41, 42
Introduction

(1986)
D. E. Rumelhart, G. E. Hinton, and R. J. Williams
Learning internal representations by error propagation

Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 1,
D. E. Rumelhart and J. L. McClelland (Eds.), Cambridge, MA: MIT Press, pp. 318–362

(1986)
David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams
Learning representations by back-propagating errors

Nature 323:533–536

These two papers describe the same model: what has become known in the neural network world as "back propagation," or, informally, "backprop." It is, as of mid-1987, the most popular learning algorithm for working with multilayer networks. It is considerably faster than the other learning algorithm that is successful with multilayer networks, the Boltzmann machine (paper 38).

In addition to the work described here, back propagation was discovered independently in two other places about the same time, an interesting case of simultaneous independent discovery of a solution, once the nature of the learning problem was made explicit (Le Cun, 1986; Parker, 1985). It also turned out that the algorithm had been described earlier by Paul J. Werbos in his Harvard Ph.D. thesis in August 1974 (Werbos, 1974).

We have included two papers on back propagation. The mathematical proof of back propagation involves repeated applications of the chain rule for partial derivatives, and can be confusing. The calculations are done in slightly different ways in the two papers. The short *Nature* paper also contains an elegant computer simulation.

Back propagation is a generalization of the Widrow-Hoff error correction rule (see paper 10). The original Widrow-Hoff technique formed an error signal, which is the difference between what the output is and what it was supposed to be. Synaptic strengths were changed in proportion to the error times the input signal, which diminishes the error in the direction of the gradient (the direction of most rapid change in the error). This algorithm can be realized locally in a neural network. It is not hard to suggest a system where the input, output, and desired output (together forming the error signal) are present at the same synapse. This does not mean that it actually exists—just that it is easy to propose a plausible circuit!

In a multilayer network containing hidden units, that is, units that are neither input nor output units, the problem is much more difficult. The error signal can be formed as before, but many synapses can give rise to the error, not just the ones at the output layers. Since we usually do not know what the hidden units are supposed to do, we cannot directly compute the error signal for hidden units.

The "generalized delta rule" gives a recipe for adjusting the strengths of synapses on internal units based on the error at the output. However, the internal units must be told how large the error is, and how strongly the internal units are connected to the output units in error. (If an internal unit can make no contribution to the error, it obviously need not modify its weights.) This involves running the synapses 'backward' so that the internal unit knows both how strongly it is connected to an output unit and the error at that unit. The internal unit then sums up all its weighted error contributions. It knows the strengths of the inputs it receives, and can modify its synaptic weights according to a rule very similar to that used by the output units:

according to the product of input and summed, weighted errors from higher layers going backward. The implementation of back propagation involves a forward pass through the layers to estimate the error, and then a backward pass modifying the synapses to decrease the error.

Practical implementations of the back propagation algorithm are not difficult, but without modification it is still rather slow, especially for systems with many layers. There is currently a great deal of work being devoted to various ways of speeding up learning, some of which are ingenious, theoretically sound ways of improving the algorithm.

Real synapses do not run backward. Making a plausible physiological model of back propagation is not easy. There are many examples of reciprocal connections between higher and lower cortical levels, but it is unclear that such projections have the correct properties, either anatomical or physiological, to serve in a back propagation network.

One of the most important aspects of both Boltzmann machines and back propagation involves the nature of the representation of information that is formed in the hidden units. The title of paper 41 in the book *Parallel Distributed Processing* places "learning internal representations" first. In the solution to the encoder problem found by the Boltzmann machine (see paper 38) the representations of the input data found by the learning algorithm, when it was successful, looked like binary counting. If a good representation exists, and use of it is necessary in order to solve a problem, then back propagation seems to be able to find it in some situations.

The representations formed in the hidden units may be effective ways of representing the important information in the input signal. It has been proposed by several groups that back propagation can be used to develop algorithms for data compression. Suppose we arrange it so that the input and output layers contain the same number of units, but that there are fewer units in the hidden layers, and that input and output layers have no direct connections. Suppose we wish to train the system so that given an input, it reproduces as accurately as possible the input pattern at the output. If it can successfully do this, then the internal representation at the hidden layers contains adequate information to reconstruct the output to some degree of accuracy. So to communicate the input pattern, we need send only the values of the hidden units.

Understanding the nature of optimal representations is a matter of great interest in both cognitive science and engineering. We have seen other papers in this book where effective internal representations were formed by neural networks, for example learning in visual cortex and topographic organization of cortex. Back propagation suggests a new and powerful way to explore good representations; it is also the most effective current learning algorithm for complex, multilayer systems.

References

Y. Le Cun (1986), "Learning processes in an asymmetric threshold network," *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman Souli, and G. Weisbuch (Eds.), Berlin: Springer.

D. Parker (1985), "Learning Logic," Technical report TR-87, Center for Computational Research in Economics and Management Science, MIT, Cambridge, MA.

P. J. Werbos (1974), "Beyond regression: new tools for prediction and analysis in the behavioral sciences," Ph.D. thesis, Harvard University, Cambridge, MA.