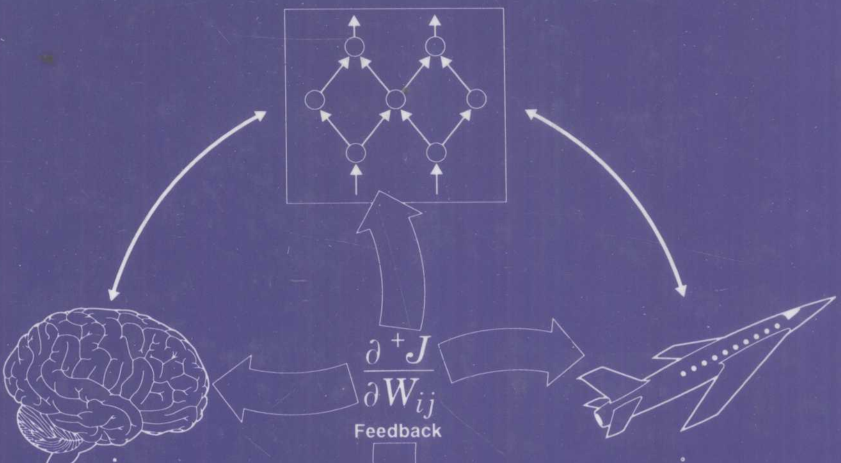
THE ROOTS OF BACKPROPAGATION

From Ordered Derivatives to Neural Networks and Political Forecasting



PAUL JOHN WERBOS

A Volume in the Wiley Series on ADAPTIVE AND LEARNING SYSTEMS FOR SIGNAL PROCESSING, COMMUNICATIONS, AND CONTROL
SIMON HAYKIN, SERIES EDITOR

THE ROOTS OF BACKPROPAGATION

From Ordered Derivatives to Neural Networks and Political Forecasting

PAUL JOHN WERBOS

A WILEY-INTERSCIENCE PUBLICATION
JOHN WILEY & SONS, INC.
NEW YORK / CHICHESTER / BRISBANE / TORONTO / SINGAPORE

This text is printed on acid-free paper.

Library of Congress Cataloging in Publication Data:

Werbos, Paul J. (Paul John), 1947–
The roots of backpropagation: from ordered derivatives to neural networks and political forecasting / Paul John Werbos.
p. cm. — (Adaptive and learning systems for signal processing, communications, and control)
Published simultaneously in Canada.
Originally presented as the author's thesis (Ph. D.—Harvard, 1974).
''A Wiley-Interscience Publication.''
Includes bibliographical references and index.
ISBN 0-471-59897-6 (cloth : alk. paper)
1. Neural networks (Computer science) 2. Regression analysis.
3. Prediction theory. I. Title. II. Series.
QA76.87.W43 1993
003.5—dc20 93-23159
CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Adaptive and Learning Systems for Signal Processing, Communications, and Control

Editor:* *Simon Haykin

Werbos / THE ROOTS OF BACKPROPAGATION: From Ordered Derivatives to Neural Networks and Political Forecasting

Foreword

As the first publicly distributed version of Paul Werbos' 1974 Harvard doctoral dissertation, *Beyond Regression,* this book has significant historical value. The work has far greater value, however, than merely as a historical reference. Indeed, the thesis, accompanied by the updated supplementary material in this volume, continues to provide an instructive account of the development of the fundamental learning rule that is now known as ''backpropagation.'' Further, the thesis is a fascinating narrative of applications of that mathematical learning method to a broad class of difficult prediction problems in the social sciences. The fact that the work was developed nearly 20 years ago simply makes the reading even more inspiring.

After some prefatory material that frames the work in an appropriate historical perspective, Werbos immediately (Chapter Two) develops his algorithm for efficiently calculating derivatives, which he has defined as ''dynamic feedback'' and which we all now call backpropagation. The mathematics is clearly developed and easily understood. The work is rigorous and complete. What results is an enlightening account, starting with simple difference equations, of the development a powerful learning rule.

With the development of the backpropagation learning algorithm complete, Werbos engages in the main thrust of the work, thoughtful discussions of its application to prediction problems. The insights gained since its original publication serve to better interpret its utility and to extend its value to previously unexplored classes of problems (notably, for example, its application to artificial neural networks).

Since the popularization of backpropagation and all of its relatives in the 1980s, there have been countless articles and a fair number of texts that describe the application of backpropagation to difficult estimation and prediction problems. Typically, these applications focus on a concrete engineering or scientific goal; for example, detecting military targets in noisy thermal images or controlling a structure in the presence of an unknown noise disturbance. Far more abstract social science problems are considered, however, in the original portions of this work. The fact that these discussions predate the more modern, less ambiguous investigations illustrate Werbos' clever perceptions about predictive theories.

The entertaining reading that follows links dynamic feedback to problems in national assimilation, Skinner's behavioral models, political mobilization, questions of ethics and rational behavior, and internal telecommunication patterns in Norway, to name a sample few. Werbos' work is remarkable in that it shows the application of the same fundamental learning rule to diverse classes of problems.

I am sure readers of this book will enjoy it, not merely as a historical record, but also as an instructive essay describing the development of the most popular learning rule in use today and as a fascinating journey through its applications to difficult and interesting socio-economic phenomena.

BERNARD WIDROW
Professor of Electrical Engineering
Stanford University

Preface

Simon Haykin, the editor of this series, offered to publish my 1974 Ph.D. thesis, *Beyond Regression*, because this thesis has become something of a classic reference in the neural network and engineering world. The thesis is now recognized as the original source of *backpropagation*, which is now the most widely used algorithm by far in the neural network world. Also, the thesis *communicates* that algorithm to an audience with *no prior understanding* of neural networks; it is divided up, de facto, into chapters written for regular engineers, for statisticians, and for political scientists (mirroring the thesis review committee). Chapters 7 to 10 are also important, because they strengthen both the historical and the pedagogical value of this book.

In actuality, the thesis contains much more than the basics of backpropagation. My real aim in the thesis was to translate some of the vast goals of the general systems and cybernetics movements into a tangible mathematical reality: to initiate a whole body of new mathematical tools capable of generating deeper theoretical understanding and practical applications across a whole range of topics, ranging from neuroscience and engineering to economics[1] and political forecasting,[2] with relevance to the deepest issues faced by human society and human individuals. The thesis gave an example of the use of these tools, by applying them to a model of nationalism and social communications proposed by Karl Deutsch,[3] by turning it into a working forecasting model, while enriching it with greater detail. The thesis also discussed the issue of how to go beyond behaviorism, in building social sciences that account for the phenomenon of *intelligence*.

In 1974, these goals seemed incredibly ambitious for a mere graduate student. However, the great blossoming of the neural network field as a large, organized intellectual disciple (or "interdiscipline") since 1987 has made it possible for many people, including myself, to build on the ideas developed

[1] G. DeBoeck, ed., *Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets*, Wiley, New York, 1994.

[2] A. Gore and M. Lloyd, in *Campaign and Election*, in preparation for 1994 submission.

[3] K. W. Deutsch, *Nationalism and Social Communications*, MIT Press, Cambridge, MA, 1966.

in the thesis and expand their applications in many directions. Also, the formulation of backpropagation in the thesis is the logical basis for fast automated differentiation (FAD), which has begun to excite interest in the more classical numerical world; by integrating such FAD programs into new designs based on backpropagation, it now seems as if the learning capabilities available for neural networks can be extended for easy use with a wider class of models, such as econometric-style models, conventional engineering models, finite element models, or production rule systems interpreted via fuzzy logic. Applications of this kind already exist but require serious technical skills. Connections have been developed between this mathematics and many strands of human psychology—cognitive science, experimental psychology or conditioning systems, and humanistic studies of psychology and social psychology. Finally, neural network methods based on backpropagation have gone back to the original domain of application studied here: the forecasting of economic[1] and political[2] variables, which has also proved highly successful. Many of these studies, cited in the Introduction can be understood without the highest level of mathematical sophistication.

These new developments have required many new techniques and refinements, based on the work of many people, which rise much higher than the foundations presented here. However, it is surprising for me to look back on my thesis and see how many of the most recent developments are echoed—sometimes briefly and sometimes at length—in this document written in 1974.

The Introduction will describe some parts of this history in more detail, in order to put the thesis (Chapters 1–6) and applications (Chapters 7–10) into perspective. It will provide citations to related sources and explain how this work relates to more advanced possibilities for future developments. First, I will describe backpropagation or derivative calculations as such; then I will describe optimization and prediction tools emerging from the thesis and from backpropagation; and, finally, I will discuss linkages to larger social and psychological issues, in a way that I hope is accessible to social scientists. (Both in the thesis and in the Introduction, the social scientist may want to jump over some of the earlier mathematically oriented discussion.)

For completeness, I should add that all this work—both the thesis and the more recent advanced work cited here—could easily have been lost in obscurity, even after 1987–1988. The neural network community and I have a very deep debt to five individuals—David Parker, Bernard Widrow, Stephen Grossberg, Nick DeClaris, and Richard Sutton—who helped us through that period.

PAUL JOHN WERBOS

Washington, D.C.

Contents

PART 2 APPLICATIONS AND EXTENSIONS

Introduction

This book shows how the lowly concepts of *feedback* and *derivatives* are the essential building blocks needed to understand *and replicate* higher-order phenomena like learning, emotion, and intelligence at all levels of the human mind, including the social level. It shows how a relatively new way of computing derivatives—now called "backpropagation"—plays a crucial role in new systems that perform tasks like automatic modeling and prediction, pattern recognition, function approximation, and optimal decision, control and planning. Some of these systems are based on artificial neural networks (ANNs); others are not. As discussed in the Preface, these system designs have *dual use*—as *engineering tools* now used for tasks ranging from reducing pollution through to political forecasting, and as *models* of intelligence.

This book presents a unified approach to this range of topics. Unfortunately, the people studying these various topics have all developed different vocabularies, and have different attitudes towards mathematical details. Reflecting this diversity, the book contains some sections written for the highest-level generalists, and others containing specific pseudocode and equations for people who need to see all the details.

The next part of this Introduction provides a general outline of the book. It describes how the different chapters fit together and fit into the needs of different audiences. The following section explains in general terms how this new mathematical concept—"backpropagation" or "ordered derivatives"—opens up the possibility of a scientific understanding of intelligence, as important to psychology and neurophysiology as Newton's concepts were to physics. The remaining sections put this book into context, by describing more of the history and by providing citations to more advanced work which builds on the fundamentals given here; they do this for the three main topics of this book, in order:

1. Derivative calculation and backpropagation as such, with an emphasis on neural networks.

2. Modeling (prediction and system identification) and intelligent optimization.
3. The methods and foundations of social science and psychology.

A GUIDE TO THIS BOOK

Part One—Chapters 1 through 6—is my 1974 Harvard Ph.D. thesis, the original source for true backpropagation. Chapter 7 contains a 1981 conference paper, also of historical and conceptual significance. Chapters 8, 9 and 10 contain more recent papers, describing the most important developments of the past few years in a tutorial fashion. Recent work on detailed circuits in the brain is discussed briefly and cited, but not included, because of its highly specialized prerequisites.

The entire thesis was written for a Harvard committee containing a political scientist, a statistician, and two applied mathematicians. However, Chapter 2 was scrutinized most carefully by the mathematicians, Chapters 3 and 4 by the statistician, and Chapters 5 and 6 by the political scientist.

Chapter 1 is a general summary. Chapter 2 explains the original backpropagation algorithm, which I then called "dynamic feedback." It also describes a number of possible applications, particularly in modeling or prediction and in optimization. Chapter 2 presents backpropagation as a *general* method, which could be applied to simple artificial neural networks (ANNs) or to any *other* ordered nonlinear system. (It also defines what an ordered system is.) Chapter 3 *applies* backpropagation to the general problem of vector ARMA estimation, which—according to conventional statistical theory—is the proper way to do prediction and modeling when the data we use may be corrupted by measurement noise. Chapters 2 and 3 both try to start out at a relatively low level, and build up gradually to formal proofs which are inherently more difficult. Those proofs are *not* a prerequisite to the remainder of the book.

Chapter 4 describes simulation studies showing how new *robust* methods for prediction over time perform better than the usual methods (ordinary least squares *or* ARMA) still in use by most social scientists and by most neural networkers. This chapter was written at the request of Dr. Mosteller, a statistician, but should be straightforward in any case.

Chapters 5 and 6 were written for the "behavioral scientist"—the psychologist or social scientist. Chapter 5 presents my own views of behavioral science (as of 1974), ranging from the role of backpropagation and neural networks, through to long-term theories of history and their value to the decision maker. (The last section of this Introduction gives an update.) Chapter 6 provides a detailed empirical study of Karl Deutsch's theory of nationalism and social communications, using the modeling tools (based on backpropagation) given earlier in the thesis; this chapter includes both an extension of Deutsch's model, and a test of alternative modeling tools based on real empirical data, further supporting the conclusions of Chapter 4.

The conference paper in Chapter 7 was, in my view, a key link in the causal chain from my thesis to the actual use of backpropagation in the neural network field. (See the later part of this Introduction.) It provides a compressed discussion of three main topics: (1) alternative forms of backpropagation, useful for calculating second derivatives and for time-forwards calculations (called ''forwards propagation'' by some researchers); (2) the use of backpropagation in providing various kinds of *sensitivity analysis* information, crucial to the users of economic models; (3) the use of backpropagation as part of a *reinforcement learning system*, a kind of neural-network optimization system with specific parallels to specific parts of the brain.

Chapter 8 provides a more straightforward, step-by-step tutorial on how backpropagation is used in neural network engineering today. Many engineers have told me that Chapter 8 is the most straightforward tutorial in existence on backpropagation for them, and that it provides certain advanced capabilities—crucial to many applications—which are missing or even muddled up in other popularized accounts. They have also told me that Chapters 8 and 9 are crucial as prerequisites to other sources, such as the *Handbook of Intelligent Control*, which provides the technical details of the most advanced neural net designs for modelling and control in existence, along with the details of many engineering applications [14].

Chapter 9 provides a relatively less technical overview of ''neurocontrol''—the study of neural networks that make decisions or control engineering systems. This chapter mentions engineering applications, but it also mentions how the brain itself is a neurocontroller.

Many papers on neural networks emphasize the dozens upon dozens of applications now on the market; for example, the everyday modem is based on work by Widrow, which is an important precursor to backpropagation. Chapter 9 places more emphasis on applications in the pipeline which could have a big impact on human society, in the context of long-term historical trends (as discussed in Chapter 5). On the positive side, neurocontrol could provide the extra capability needed for a big reduction in the cost of space travel which, in turn, would help permit a Rostow-style economic takeoff in space. It could help solve the complex integration and control problems involved in the transient cycles and startup of fuel cell cars, which might in turn free the world from its dependence on oil. On the negative side, the dangers depicted in movies like *Terminator 2* are more realistic than one might imagine.

Chapter 10 comes back full circle to the human mind, to the basic issues in psychology which inspired all of this work in the first place. Many of us believe that this is the real payoff to this entire enterprise. Chapter 10 provides a deeper, more intuitive explanation of backpropagation, which can be useful even in engineering applications. It also provides a condensed discussion of a very wide range of issues, which may merit a careful line-by-line consideration by those of us concerned with humanistic psychology.

LEARNING AND BACKPROPAGATION AS KEYS TO THE BRAIN

This section explains how new developments in mathematics and engineering—including backpropagation—make it possible to create a ''Newtonian revolution'' in our understanding of intelligence as it exists in the human brain. It will focus on the brain itself—on neuropsychology—because that is the area where theory and experiment have the greatest chance to come together in a precise way in the near future, and because later parts of this book already deal with the larger aspects of human intelligence. As this book goes to print, crucial new experiments are now being reported or initiated; but, because of inertia and overspecialization, both in universities and in government, it is far from clear that this opportunity for a Newtonian revolution will in fact be grasped.

The analogy here is to Issac Newton. Before Newton, physics (or ''natural philosophy'') was essentially an anecdotal science, a very complex melange of historical ideas and empirical observations, without any true unification—like neuroscience today. Newton's new mathematical concept—the *derivative*—was the key to Newton's laws of gravity, which in turn *created* physics as a modern science. Before Newton, natural philosophy *described* nature; after Newton, there were laws of physics used to *explain* nature.

Can we uncover mathematical laws for intelligence in the brain, laws which could play the same role as Newton's laws did in physics, both in unifying and simplifying the subject?

At first glance, most neuroscientists would regard this as a rather wild idea for us to think about today. Most neuroscientists have become overwhelmed by the mind-numbing complexity of the literature on the brain, including the literature on the dynamics of thousands upon thousands of local circuits in the dozens upon dozens of the species they study very carefully. A few researchers have even tried to translate their despair into a general, fundamental principle: if the information content in any *one* brain—the brain under study—is as great as the capacity of the brain trying to study it, then it is impossible for the second brain to hold all that information. Based on this principle, they argue that a scientific understanding of the brain is impossible.

What would Newton have said about this despair? Certainly, the complexity of the human brain at any one time is very great. But so is the complexity of the *entire physical universe*, which is what the physicists study. Newton did not find a clever way of summarizing all our knowledge about the *present state* of the universe; instead, he *changed the focus of attention* toward an effort to understand the *dynamic laws* which *govern* the state of the universe, as it changes over time. The *catalog* of neural circuits—including the recurrent connections and dynamics at any one time—will always be too complex for a human to absorb in full. But the *laws of learning*—the dynamic principles which *change* the strengths of connections between cells (and other long-term parameters)—are a very different matter.

Dozens and dozens of studies of ''plasticity'' or learning have shown that the *higher* centers of the brain—centers like the cerebral cortex, which are responsible for higher intelligence—have a highly flexible and uniform (''modular'') kind of structure. One part of the cortex will learn to take over the functions of another, if it is given the right inputs; for example, cells in temporal cortex (which normally performs language functions in humans) can learn to develop edge detecotrs, when they are hooked up to visual inputs. We should *not* expect that the mathematical laws of learning are the same for *different* organs of the brain, such as the cerebral cortex and the cerebellum, even when similar molecules are involved; however, we can expect that the same basic laws will apply across all cells of similar types within each major organ. *The effort to undestand these laws of learning will be the basic foundation of our Newtonian revolution, if in fact that revolution happens.*

What is the role of engineering and of backpropagation—the use of our new kind of derivative—in that revolution?

In the past, the neural network community has mainly been divided into three or four subgroups, who use similar-looking models at times, but often find it difficult to understand each other because they use *totally different standards of validation* for their models. Biological modellers usually consider how well their models fit local-circuit physiological data. Psychologists (including cognitive scientists) usually study how well their models replicate a few selected aspects of human or animal behavior, at an aggregate input–output level. Engineers study how well different designs actually work in learning to perform difficult tasks. Humanistic psychologists would ask how well different models would fit their direct knowledge of what goes on in their own minds and in the minds of others. Yet a true description of human intelligence should meet *all four validation criteria.* The human brain *works*, as a highly capable engineering system, at least after learning; this is an extremely important piece of information, which we need to exploit to the utmost when developing models.

Chapters 9 and 10 explain in more detail why backpropagating itself is crucial in building any system—artificial or natural—that *works* in certain tasks (like planning or optimization over time) which the human brain does handle. Those chapters, and the following section, cite more detailed discussions of relevant biological experiments. Among the important recent experiments are those showing that NO acts as a reverse transmitter, the work by Sclabassi showing that conventional LTP experiments suppress the nonlinear capabilities of the hippocampus, and the work by David Gardner apparently demonstrating backpropagation-based learning in Aplysia. Among those in the pipeline are engineering-style tests of learning by cells in (co)cultures, such as inferior olive cells at Northwestern University and Purkinje cells trained to learn time-lags at the University of North Texas. Crucial as this work may be, it is still only a small beginning.

HISTORY AND CAPABILITIES OF BACKPROPAGATION

This thesis uses the term ''dynamic feedback'' to describe a method which is now called ''backpropagation.'' No one I know is sure where the word ''backpropagation'' comes from; however, some people speculate that someone