

The study of the principles behind information processing in complex networks of simple interacting decision-making units, be these units cells ('neurons') in brains or in other nervous tissue, or electronic processors (or even software) in artificial systems which have been inspired by biological neural networks, is one of the few truly interdisciplinary scientific enterprises. The field involves biologists (and psychologists), computer scientists, engineers, physicists and mathematicians; over the years these have all moved in and out of the centre stage in various combinations and permutations, modulated and triggered by advances in experimental, mathematical or computational techniques. The reason for its unique interdisciplinary character is that this multi-faceted area of research, which from now on we will simply denote as the study of 'neural information processing systems', is one of the few which meets the fundamental requirement of fruitful interdisciplinary science: all disciplines get something interesting and worthwhile out of the collaboration. The biologist benefits from tapping into the mathematical techniques offered by the more quantitative sciences, the computer scientist or engineer who is interested in machine learning finds inspiration in biology, and the theoretical physicist or applied mathematician finds new and challenging application domains for newly developed mathematical techniques.

We owe the knowledge that brain tissue consists of complicated networks of interacting brain cells mainly to the work (carried out towards the end of the nineteenth century) of two individuals, who shared the 1906 Nobel Prize in medicine in recognition of this achievement: Camillo Golgi, who invented a revolutionary staining method that for the first time enabled us to actually *see* neurons and their connections under a microscope, and Santiago Ramón y Cajal, who used this new technique to map out systematically and draw in meticulous and artful detail the various cell types and network structures which were now being revealed (in fact Cajal had originally wanted to be an artist). Initially and for several decades neural networks continued to be regarded as a branch of medicine and biology. This situation changed, however, with the birth of programmable computing machines around the time of the second world war, when the word 'computer' was still used to denote a *person* doing computations. It came to be realized that programmable machines might be made to 'think', and, conversely, that human thinking could perhaps be understood in the language of programmable machines. This period also saw the conception of 'information theory', which was largely the brain child of Claude Shannon. Probably the first to focus systematically on the information processing capabilities of neural networks were Warren McCulloch and Walter Pitts, who published in 1943 a paper ('A Logical Calculus of the Ideas Immanent in Nervous Activity') that can safely be regarded as the starting point of our research field. Looking back, one cannot help observing that McCulloch and Pitts were surprisingly typical of the kind of scientist that henceforth would tend to be drawn into this area. McCulloch had studied philosophy and psychology, then moved into medicine, and ended up in a laboratory of electronic engineering. Pitts, who was only twenty at the time when 'A Logical Calculus' was published, initially studied mathematics and also ended

up in electronic engineering, but he never received a formal academic degree. It is not unreasonable to take the view that bringing together these disparate scientific backgrounds and interests was crucial to the achievement of McCulloch and Pitts.

The field never lost the interdisciplinary flavour with which it was born. Since the 1940s its popularity peaked at (roughly) twenty-year intervals, with a second wave in the 1960s (the launch of the perceptron, and the exploration of learning rules for individual neurons), and a more recent wave in the 1980s (which saw the development of learning rules for multi-layer neural networks, and the extensive application of statistical mechanics techniques to recurrent ones). Extrapolation of this trend would suggest that interesting times might soon be upon us. However, the interdisciplinary character of neural network research was also found to have drawbacks: it is neither a trivial matter to keep the disciplines involved connected (due to language barriers, motivation differences, lack of appropriate journals, etc.), nor to execute effective quality control (which here requires both depth and unusual breadth). As a result, several important discoveries had to be made more than once, before they found themselves recognized as such (and hence credit was not always allocated where in retrospect it should have been). In this context one may appreciate the special role of textbooks, which allow those interested in contributing towards this field to avoid first having to study discipline specific research papers from fields in which they have not been trained.

Following the most recent wave of activity in the theory of neural information processing systems, several excellent textbooks intended specifically for an interdisciplinary audience were published around 1990. Since then, however, the connectivity between disciplines has again decreased. Neural network research still continues with energy and passion, but now mostly according to the individual scientific agendas, the style, and the notation of the traditional stake-holding disciplines. As a consequence, those neural network theory textbooks which deal with the progress which has been achieved since (roughly) 1990, tend to be of a different character. They are excellent expositions, but often quite specialized, and focused primarily on the questions and methods of a single discipline.

The present textbook aims to partly remedy this situation, by giving an explicit, coherent and up-to-date account of the modern theory of neural information processing systems, aimed at students with an undergraduate degree in any quantitative discipline (e.g. computer science, physics, engineering, biology, or mathematics). The book tries to cover all the major theoretical developments from the 1940s right up to the present day, as they have been contributed over the years by the different disciplines, within a uniform style of presentation and of mathematical notation. It starts with simple model neurons in the spirit of McCulloch and Pitts, and includes not only the mainstream topics of the 1960s and 1980s (perceptrons, multi-layer networks, learning rules and learning dynamics, Boltzmann machines, statistical mechanics of recurrent networks, etc) but also the more recent developments of, say, the last fifteen years (such as the applica-

tion of Bayesian methods, Gaussian processes and support vector machines) and an introduction to Amari's information geometry. The text is fully self-contained, including introductions to the various discipline-specific mathematical tools (e.g. information theory, or statistical mechanics), and with multiple exercises on each topic. It does not assume prior familiarity with neural networks; only the basic elements of calculus and linear algebra, and an open mind. The book is pitched at the typical postgraduate student: it hopes to bring students with an undergraduate degree to the level where they can actually contribute to research in an academic or industrial environment. As such, the book could be used either in the classroom as a textbook for postgraduate lecture courses, or for the training of individual PhD students in the first phase of their studies, or as a reference text for those who are already involved in neural information processing research. The material has been developed, used and tested by the authors over a period of some eight years, split into four individual one semester lecture courses, in the context of a one-year inter-disciplinary Master's programme in Information Processing and Neural Networks at King's College London.

London, January 2005

Ton Coolen, Reimer Kühn, Peter Sollich

ACKNOWLEDGEMENTS

It is our great pleasure to thank all those colleagues who, through discussions, joint research, joint teaching or otherwise, have directly or indirectly contributed towards this textbook. Especially we would like to thank (in alphabetical order): David Barber, Michael Biehl, Siegfried Bös, Zoubin Ghahramani, Dieter Gensing, Leo van Hemmen, Heinz Horner, Wolfgang Kinzel, Jort van Mourik, Hidetoshi Nishimori, Manfred Opper, Hamish Rae, David Saad, David Sherrington, Nikos Skantzos, Chris Williams and Annette Zippelius.

CONTENTS

I INTRODUCTION TO NEURAL NETWORKS	
1	General introduction 3
1.1	Principles of neural information processing 3
1.2	Biological neurons and model neurons 7
1.3	Universality of McCulloch-Pitts neurons 20
1.4	Exercises 24
2	Layered networks 26
2.1	Linear separability 26
2.2	Multi-layer networks 30
2.3	The perceptron 33
2.4	Learning in layered networks: error backpropagation 41
2.5	Learning dynamics in small learning rate perceptrons 48
2.6	Numerical simulations 53
2.7	Exercises 59
3	Recurrent networks with binary neurons 62
3.1	Noiseless recurrent networks 63
3.2	Synaptic symmetry and Lyapunov functions 70
3.3	Information processing in recurrent networks 73
3.4	Exercises 78
4	Notes and suggestions for further reading 81
II ADVANCED NEURAL NETWORKS	
5	Competitive unsupervised learning processes 85
5.1	Vector quantization 85
5.2	Soft vector quantization 96
5.3	Time-dependent learning rates 103
5.4	Self-organizing maps 106
5.5	Exercises 113
6	Bayesian techniques in supervised learning 117
6.1	Preliminaries and introduction 117
6.2	Bayesian learning of network weights 126
6.3	Predictions with error bars: real-valued functions 133
6.4	Predictions with error bars: binary classification 140
6.5	Bayesian model selection 143
6.6	Practicalities: measuring curvature 148

6.7	Exercises	150
7	Gaussian processes	153
7.1	The underlying idea	153
7.2	Examples of networks reducing to Gaussian processes	156
7.3	The ‘priors over functions’ point of view	159
7.4	Stationary covariance functions	160
7.5	Learning and prediction with Gaussian processes	162
7.6	Exercises	165
8	Support vector machines for binary classification	167
8.1	Optimal separating plane for linearly separable tasks	167
8.2	Representation in terms of support vectors	171
8.3	Preprocessing and SVM kernels	178
8.4	Exercises	182
9	Notes and suggestions for further reading	185
III INFORMATION THEORY AND NEURAL NETWORKS		
10	Measuring information	189
10.1	Brute force: counting messages	189
10.2	Exploiting message likelihood differences via coding	192
10.3	Proposal for a measure of information	197
11	Identification of entropy as an information measure	202
11.1	Coding theory and the Kraft inequality	202
11.2	Entropy and optimal coding	208
11.3	Shannon’s original proof	211
12	Building blocks of Shannon’s information theory	213
12.1	Entropy	213
12.2	Joint and conditional entropy	218
12.3	Relative entropy and mutual information	222
12.4	Information measures for continuous random variables	228
12.5	Exercises	234
13	Information theory and statistical inference	236
13.1	Maximum likelihood estimation	236
13.2	The maximum entropy principle	239
13.3	Exercises	245
14	Applications to neural networks	246
14.1	Supervised learning: Boltzmann Machines	246
14.2	Maximum information preservation	253
14.3	Neuronal specialization	257
14.4	Detection of coherent features	264

14.5	The effect of nonlinearities	267
14.6	Introduction to Amari's information geometry	268
14.7	Simple applications of information geometry	274
14.8	Exercises	278
15	Notes and suggestions for further reading	281
IV MACROSCOPIC ANALYSIS OF DYNAMICS		
16	Network operation: macroscopic dynamics	285
16.1	Microscopic dynamics in probabilistic form	286
16.2	Sequential dynamics	292
16.3	Parallel dynamics	304
16.4	Exercises	311
17	Dynamics of online learning in binary perceptrons	314
17.1	Probabilistic definitions, performance measures	314
17.2	Explicit learning rules	317
17.3	Optimized learning rules	331
17.4	Exercises	346
18	Dynamics of online gradient descent learning	348
18.1	Online gradient descent	348
18.2	Learning from noisy examples	354
18.3	Exercises	356
19	Notes and suggestions for further reading	358
V EQUILIBRIUM STATISTICAL MECHANICS OF NEURAL NETWORKS		
20	Basics of equilibrium statistical mechanics	363
20.1	Stationary distributions and ergodicity	363
20.2	Detailed balance and interaction symmetry	368
20.3	Equilibrium statistical mechanics: concepts, definitions	372
20.4	A simple example: storing a single pattern	378
20.5	Phase transitions and ergodicity breaking	383
20.6	Exercises	388
21	Network operation: equilibrium analysis	392
21.1	Hopfield model with finite number of patterns	392
21.2	Introduction to replica theory: the SK model	401
21.3	Hopfield model with an extensive number of patterns	413
21.4	Exercises	430
22	Gardner theory of task realizability	437
22.1	The space of interactions	437
22.2	Capacity of perceptrons – definition & toy example	442

22.3 Capacity of perceptrons – random inputs	445
23 Notes and suggestions for further reading	453
A Probability theory in a nutshell	457
A.1 Discrete event sets	457
A.2 Continuous event sets	458
A.3 Averages of specific random variables	460
B Conditions for the central limit theorem to apply	462
C Some simple summation identities	465
D Gaussian integrals and probability distributions	466
D.1 General properties of Gaussian integrals	466
D.2 Gaussian probability distributions	470
D.3 A list of specific Gaussian integrals	473
E Matrix identities	478
F The δ-distribution	480
G Inequalities based on convexity	483
H Metrics for parametrized probability distributions	488
I Saddle-point integration	491
References	493