THE *Mathematical Foundations* OF LEARNING MACHINES

NILS J. NILSSON
Stanford University

Introduction by

Terrence J. Sejnowski
The Salk Institute
and
Halbert White
The University of California at San Diego

Morgan Kaufmann Publishers
San Mateo, California

INTRODUCTION

Terrence J. Sejnowski and Halbert White

This book is about machines that learn to discover hidden relationships in data. A constant stream of data bombards our senses and millions of sensory channels carry information into our brains. Brains are also learning machines that condition, combine, parse, and store data. Is it possible to learn something about learning by observing the style of computation used by brains? This is the motivation for research into computational devices that today are called "neural networks." Neural networks are nonlinear dynamical systems with many degrees of freedom that can be used to solve computational problems. The mathematical foundations for learning in this class of machines was laid by a group of researchers in the 1940s and 1950s.

The achievement documented in this book is the thorough study of one of the simplest members of this class, feedforward networks with one layer of modifiable weights connecting input units to output units. In a sense, these might be called reflex machines. The knee-jerk reflex, for example, is mediated by synaptic connections from the sensory receptors in your knee directly onto motoneurons in your spinal cord that in turn activate leg muscles. There are limits to how much computation can be accomplished by such reflexes, and these limits have been carefully delineated in this book. Just as more complex creatures evolved by layering control loops on the primitive reflexes, network models have also evolved in recent years and now have achieved vastly greater capabilities than reflex machines by making use of multilayered architectures with feedback connections. Nonetheless, recent work could not have been accomplished without building on these foundations.

Despite the early promise of research on neural networks, there was a period of about 20 years, from the mid 1960s to the mid 1980s, when interest in neural networks as computational devices and models of human behavior waned in favor of models based on symbol processing. There are many reasons for this, some of them now evident in this book. Still, *Learning Machines* was an underground classic among the neural network modelers who were active during this "dark age" and deserves to be better known to the generation that is "relearning" what was once known about statistical learning machines. The intuitive geometric explanations and the mathematical foundations in this monograph are as invaluable today as they were when it was first written.

In the beginning ...

Nilsson is one of a group of researchers who explored the potential of network models in the late 1950s and early 1960s. These included Frank Rosenblatt, H. David Block, Bernard Widrow, Ted Hoff, Marvin Minsky, Seymour Papert, Karl Steinbuch, Ross Ashby, Oliver Selfridge, Thomas Cover, Woody Bledsoe, Richard Duda, Peter Hart, and many more. There was great excitement about the potential of network learning techniques, but within a few years this line of research nearly disappeared as many of the principals took up other problems and used other approaches. Why this happened can be read between the lines of this book, and in other classics from the same era (Minsky and Papert 1969; Rosenblatt 1959; Duda and Hart 1973). Short but substantial, this monograph is so clearly written that, 25 years later, one can precisely pinpoint the premises and roadblocks that held back further progress.

Chapter 6 on layered machines goes to the very heart of the matter — how to handle additional layers of processing units in a multilayered architecture. These units are now called "hidden units" to distinguish them from units in the input and output layers that receive information or directly interact with the world outside of the network (Hinton and Sejnowski 1983). The hidden units of a network model code the higher-order structure in a problem and make possible the discovery of the relevant features for invariant pattern recognition. It is a difficult problem to discover these features directly from the data. This problem is not tackled head-on in this book, but rather, several interesting examples are analyzed, such as the committee machine studied in Section 6.2, in which the output unit always takes the majority vote of the hidden units.

Despite some success at analyzing special cases, the conclusions reached in the 1960s by the field as a whole were pessimistic. The field saw the problems of dealing with the unconstrained nonlinearities arising in multilayer machines as intractable, and in particular the difficulties of determining properties and training methods for multilayered machines as effectively insurmountable. For example, the judgement of Minsky and Papert, page 232 (Minsky and Papert 1969) was that "The perceptron [a single layer machine] has shown itself worthy of study despite (and even because of!) its severe limitations. It has many interesting features to attract attention: its intriguing learning theorem; its clear paradigmatic simplicity as a kind of parallel computation. There is no reason to suppose that any of these virtues carry over to the many-layered version. Nevertheless, we consider it to be an important research problem to elucidate (or reject) our intuitive judgment that the extension is sterile. Perhaps some powerful convergence theorem will be discovered, or some profound reason for the failure to produce an interesting "learning theorem" for multilayered machine will be found." This pessimism was not an offhand remark, but was the result of years of research culminating in a book that remains one of the best available treatises on neural networks.

Largely for these reasons, mainstream attention shifted to alternative approaches in artificial intelligence that were more promising. The present and now fairly mature field of artificial intelligence based on logic and rules owes much to this shift, and Nilsson has made important contributions to this field. Only when the limitations of a strictly rule-based approach became apparent in the early 1980s was attention drawn again to the potential of massively parallel networks for modeling the complexities of the world. As we summarize in this introduction, some but not all of the problems raised in the book have now been surmounted. The "sterility" predicted by Minsky and Papert was a failure to imagine interesting architectures more powerful than the simple perceptron and less powerful than a general purpose computer. Today, there is an explosion of interesting architectures and demonstrable fecundity.

Threshold logic units and sigmoids

This book is based almost entirely on processing units that have threshold nonlinearities. McCulloch and Pitts (McCulloch and Pitts 1943) had shown that, in principle, networks of these units could compute any computable function. This existence proof was influential in the evolution of general-purpose digital computers (Minsky and Papert 1969); however, it was no help when attempts were made to design learning networks based on threshold logic units (TLUs). The problem was that the class of networks for which learning theorems existed was too weak to support difficult computational problems, and no learning theorems were then known for more complex networks that could solve these problems. This book provides a clear exposition of the results for the class of functions that can be represented by feedforward networks of TLU's with a single layer of modifiable weights, culminating in the fundamental training theorem in Chapter 5. However, there are many points in the book where unsolved problems arose in trying to extend these results to more general contexts, particularly in the last two chapters.

For example, unsolved difficulties arose in Chapter 6 in obtaining weight adjustment procedures for multilayered architectures; the techniques considered pertain to training only a single layer of TLUs. In Chapter 7, the difficulties of error-correction training from "overlapping" patterns generated by probabilistic phenomena are acknowledged; error correction training is then set aside for nonparametric techniques based on mode-seeking. The latter is certainly appealing, but presents its own difficulties, especially in cases with high-dimensional input spaces. Incidentally, the use of the terms parametric and nonparametric has shifted since this book was written, and all the methods in this book would now fall into the class of parametric models — those models with a fixed, finite set of adjustable parameters. Nonparametric techniques now include those that do not have a fixed set of parameters, such as kernel methods (Wertz 1978; Marron 1985) and spline techniques (Wahba and Wold 1975; Cox 1984). The possibility of adding processing units to networks has been raised by Baum (Baum 1989).

In Chapter 2, Nilsson lucidly addresses the capacity of networks built from TLUs. Important theorems by Cover and others are presented and generalized. Researchers who later rediscovered these results could have been spared the effort by reading this chapter. The origin of the "rule of thumb" that each weight in a network of TLU's can store about 2 bits of information can be found here, a result that has subsequently been generalized to processing units with smooth nonlinearities (Mitchison and Durbin 1989; Baum and Haussler 1989). Major advances have been made in the last few years in analyzing networks of TLUs and units that use sigmoids and other smooth functions instead of discontinuous threshold functions. We now understand the ability of this class of feedforward networks to represent multivariate continuous functions (for example, Hornik, Stinchcombe et al. 1989a and 1989b in press) and we now have learning algorithms that can be used to train them. By replacing the discontinuous threshold function with a smooth one, it becomes possible to compute error gradients in multilayer feedforward networks. The method of error backpropagation (Werbos 1974; Parker and Denker 1986; Rumelhart, Hinton et al. 1986) exploited this analytical flexibility to remove the final obstacles to training multilayer machines. Many variant learning procedures for feedforward networks have been developed, including networks using radial basis functions (Moody and Darken 1989; Poggio and Girosi 1989). Convergence results for such learning procedures have been rigorously established by White (White 1989 in press).

This new class of learning algorithms has been applied to a number of difficult problems in speech recognition (Lippmann 1989; Waibel 1989) , optical character recognition (LeCun, Boser et al. 1990), and games (Tesauro 1989). However, there is a severe restriction on the complexity of learning for some problems (Judd 1988). For many problems, such as the parity problem and simple geometrical problems such as connectedness (Minsky and Papert 1969) , the training time and the number of hidden units grows rapidly with the size of the problem. These are problems that require modifications in the architecture to achieve practical solutions. For example, the introduction of multiplicative synapses allows a practical solution to the parity problem for a feedforward network (Durbin and Rumelhart 1989). The extension of feedforward networks by including feedback (recurrent) connections allows the short-term memory of partially processed information to be used as part of the computation. Such networks are being analyzed with techniques from nonlinear dynamical systems (Hirsh 1989).

The study of recurrent networks has been greatly aided by the generalization of the backpropagation learning procedure to recurrent networks (Pineda 1987; Pineda 1989; Almeida 1987; Jordon 1986; Elman 1988). For example, these new learning procedures have recently been used to solve the correspondence problem for matching random-dot stereograms of transparent surfaces (Qian and Sejnowski 1988). Network architectures are now being explored that exhibit relaxation on multiple time scales and others that produce limit cycles rather than fixed-point solutions. Learning algorithms have also been developed for training recurrent networks to recognize temporal sequences and to produce spatio-temporal trajectories (Pearlmutter 1989; Williams and Zipser 1989). A rigorous study of the convergence properties of learning in recurrent networks has been initiated by Kuan (Kuan 1989). Many open questions remain to be examined.

The complexity of layered networks is an active area of research (Abu-Mostafa 1989; Baum 1989). Perhaps the most important open questions concern the rate of convergence of multilayered learning machines as a function of the number of processing units in the middle or "hidden" layers. How does the approximation of the network to the desired function improve with the number of hidden units? How much is required to achieve a particular level of performance as the difficulty of the problem increases? Practical applications of multilayered neural networks to problems in the real-world will depend the answers to these questions.

Probabilistic techniques

The emphasis of this book is primarily on deterministic methods for solving deterministic classification problems. This is an appropriate place to start, but the treatment of probabilistic methods is of paramount importance for realistic applications. Chapter 3 covers the problem of choosing a training set when the distribution of input patterns is known. The special case of multivariate normal distributions is treated in sections 3.7 and 3.11, and further consideration given in

Chapter 7, but the treatment given to this part of the learning paradigm is too brief. In most real-world applications the distributions are unknown and are rarely normal. One of the strengths of the recent advances in learning algorithms is their ability to adapt to a wide range of distributions. Probabilistic methods are now a major tool for constructing and analyzing nonlinear network models.

One of the first stochastic network models to be studied in the modern era was the Boltzmann machine (Hinton and Sejnowski 1983), based on the associative networks introduced by Hopfield (Hopfield 1982). The Boltzmann machine uses binary units and a stochastic update rule to avoid local minima. The behavior of the network can be analyzed with techniques borrowed from statistical mechanics. A learning algorithm was discovered for the Boltzmann machine that provided the first counterexample to the conjecture by Minsky and Papert that extensions of the perceptron learning rule to multilayered networks of TLU's was not possible (Hinton and Sejnowski 1986). The gradients of the weights are estimated by computing the local statistical averages for co-occurrences between pairs of units. The process of statistical averaging is slow to simulate on a sequential digital machine, which must also simulate noise, but there are now VLSI chips that can perform these operations in parallel very quickly (Alspector, Gupta et al. 1989). A deterministic version of the Boltzmann machine based on the mean field approximation also looks very promising (Peterson and Hartman 1989; Hinton and Sejnowski 1986).

Another important use of probabilities is in the specification of input target patterns and in proving asymptotic results for learning algorithms when inputs are presented as a sample from some probability distribution. There is a rich connection with the field of stochastic approximation and powerful asymptotic results are now available for analyzing the convergence of learning rules (White 1990 in press). These results provide a solution for the problems arising in error-correction training from overlapping probabilistic patterns encountered in Chapter 7. In particular, theorems on stochastic approximation establish that use of a learning rate ("correction increment" in Nilsson's terminology) declining to zero at an appropriate rate with the accumulating presentation of training examples leads to a correct probabilistic classification in the limit. It is the constant learning rate that created the difficulties that Nilsson notes in section 7.3.

Neurons and VLSI

When this book was published in 1965, computers were built from discrete components. This has dramatically changed with the development of very large-scale integrated circuits (VLSI), a technology that has contributed to the exponential rise in the computational power of digital machines. It is now possible to simulate large network models; however, these simulations do not take advantage of the inherent parallelism of a neural network, nor do they exploit the fundamental analog nature of the processing units. However, conventional VLSI technology can also be used to directly implement networks using analog circuits. Already, it is possible to build dynamical networks for early sensory processing using analog VLSI (Mead 1989), and learning chips are sure to follow (Schwartz, Howard et al. 1989; Alspector, Gupta et al. 1989). There are still technical problems inherent in analog processing, such as variability, low accuracy and narrow dynamic range. These problems must be solved, or rather understood and exploited, before real learning machines become a reality. Neurons are also limited by low accuracy and limited dynamical range, so there is some hope that these problems are surmountable.

Neural network models have been simulated on a wide variety of digital machines, including coarse-grain and fine-grain parallel architectures. Learning machines made from silicon and optics are also being built that will be far faster and more powerful than current simulations of networks. Special purpose hardware should reduce the power dissipation per unit computation performed by networks by factors of thousands to millions (Mead 1989). At present, this research is exploratory, but there are already a number of VLSI chips that demonstrate feasibility. The promise of fast, adaptive, and inexpensive but powerful computers based on the principles of neural computation, is a driving force behind much of the current research in this field.

There is also the hope, expressed in the term "neural network", that understanding this class of computing devices will provide insights into the function of the brain (Sejnowski, Koch et al. 1988). The network level, one of many in the brain, comes between the level of neurons and structured groups of interacting networks such as columns and maps as illustrated in Figure 1 on page xiv. The principles of neural computation span all of these levels. The current network architectures are relatively simple compared to those that are found in brains, which are highly evolved and often are very specialized (Shepherd 1990).

Some general lessons have been learned concerning the style of distributed representation in populations of neurons by using learning algorithms (Zipser and Andersen 1988 ; Lehky, Sejnowski et al. 1988; Anastasio and Robinson 1989). The learning algorithms are used to construct networks that perform particular tasks, not to model learning in real neural systems. The advantage is that afterwards the modeler can test the properties of the model neurons and compare them with those found in real brains. In the model networks it is often not easy to guess the function of a single neuron from its response properties or even from the distribution of properties in a population of neurons. Models may be indispensible for improving our intuition about distributed processing and demonstrating computational principles. But before we can make detailed comparisons between the brain and simplifying network models, the complexity of the processing units in these models must approach that of real neurons (Shepherd 1990). It may be possible to model

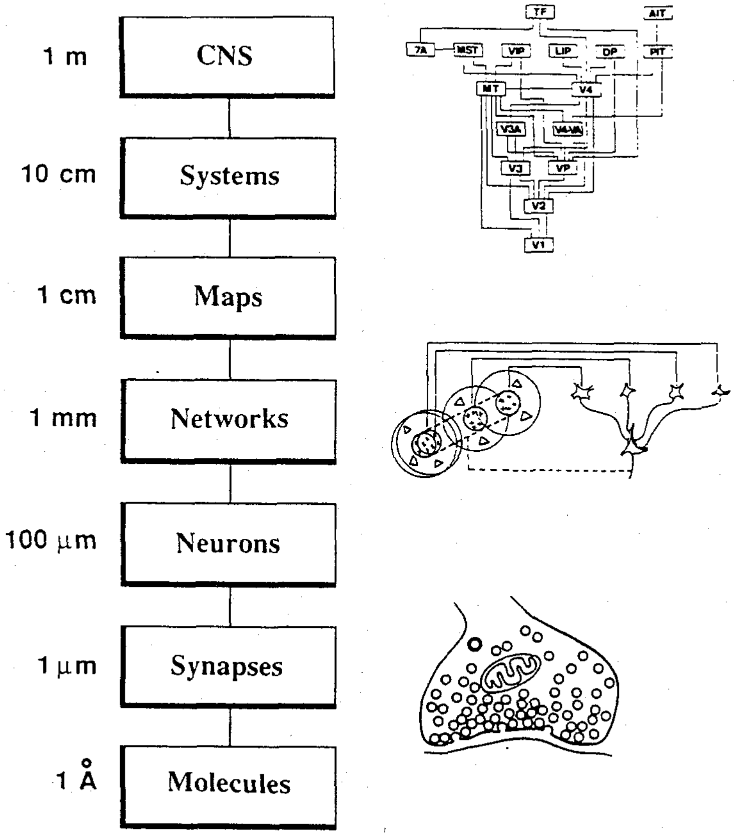

Figure 1. Structural levels of organization in the nervous system. The spatial scale at which anatomical organizations can be identified varies over many orders of magnitude. Schematic diagrams illustrate (top) a cortex; (center) a small network model for the synthesis cortex; and (bottom) the structure of a chemical synapse. Relatively little is known about the properties at the network level in comparison with the detailed knowledge we have of synapses and the general organization of pathways in sensory and motor system.

the dynamic nonlinearities in a single neuron by a micro-network model. Ironically, this would lead to a view of the neuron itself as a "neural network."

Past as prologue

The 20-year hiatus in the development of neural networks is difficult to explain. Many sociological explanations have been offered, such as overselling, underfunding and conspiracy. None of these seem as important, however, as the simple explanation that the problems encountered were technically very difficult. The lack of computer power to simulate large networks is often cited as a crucial factor, but there were severe theoretical limitations as well. What could have prevented this delay? For one, it is likely that interactions between fields would have greatly improved the likelihood of hitting on a profitable approach. The theory of stochastic approximation, which has flourished over the last 35 years, has had a major impact on the modern formulation of learning problems (White, 1990). Methods from the physics of collective phenomena have been critically important conceptually as well as technically (Hopfield 1982; Amit 1989). Biological systems are a source of ideas that will not soon dry up (Sejnowski, Koch and Churchland 1988). This cross-fertilization is just now beginning, but it could have happened decades ago. The problem of communication between the sciences may be a sociological explanation that is even more fundamental.

Several methodological lessons are apparent upon rereading the book in the light of what we now know. The first lesson to be drawn here is that small reformulations of a problem can greatly change the possibilities of making progress. The change from TLUs to sigmoids might not seem like a major reformulation, but by using continuous rather than discontinuous functions, it became possible to generalize the Widrow-Hoff and perceptron learning algorithms to multilayered networks. This has had a dramatic impact not only on the mathematical analysis, but on practical applications as well. The general problem of learning the structure of complex real-world data using neural networks has by no means been solved. It is likely that further changes need to be made to the current generation of models to make them more powerful, just as the generalization of TLUs to sigmoid nonlinearities made it possible to overcome other difficulties.

Another lesson is the importance of experimentation. Nonlinear network models can now be explored through simulations to get intuitions for what works and what does not. Computer power sufficient to explore multilayer network models was not available until quite recently, and without empirical successes it is unlikely that anyone would have looked for proofs of convergence. On the other hand, when experiments fail, it may not be obvious why, or in what direction to go next. Interaction between practical experimentation and formal theory is essential for long-term progress. This interaction is flourishing and the continued health of the field depends on it.

The recent revival of neural networks was strongly influenced by the publication of the books on *Parallel Distributed Processing* (PDP) by Rumelhart and McClelland (Rumelhart and McClelland 1986). Nilsson's book did not have such an influence, probably because it came at a time of transition, away rather than toward neural computation. Also, it ends not on a high point, but on a confusing note regarding the status of multilayer machines. No general guidance is offered except the hope that some progress could be made by looking at special cases. Nonetheless, the book influenced a generation of neural network modelers who were concerned about the mathematical foundations of their field. In this way it may continue to serve the next generation.

Anyone interested in the foundations of neural networks and learning in parallel distributed processing systems should read this monograph. There are many valuable insights and also significant lessons to be learned. As we tackle more and more difficult problems in representing and learning in nonlinear dynamical systems, the strength of our foundations becomes more and more important.

Bibliographic References

Hinton and Anderson (Hinton and Anderson 1981) is a good place to start the modern era. This collection is based on a meeting at San Diego in 1979 that included many of those who revived the field. See especially the chapter by Feldman, who has explored structured network models. The PDP volumes (Rumelhart and McClelland 1986) and the accompanying computer problem book (McClelland and Rumelhart 1988), cover a wide range of topics and network models. Important network models based on unsupervised learning have been a topic that is not covered in Nilsson's book (Willshaw and von der Malsburg 1976; Grossberg 1976; Takeuchi and Amari 1979; Kohonen 1984). Amit (Amit 1989) has extensively analyzed the Hopfield model and its close ties with spin-glass physics. A discussion of the links between network learning and statistics is given by White (White 1990 in press). For collections of old and new papers, see Anderson and Rosenfeld (Anderson and Rosenfeld 1988) and Shaw and Palm (Shaw and Palm 1988). A survey of practical applications of neural networks was published by DARPA (1988).

A good sample of recent research can be found in the annual *Proceedings of the Neural Information Processing Systems Conference* (Touretzky 1989). The *International Joint Conference on Neural Networks* (IJCNN), held twice yearly,

represents a broad selection of research. The *Proceedings of the Annual Meeting of the Cognitive Science Society* (1989) has many papers on the application of network models to cognitive processing. A summer school for neural networks is held every two years and the research of the young investigators attending this school can be found in the *Proceedings of the Connectionist Models Summer School* (Touretzky, Hinton and Sejnowski 1989).

Network models are also being applied to problems in neuroscience (Sejnowski, Koch and Churchland 1988; Koch and Segev 1989; Gluck and Rumelhart 1989; Durbin, Miall and Mitchison 1989; Edelman 1987). Ideas and techniques are flowing in both directions. General mathematical and computational techniques developed in the engineering realm are being applied to specific problems in biological information processing; conversely, advances in understanding the distributed representations and processing in brain circuits can be applied to many practical problems.

Papers on neural networks can be found in an astonishing array of established scholarly journals as well as many new journals that specialize in neural networks. A good place to sample the most recent advances in biological and artificial neural networks is *Neural Computation*, a quarterly journal that publishes short research letters and long reviews. A cornucopia of new journals publish full-length research papers, such as *Neural Networks, Network, International Journal of Neural Systems, Concepts in Neuroscience, Neural Network Review, Journal of Neural Network Computing*, and *IEEE Transactions on Neural Networks*.

Abu-Mostafa, Y. S. (1989). "The Vapnik-Chervonekis dimension: Information versus complexity in learning." Neural Computation 1(3): 312-317.

Almeida, L. B. (1987). "A learning rule for asynchronous perceptrons with feedback in a combinatorial environment." *Proceedings of IEEE First International Conference on Neural Networks.* 2, 609-618. San Diego, CA.

Alspector, J., B. Gupta and R. B. Allen. (1989). "Performance of a stochastic learning microchip." *Advances in Neural Information Processing Systems I.* Morgan Kaufmann Publishers, San Mateo, CA.

Amit, D. J. (1989). *Modeling Brain Function.* Cambridge University Press, Cambridge, MA.

Anastasio, T. J. and D. A. Robinson. (1989). "Distributed parallel processing in the vestibulo-oculomotor system." Neural Computation. 1(2): 230-241.

Anderson, J. A. and E. Rosenfeld. (1988). *Neurocomputing Foundations of Research*, MIT Press, Cambridge, MA.

Baum, E. B. (1989). "A proposal for more powerful learning algorithms." Neural Computation. 1(2): 201-207.

Baum, E. B. and D. Haussler. (1989). "What size net gives valid generalization?" Neural Computation. 1(1): 151–160.

Cox, D. (1984). "Multivariate smoothing spline functions." SIAM Journal of Numerical Analysis. 21: 789-813.

Duda, R. O. and P. E. Hart. (1973). *Pattern Classification and Scene Analysis.* Wiley, New York, NY.

Durbin, R., C. Miall and G. Mitchison. (1989). *The Computing Neuron.* Addison Wesley, Reading, PA.

Durbin, R. and D. E. Rumelhart. (1989). "Product units: A computationally powerful and biologically plausible extension to backpropagation networks." Neural Computation. 1(1): 133-142.

Edelman, G. M. (1987). *Neural Darwinism.* Basic Books, New York, NY.

Elman, J. L. (1988). CRL Technical Report No. 8801: *Finding structure in time.* Center for Research in Language. University of California, San Diego.

Gluck, M. A. and D. E. Rumelhart. (1989). *Neuroscience and Connectionist Theory.* The Developments in Connectionist Theory. Erlbaum Associates, Hillsdale, N.J.

Grossberg, S. (1976). "Adaptive pattern classification and universal recoding: I: Parallel development and coding of neural feature detectors." Biological Cybernetics. 23: 121-134.

Hinton, G. and T. Sejnowski. (1983). Optimal perceptual inference. Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC 448-453.

Hinton, G. and T. Sejnowski. (1986). Eds. J.L. McClelland and Rumelhart, D.E. Learning and relearning in Boltzmann machines. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Psychological and Biological Models.* 2: 282–317.

Hinton, G. E. and J. A. Anderson. (1981). *Parallel models of associative memory.* Erlbaum Associates, Hillsdale, NJ.

Hirsh, M. W. (1989). "Convergent activation dynamics in continuous time networks." Neural Networks. 2: 331-349.

Hopfield, J. J. (1982). Neural neworks and physical systems with emergent collective computational abilities. National Academy of Sciences USA. **79,** 2554-2558.

Hornik, K., M. Stinchcombe and H. White. (1989a in press). "Multilayer feedforward networks are universal approximations." Neural Networks. 2(5):359–366.

Hornik, K., M. Stinchcombe and H. White. (1989b). NC Technical Report. *Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks.* University of California at San Diego.

Jordon, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. Eighth Annual Conference of the Cognitive Science Society. 531-546. Erlbaum Associates, Hillsdale, NJ.

Judd, S. (1988). "On the complexity of loading shallow neural networks." Journal of Complexity. 4:177–192.

Koch, C. and I. Segev. (1989). *Methods in Neuronal Modeling: From Synapse to Networks.* MIT Press, Cambridge, MA.

Kohonen, T. (1984). *Self-Organization and Associative Memory.* Springer Verlag, New York, NY.

Kuan, C.-M. (1989). *Estimation of Neural Network Models.* Thesis. University of California at San Diego.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. (1990). "Backpropagation applied to handwritten zip code recognition." Neural Computation. 1(4): 540-566.

Lehky, S., T. Sejnowski . (1988). "Neural network model for the cortical representation of curvature from images of shaded surfaces." In: J. Lund (Ed.) *Sensory Processing in the Mammalian Brain Neural Substrate and Experimental Strategies.* Oxford University Press, 1988.

Lippmann, R. P. (1989). "Review of neural networks for speech recognition." Neural Computation. 1(1): 1-38.

Marron, J. (1985). "An asymptotically efficient solution to the bandwidth problem of kernel density estimation." Annals of Statistics. 13: 1011-1023.

McClelland, J. L. and D. E. Rumelhart. (1988). *Explorations in Parallel Distributed Processing. A Handbook of Models, Programs, and Exercises.* MIT Press, Cambridge, MA.

McCulloch, W. and W. Pitts. (1943). "A logical calculus of ideas immanent in nervous activity." Bull. Math. Biophysics. 5: 115-133 .

Mead, C. (1989). *Analog VLSI and Neural Systems.* Addison-Wesley, Reading, MA.

Minsky, M. and S. Papert. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

Mitchison, G. and R. Durbin. (1989). “Bounds on the learning capacity of some multilayer networks.” Biological Cybernetics. **60**: 345-356.

Moody, J. and C. J. Darken. (1989). “Fast learning in networks of locally-tuned processing units.” Neural Computation. **1**(3): 281-294.

Parker, D. B. (1986). Ed: J. S. Denker. A comparison of algorithms for neuron-like cells. *Neural Networks for Computing*. American Institute of Physics, New York, NY.

Pearlmutter. (1989). “Learning state space trajectories in recurrent neural networks.” Neural Computation. **1**(2): 263-269.

Peterson, C. and Hartman, P. (1989). “Explorations of the mean field theory learning algorithm.” Neural Networks. **2**(6):475–494.

Pineda, F. J. (1987). “Generalization of backpropagation to recurrent neural networks.” Physical Review Letters. **18**: 2229-2232.

Pineda, F. J. (1989). “Recurrent backpropagation and the dynamical approach to adaptive neural computation.” Neural Computation. **1**(2): 161-172.

Poggio, R. and F. Girosi. (1989). *A theory of networks for approximation and learning*. A.I. Memo Na1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Qian, N. and T. J. Sejnowski. 1988. Learning to solve random-dot stereograms of dense transparent surfaces with recurrent backpropagation. Proceedings of the 1988 Connectionist Models Summer School. Pittsburgh, PA 235-443. Morgan Kaufmann, San Mateo, CA.

Rosenblatt, F. (1959). *Principles of Neurodynamics*. Spartan Books, New York, NY.

Rumelhart, D. E., G. E. Hinton and R. J. Williams. (1986). Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. MIT Press, Cambridge, MA.

Rumelhart, D. E. and J. L. McClelland. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. MIT Press, Cambridge, MA.

Schwartz, D. B., R. E. Howard and W. E. Hubbard. (1989). Adaptive neural networks using MOS charge storage. *Advances in Neural Information Processing Systems I*. Morgan Kaufmann Publishers, San Mateo, CA.

Sejnowski, T., C. Koch and P. Churchland. (1988). “Computational Neuroscience.” Science. **241**: 1299-1306.

Shaw, G. L. and G. Palm. (1988). *Brain Theory*. Advanced Series in Neuroscience. Volume 1, World Scientific Press, Singapore.

Shepherd, G. M. (1990). The significance of real neuron architectures for neural network simulations. In: E. Schwartz (Ed.). *Computational Neuroscience*.MIT Press, Cambridge, MA.

Takeuchi, A. and Amari, S. (1979). Formation of topographic maps and columnar microstructures in nerve fields. Biological Cybernetics. 35:63–72.

Tesauro, G. (1989). “Neurogammon wins computer olympiad.” Neural Computation. 1(3): 321-323.

Touretzky, D., G. Hinton and T. Sejnowski. (1989). *Proceedings of Connectionist Models Summer School*. Pittsburgh, PA. Morgan Kaufmann Publishers, San Mateo, CA.

Touretzky, D. S. (1989). *Advances in Neural Information Processing Systems I*. Morgan Kaufmann Publishers, San Mateo, CA.

Wahba, G. and S. Wold. (1975). “A completely automatic French curve: Fitting spline functions by cross validation.” Communications in Statistics. **4**: 1–17.

Waibel, A. (1989). “Modular construction of time-delay neural networks for speech recognition.” Neural Computation. **1**(1): 39-46.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Thesis. Harvard University.

Wertz, W. (1978). “Statistical density estimation: A survey.” Angewandete Statistique und Okonometrie. Volume 13. Vandenhoeck und Ruprechat, Gottingen, Germany.

White, H. (1989 in press). “Some asymptotic results for learning in single hidden layer feedforward networks.” Journal of the American Statistical Association. Volume 85.

White, H. (1990). “Learning in artificial neural networks: A statistical perspective.” Neural Computation. **1**(4): 425-464.

Williams, R. J. and D. Zipser. (1989). “A learning algorithm for continually running fully recurrent neural networks.” Neural Computation. **1**(2): 270-280.

Willshaw, D.J., Malsburg, C. von der. (1976). “How patterned neural connections can be set up by self-organization.” *Proceedings of the Royal Society*. London, Ser. B 194:431–445.

Zipser, D. and R. Andersen. (1988). “Back propagation learning simulates response properties of a subset of posterior parietal neurons.” Nature. **331**: 679-684 .