neurons was constructed. The 32×64 matrix of optical synapses was a computer generated transparency.

A standard connection matrix, Eq. 4.3, was computed from a set of patterns to be stored. It was then clipped into 0, +1 or −1 depending on whether J_{ij} was zero, positive or negative, respectively. The initial state was chosen by selecting some subset of the LED lights to be on. The LED's are then driven by the feedback and eventually they relax into an attractor, performing the task of associative memory as they should.

The promise of the approach appears to lie in the possibility that

- The electronic feedback be replaced by a non-linear optical mechanism.
- The matrix – the *spatial light modulator* (SLM) – be modifiable, either for synchronized processing of different sets of memories or in order to allow learning.
- The number of 'neurons' be very significantly increased.

The improvements, we are told, are within technological reach[3].

10.4 Shift Register (CCD) Implementation

The last scheme to be described[9] is again a micro-electronic device. Much like the electro-optical effort it aims at circumventing the double difficulty of high connectivity and synaptic inaccessibility. The basic scheme is presented in Figure 10.10. It is based on a solid-state micro-electronic Charge Coupled Device (CCD) which can store discrete groups of electrons is arrays of highly localized sites. These groups of charges can be shifted around at high speed, by the application of external potentials, keeping their local values as they go. This is a *shift register.*

The mode of operation of the circuit is as follows:

- The elements J_{ij} (denoted by T_{ij} in Figure 10.10) are computed externally and are loaded into the CCD arrays in part SYNP in Figure 10.10.
- The states $V_i = 0, 1$ of the CCD array in part X of the design can be considered as axonal states of the N neurons.

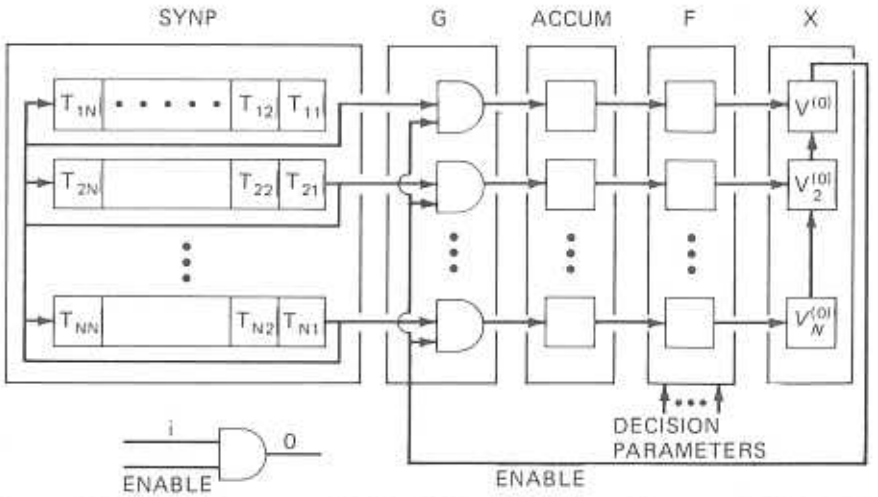


Figure 10.10: Architecture of CCD ANN. (SYNP) are N rows of N CCD's, which contain the synaptic matrix; (G) N analog switches, multiplying synapses by neural states; (ACCUM) summing elements - for 'dendritic' inputs; (F) threshold functions for each neuron; (X) N CCD's representing the neurons. (After ref. [9], by permission.)

- The line marked ENABLE communicates, one by one, the axonal state of *one* neuron, the one appearing at the top slot in X, to the synapses of *all* post-synaptic neurons, which are present, at the right moment, to the right of all CCD arrays in SYNP. The appearance of the correct bit at the top of X and its corresponding line on the right of SYNP is arranged by careful timing.

- The elements in column G multiply the arriving pre-synaptic axonal bit by the numeric value of the synaptic efficacy. This multiplication is most easily envisaged if the 0 or 1 value of the arriving neural state serves as an ENABLE-r (see figure), to either block or pass the synaptic value on its way into the corresponding 'soma' in ACCUM.

- The elements in ACCUM are *integrators*, collecting the synaptic inputs into the post-synaptic neuron, one after the other. At the n-th cycle the value in the k-th ACCUMulator is incremented by $J_{kn}V_n$. The value in this ACCUMulator following this step would be

$$\sum_{j=1}^{n} J_{kj}V_j.$$

- Following each cycle, the $N+1$ CCD arrays are shifted synchronously in a circular way, as is indicated in the figure. This enables the next pre-synaptic axon to feed, via its corresponding synapses, the N post-synaptic somas.

- After $N+1$ clock-cycles the decision elements in column F are activated. Each one produces a zero or a one which is inscribed into the corresponding 'axon' in X.

- This completes an updating cycle and opens a new one until the set in X reaches an attractor, i.e., a repeating configuration of zeroes and ones.

The arrangement manifestly avoids most of the connectivity issues. It also has a synaptic matrix which is very easily accessible and modifiable. On the other hand it sacrifices the complete asynchrony and parallelism of the original idea. The relatively high speed of CCD devices compensates partially for the lack of full parallel processing. The state of the art in that technology are shift-registers of 2,000 CCD's, operating at a frequency of 10MHz. A network of 1,000 neurons can perform 10,000 sweeps per second. This is not a very high speed and looks even worse when it is recalled that the sweep time scales linearly with the size of the network.

Bibliography

[1] J.S. Denker, Neural network models of learning and adaptation, *Physica*, **22D**, 216(1986).
[2] Seeking the mind in pathways of the machine, *The Economist*, June 29, 1985.
[3] J. Kinoshita and N.G. Palevsky, Computing with neural networks, *High Technology*, May 1987.
[4] T. Williams, Optics and neural nets: trying to model the neural brain, *Computer Design*, March 1, 1987.

[5] W. Rall and I. Segev, Functional possibilities for synapses on dendrites and on dendritic spines, in *Synaptic Function*, G.M. Edelman, W.E. Gall and W.M. Cowan, eds. (Wiley, NY, 1987).

[6] M.A. Sivilotti, M.R. Emerling and C.A. Mead, VLSI architecture for implementation of neural networks, *AIP Conf. Proc.*, **151**, 408(1986).

[7] R.E. Howard, D. Schwartz, J.S. Denker, R.W. Epworth, H.P. Graf, W.E. Hubbard, L.D. Jackel, B.L. Straughn and D.M. Tennant, An associative memory based on an electronic neural network architecture, *IEEE Trans. ED*, **34**, 1553(1987).

[8] D. Psaltis and N. Farhat, Optical information processing based on an associative-memory model of neural nets with thresholding and feedback, *Optics Letters*, **10**, 98(1985) and N.H. Farhat, D. Psaltis, A. Prata and E. Paek, Optical implementation of the Hopfield model, *Applied Optics*, **24**, 1469(1985).

[9] A. Agranat and A. Yariv, A new architecture for a microelectronic implementation of neural network models, *Proceedings of IEEE Conference, San Diego June 1987*, **3**, 403(1987) and Semiparallel microelectronic implementation of neural network models, *Electronics Letters*, **23**, 580(1987).

[10] H.P. Graf, L.D. Jackel, R.E. Howard, B. Straughn, J.S. Denker, W. Hubbard, D.M. Tennant and D. Schwartz, VLSI implementation of a neural network memory with several hundreds of neurons, *AIP Conf. Proc.*, **151**, 182(1986).

[11] M. Meidan, D.J. Amit, H. Gutfreund and H. Sompolinsky, Electronic analog neural network with noise and time delays, *Neural Networks: From Models to Applications*, L. Personnaz and G. Dreyfus, eds. (IDEST, Paris, 1989).

Glossary

Absolute refractory period, A period of a few milliseconds following the emission of a *spike* within which no amount of current entering the Neuron's soma will cause that neuron to emit another spike.

Afferent, Incoming to a neuron.

Analog depth, The number of different values which have to be faithfully carried by an element of the network for proper functioning.

Associative recall, *Retrieval* of a particular memory by a large class of initial states which are interpretted as similar stimuli.

Asynchronous dynamics, A dynamical process in which individual *neurons* are picked in a random sequence for updating, and each is updated based on the new situation that has been created by the updating of the previous neuron.

Attractor, A special *network state*, or a restricted set of states, to which the dynamical process, governing the time evolution of the network, brings the network, after a long enough time, from large classes of initial network states.

Attractor neural network (ANN), A network of interacting *formal neurons*, with a high degree of feedback, whose dynamics is governed at long times by *attractors*.

Axon, The output part of a *neuron*. Has sensitive membrane, which can carry the *spike* (or action potential) as a localized, high-speed signal.

Basin of attraction, The set of network states which are attracted by the dynamics to the same *attractor* state.

Burst, A rapid train of spikes, at rates much higher than the mean spike activity.

Canonical ensemble (Gibbs ensemble). A probability distribution of system states for a system at a fixed noise level. The probability of a state is proportional to $\exp(-E/T)$, where E is the energy of the state and T the system's temperature.

Central pattern generator (CPG), A well identified small set of neurons, usually in invertebrates, which can enter into an oscillating activity (*burst*) pattern, without containing *pace makers* nor strong feddback from the motor system.

Cognition time, The shortest time required by the biological system for ascertaining that the dynamics of the network has entered a pattern which corresponds to a cognitive event, e.g. an *attractor*.

Content addressabilty, The ability to recall an item from memory using partial content of the memorized item, instead of a memory address.

Cooperative properties, Properties of a system with a large number of degrees of freedom which are due to the mutual interactions between the constituent elements and which are robust to noise by virtue of their very large number.

Cycle-time, The shortest time interval after which a *neuron*'s state can be modified. Usually taken to be the *absolute refractory period*.

Dale's law, A neurobiological regularity by which neurons emit only one type of synapse, either excitatory or inhibitory.

Degeneracy, The existence of different *network states* with the same energy.

Dendrite, The input part of a neuron. Usually composed of an extensive arbor of fibers.

Detailed balance, A property of a stochastic dynamical process ensuring that the ratio of the probabilities of a transition between two states in one direction and in the opposite direction is the ratio of the values of the same function, evaluated at the two different states.

Deterministic process, In which each state determines uniquely its successor state.

Diversity, The existence of a large number of mutually independent *attractors*.

Efferent, Outgoing from a neuron.

Emergent properties, Properties of a collection of elements which appear as special ideal limiting behaviors. Such properties cannot be simply deduced from the properties of the individual elements, nor from those of small collections of them. See also Cooperative properties.

Ergodicity, A property of a dynamical process which allows the system it governs to arrive from any state to any other state, in the course of time.

Excitatory, An influence on a *post-synaptic neuron* which enhances the probability that the neuron emit a *spike*.

Fast noise, A random variable that changes rapidly on the time scale of the dynamical changes in the network.

Feed forward network, A network in which the interactions between the *formal neurons* are such that the neurons can be divided into groups (layers) and the neural activities in one group can only influence the future activities of neurons in consecutive layers.

Firing rate, Number of *spikes* emitted, by a given neuron, per second.

Fixed-point, A *network state* which repeats itself under the dynamical process.

Formal neuron, A neuron represented by a schematic logical element which acts as a linear threshold function of its inputs.

Free-energy, A quantity which decreases monotonically as a system follows a dynamical process at fixed noise (temperature) level, provided the dynamical process satisfies the condition of *detailed ballance*.

Frustration, The tendency of inter neural interactions to provoke contradictory effects in a given neuron. This is brought about by *excitatory* and *inhibitory* inputs to the same neuron. In the magnetic analog, it is the existence of both ferromagnetic and antiferromagnetic interactions between pairs of spins.

Gain, The slope of the linear part of a response function.

Grandmother cell, A *neuron* which is active *bursting* in response to a very restricted set of stimuli only.

Ground state, State of lowest energy.

Hamming distance, The distance between two N-bit words which amounts to the number of positions in which they differ.

Heat bath (Glauber dynamics), A stochastic dynamical process in which each element in the system chooses its consecutive state as if it itself were in eqilibrium at a given temperature.

Hebbian learning, *Synaptic* modification process in which an

excitatory synapse is strengthened if the two neurons it connects have a strongly correlated activity.

Homunculus, Literally, little man. Often invoked, or implied, in modeling cognitive processes by material elements, as the agent who can determine that an answer has been reached and to interpret its meaning.

Inhibitory, An influence on a *post-synaptic neuron* which reduces the probability that the neuron emit a *spike*.

Lyapunov function, A function of the *network states* which decreases with every step of the dynamical process and is bounded from below. The dynamical process necessarily leads to the minima of such a function.

Meaning, The ability to produce special biological responses for special stimuli, ignoring others as irrelevant transients.

Monte-Carlo, A stochastic dynamical process which ensures the generation of a sequence of states distributed according as in the *canonical ensemble*.

Network state, The specification of the instantaneous *neural states* of an assembly of *neurons* which are considered as a network.

Neural state, One of two, mutually exclusive, situations: either the *neuron* has or has not a *spike* traveling down its *axon*.

Neuron, A cell composing the nervous system. Its salient features are its ability to open to ionic current when stimulated on its *dendrites*; to sum the incoming potentials in its cell body and to emit an *action potential* down its *axon*ic part, if the potential accumulated is high enough.

Overlap (similarity), A parameter which measures the nearness of two N-bit words. It is the angle cosine between the two vectors represented by the two words, and is linearly related to the *Hamming distance* between them.

Pacemaker cell, A neuron which can spontaneously emit spikes at a given rate.

Parallel dynamics, See Synchronous dynamics.

Parallel processing, The performance of a set of related computations, each on a different computing unit, at the same time and in disregard of each other.

Perceptron, A device comprising a set of input channels, a linear threshold calculator and an output channel. The truth values (0, 1) of elemetary propositions arrive via the input channels at the calculator.

Each is weighed by a channel weight, then summed by the calculator, compared to the threshold and the result (0, 1) is communicated along the output channel.

Phase transition, A qualitative change in the dynamical properties of a system of many degrees of freedom due to a chnage of externally controlled parameters, such as the noise level. In particular, there is a large scale modification of the correspondence between initial *network states* and the asymptotic trajectories, as determined by the dynamical process.

Post synaptic potential (PSP), Generated in a *neuron* by the spiking activity of neurons synaptically connected to it.

Pre-synaptic, The output activity of a *neuron*, related to the *action potential*, prior to the secretion of the neurotransmitter. See also Synapse.

Quasi-attractor, A network state which acts as an attractor for a limited amount of time.

Quenched, Effectively fixed on the time scale of the relevant dynamics.

Relative refractory period, A period following the emission of a *spike* within which the *neuron* has, effectively, a higher threshold for emitting a spike.

Retrieval, The process by which a stimulus leads a network to an *attractor* which is a pattern stored in memory. Such an attractor will represent a special pattern of *burst*ing neurons.

Self-recognizability, The ability of the biologocal system itself to assign or detect *meaning*, without the intervention of a *homunculus*.

Sequential dynamics, Same as Asynchronous dynamics, except that the order in which the neurons are selected for updating is fixed.

Simulated annealing, A procedure for finding states close to absolute energy minima, which starts with high temperature, to avoid local minima, and gradually lowers the temperature according to a cooling prescription. Designed in analogy with practical methods used in solidification, for example.

Slow noise, A random variable that changes slowly on the time scale of the dynamical changes in the network.

Soma, The body of a *neuron*. Acts approximately as a linear threshold element.

Sparse coding, A coding of information in which each item uses a small fraction of the available bits.

Spike (Action potential), A localized signal in the local potential difference between the inside and outside of a neuron. It lasts for about a millisecond and travels down the *axon* at about 10 meters per second.

Spurious state, *Attractor* of the dynamics which has not been introduced as memorized patterns.

Stability, A property of an *attractor* and of the dynamical process: if the network is shifted by an arbitrary small amount from a stable attractor, the dynamical process will drive it back to the attractor.

Stochastic process, Each state determines only the relative probabilities of its successor state.

Symmetry breaking, A *phase transition* in which the asymptotic trajectories of the *network states* do not have the symmetry implied by the dynamical process, even in the presence of noise.

Synapse, A junction between the axon of one *neuron* (*pre-synaptic*) and the *dendrite* of another (*post-synaptic*). The arrival of an *action potential* on the pre-synaptic axon, provokes the secretion of neurotransmitter at the synapse, which in turn opens up ionic flow channels on the post-synaptic side of the synapse.

Synaptic efficacy, The amount of *post-synaptic* potential generated in a post-synaptic *neuron* per *spike* emitted in a *pre-synaptic* neuron.

Synaptic plasticity, The change in the properties of a *synapse*, notably its *efficacy*, due to previous activity.

Synchronous dynamics, A dynamical process in which all *neurons* in the network determine their new *neural state* based on the same previous states of all the other neurons.

Index

One of the most exciting and potentially rewarding areas of scientific research is the study of the principles and mechanisms underlying brain function. It is also of great promise to future generations of computers. A growing group of researchers, adapting knowledge and techniques from a wide range of scientific disciplines, have made substantial progress understanding memory, the learning process, and self organization by studying the properties of models of neural networks – idealized systems containing very large numbers of connected neurons, whose interactions give rise to the special qualities of the brain.

About five years ago researchers in this area realized that there are many important parallels between the properties of statistical, nonlinear cooperative systems in physics (where successful modeling techniques were well established) and neural networks. The adaptation of these techniques from physics to the particular problem of neural networks has resulted in enormous progress.

This book introduces and explains the techniques brought from physics to the study of neural networks and the insights they have stimulated. It is written at a level accessible to the wide range of researchers working on these problems – statistical physicists, biologists, computer scientists, computer technologists and cognitive psychologists. The author presents a coherent and clear nontechnical presentation of all the basic ideas and results. More technical aspects are restricted, wherever possible, to special sections and appendices in each chapter. The book is suitable as a text for graduate courses in physics, electrical engineering, computer science and biology.

Daniel Amit is a professor at the Racah Institute of Physics, Hebrew University, Jerusalem. Since 1983 he has made neural networks his central subject of investigation. He is the Honorary Editor of the interdisciplinary journal NETWORK. For many years he has been the chairman of the Committee for Solidarity with BirZeit University in the West Bank.

CAMBRIDGE
UNIVERSITY
PRESS

ISBN 0-521-42124-1
90000
9 780521 421249

design by Dennis M. Arnold