Modeling brain function

Modeling brain function

The world of attractor neural networks

DANIEL J. AMIT

Racah Institute of Physics

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1989
First paperback edition 1992
Reprinted 1994, 1997

Typeset in

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing-in-Publication Data is available

ISBN 0-521-36100-1 hardback
ISBN 0-521-42124-1 paperback

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

Transferred to digital printing 1999

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. Finally, special gratitude is due to A. Herz and R. Kuhn, who carefully read the first printing of this manuscript and, with their comments, contributed so much to the improvement of the second printing.

Rome
November, 1988

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