tron. They should not use a complicated model where a simple one will do.

The history of the projection pursuit regression (PPR) model illuminates some of the differences between the two fields. As noted by Cheng and Titterington, the PPR model has the same form as a single layer perceptron. When PPR was introduced into the statistics field by Friedman and Stuetzle in 1981, it did not have much practical impact. There are a number of possible reasons for this. Computationally, it was ahead of its time: many statisticians still do not feel comfortable using very large amounts of computation in an analysis. In addition, statisticians do not often tackle the large prediction problems that can often benefit from such an approach. Finally, the particular fitting (learning) procedure might have been too greedy to work effectively with large number of projections.

In contrast, neural network researchers have developed and applied the PPR model to some difficult problems with considerable success. In recent years, they have further improved their results by applying classical statistical techniques such as regularization, cross-validation and Bayesian modelling. This suggests that both fields should be listening and learning from each other. Cheng and Titterington's paper will help this cause.

# Rejoinder

**Bing Cheng and D. M. Titterington**

We are very grateful to the discussants for the time and effort they have expended in commenting on our paper. When we submitted the revised version of the paper, we felt some trepidation that, in spite of our best effort at brevity, the paper still seemed very long in comparison to many other contributions to the journal, and yet we were fully aware that we had not done justice to important aspects of the field. Fortunately, some of our sins of omission have been absolved by the choice of discussants, and we are happy to regard many of their comments as complementary to our presentation. *Statistical Science* can, therefore, be said to be publishing a 10-author review of the interface between statistics and neural network research rather than a two-author review plus discussion. We are glad that the discussants include representives from what one may call (against Breiman's advice) the mainstream neural-network community (McClelland), as well as distinguished statisticians with both short and long (in terms of time) records of involvement in the area. We apologize to all discussants for not having space to respond to each of the many points they have made.

Later in our rejoinder we shall remark on some points raised by individual discussants, and we shall finish by pulling together views about the future of the interface. First, we mention three areas of research on which several discussants expressed views. These areas were implicitly identified by Amari and, in slightly different form, by Breiman.

- Mathematical modeling of real cognitive processes
- Theoretical investigations of networks and neurocomputing
- Development of useful tools for practical prediction and pattern recognition

## MODELING OF REAL COGNITIVE PROCESSES

The dominant discussant here is McClelland. He emphasizes the fact that machine intelligence still has far to go to emulate many human mental processes, a view echoed by Bienenstock and Geman. McClelland sounds more hopeful than they do that concepts closely akin to artificial networks, presumably as known today, might prove to be key aspects. Furthermore, he suggests that the mechanics of statistics will be important in the development of such realistic cognitive machines: first, manipulation of probability models using Bayes' theorem could be the way to mimic the brain's approach to data analysis ("interpretation"); second, nondeterministic elements seem to be inevitable in modeling any realistic learning process. Practical realization of such models does, however, seem to be a daunting prospect. In the "interpretation" question, for instance, the equivalent of a prior distribution will have to include representation of all useful contextual and background information.

However, it seems clear from McClelland's penultimate paragraph that there are important new developments in areas such as speech processing, even in irritatingly irregular languages such as English and even then in the arguably less irregular American version. We are, nevertheless, doubtful about regarding pronunciation as necessarily statistical. Given enough context, in terms of neighboring letters, the correct pronunciation is presumably determined. Our very naive view would, therefore, be that a system of general rules with exceptions may be more natural; but perhaps the necessary complexity renders the approach too unwieldy.

## MATHEMATICAL THEORY OF NETWORKS AND NEUROCOMPUTING

Major contributors under this heading are Amari and Barron, both of whom feel the lack of theoretical underpinning of certain aspects of the subject, in spite of what Amari calls the superficial, if positive, achievements of ANN models in applications. He mentions various approaches to the question of estimating generalization error, and he reemphasisizes the relevance of information geometry and the EM algorithm in studying and training Boltzmann machines. Also see Titterington and Anderson (1994).

Barron provides important details of the latest results in approximation theory, and his discussion of the case of feed-forward networks with two hidden layers is of particular interest. The ability to bound the risk associated with estimation procedures is of great value, although it appears that much still needs to be achieved in the practical area of bounding the risk associated with data-based model selection criteria: the practitioner wants to know how well he or she is doing in particular applications. Barron emphasizes the fact that the scale of computation time required for network estimation is still a problem and earmarks this area as "the most important task for theoretical research in neural networks." For instance, he solicits theoretical work both to tidy up the fuzzy (in a nontechnical sense) and sometimes contradictory folklore about techniques such as gradient search and to identify whether or not associated optimization techniques can be shown to produce a good solution in a reasonable time.

## USEFUL TOOLS FOR PATTERN RECOGNITION

We shall use this heading to draw together the discussion about the practicalities associated with, in particular, the use of feed-forward networks in prediction and classification.

Breiman sets the scene with a clear description of what he calls the single (hidden) layer feed-forward network. (In the main paper, of course, we use the nomenclature "two-layer" for this architecture.) He highlights the "tinkering and tailoring" approach to network design and the practical awkwardnesses encountered when trying to estimate parameters by optimizing a multi-minima surface. Should one try several starting points and compare the resulting local minima? Should one use a validation set to determine the stopping point of the algorithm? Or, should one prevent overfitting by regularization? These are the sorts of questions to which Barron would like some theoretical answers to reinforce guidelines formulated from empirical studies. They are clearly relevant in complicated optimization problems beyond the estimation of parameters in feed-forward neural networks. However, the practice of optimization is clearly a messy bussiness. Ripley issues caveats about the use of a validation set and comments that the empirical experiences of himself and other investigators are different. At best, this suggests that behavior of optimization procedures in this area depends more than one might like on the particular application. Systematic recommendations may, therefore, be hard to come by, leading us to fall back on Breiman's experience-backed tinkering. To this end, well-designed and carefully assessed empirical studies such as those in Ripley's papers are clearly valuable, especially if they can include a wide range of really large problems.

Amari and Ripley highlight the issues of model complexity and assessment of generalization as important questions, reinforcing feedback we gleaned from our spy at the NATO Advanced Study Institute at Les Arcs. Although these are both theoretical questions, they are clearly related to practice, and we noted with interest the mention of stepwise model construction in the contributions by Barron and, in particular, Breiman. This is surely an important area for development.

Related to this is the question of the resulting classifier's interpretability. This is clearly one area where Ripley feels that multilayer perceptrons fail in comparison to some of their competitors. In addition, he is concerned about the computational demands they make and their complexity relative to other approaches. Classification trees, in contrast, have easily interpreted rules. Sometimes, however, interpretability can be a double-edged sword, especially if the classification tree is interpreted as portraying a physical explanation of the differences among the classes. In multiple regression with many covariates, many regressions based on different subsets of the covariates may provide predictors of comparable abilities; however, some of the models may not include covariates that influence the response in a direct, physical way. Similarly, many classification trees are likely to be closely comparable in terms of performance, and there is no guarantee that the particular one chosen by a computer package represents a rule that directly reveals a physical process.

As promised earlier, we shall come back to some general issues at the end, but we now highlight a few points made by some of the individual discussants.

## AMARI

We appreciated the extra historical perspective provided by Amari. We readily agree that it is wrong to lay all the blame for the dark period on Minsky and Papert. We are grateful for the unsurprising information that some of the basic ideas appeared in papers that predate the most familiar references. We too have recently read with interest the work of Jordan and Jacobs (1993), and we have a general interest in the various roles that the EM algorithm plays in this area.

## BIENENSTOCK AND GEMAN

The comments of Bienenstock and Geman will strongly influence the final part of our discussion. Here, we merely note their discussion of the appeal of generalization and highlight their warning that it seems virtually impossible that adequate training sets will be available for the training of such sophisticated devices as fully successful object-recognizers. The demands of generalizability are too high. Bienenstock and Geman recommend that future emphasis should be on modeling rather than training. This is somewhat related to Ripley's point about using a family of functions well rather than worrying about which family of functions to choose.

## BREIMAN

We apologize to Breiman that we have continued to use the phase "neural-network community" even in this rejoinder if only for the fact that we prefer not to write "nonmainstream-statistical community"! Breiman and others emphasize the problem-oriented approach of the "other" community. One might be indignant, sad or indifferent about the fact that they have the "good fortune not to have any formal statistical training". We hope (and assume) that this is not meant to imply that formally trained statisticians should not try to get involved!

## RIPLEY

We are glad that Ripley has included the formulation of a more general projection-pursuit regression, and we are grateful for the recent references on various topics, including handwritten digit recognition. To the latter, we contribute the recent special issues of *Pattern Recognition* (March 1993) and *Pattern Recognition Letters* (April 1993). No doubt, several dozen further references relevant to our review will appear before its publication date: a measure of the speed of current development at this interface!

## TIBSHIRANI

We look forward with great interest to the fruits of the Hinton-Tibshirani collaboration. We note the two-way flow of benefits between the two communities and are mildly surprised that statisticians appear to have something to gain in more ways than do neural network researchers. The two points labelled "1" have bearing on the final part of our rejoinder. So, in a way, does point 6 about statisticians being good self-sellers, which echoes the spirit of remarks of Bienenstock and Geman. At the risk of offending many nonarchetypal (and other) friends on both sides of the Atlantic, we should be less surprised about point 6 if most statisticians were British and most neural-network researchers non-British. Perhaps statisticians have just been around longer and are perceived as nondynamic; or perhaps, yet again, the title of their profession engenders an image of unimaginativeness.

## STATISTICS AND NEURAL NETWORKS: THE WAY FORWARD

In this final part of our rejoinder, we try to formulate a perspective of the future synergy, if any, between research in statistics and neural networks.

The development and application of (artificial) neural networks has clearly been explosive and accompanied by much hype. Hype can cause two types of reaction. First, it can, like any successful and energetic advertising campaign, stimulate great interest and many acolytes as a result of appealing packaging, ambitiously stated goals and apparently successful applications, as described by Bienenstock and Geman. On the other hand, hype can be off-putting. Breiman appears to have been affected this way initially, and we must confess that an instinctive suspicion of glossy advertising was the stimulus of our early reading in the area a few years ago. We felt that there must be something of statistical interest going on but surely nothing fundamentally novel.

At one level, the conclusion is anticlimactic. In the context of classification, in particular, neural-network models provide nonlinear predictors that are, under certain weak conditions, universal approximators. However, they both overlap with procedures that are well known to statisticians and, in many applications, seem to have no advantage over simpler methods. In addition, implementation of the predictors in practice appears to involve either simply recourse to a black-box application or a development process that is much less systematic than would satisfy statisticians. We might, therefore, decide to refer to the lessons offered by the comparative experiments of Ripley that they would be better advised to use different tools, and abandon the field.

We feel that this would surely be a mistake. While feed-forward networks, trained by the generalized delta rule, may not be cure-all classifiers let alone a realistic prototype for real neural structures, some of the contexts in which they have been used involve data of great volume and complexity. It is clear that the frontiers of complexity will continue to be attacked and, in principle, statisticians ought to be involved. Perhaps, as Breiman and Tibshirani indicate, statisticians will have to stray from their traditional paradigms in order to make meaningful impact, and many people will, no doubt, be reluctant to do so. Perhaps the paradigms can be suitably adapted; they are, after all, increasingly permeating the neural-network literature. In any case, many of the underlying problems of interest are of deep practical significance. As Bienenstock and Geman remark, they are attracting extremely able people who will inevitably have very clever ideas. It would be unworthy of statisticians to dismiss all of these as being too ad hoc, and it would certainly be foolish to be so blinkered as not to become informed about and involved in the key developments in this area.

We are pleased that all the discussants were positive about the involvement of statisticians at the interface, that McClelland saw the possible benefit of this to the cognitive science community and that the others indicated that statisticians would be enriched by participating. It seems to us that much of the challenge of the future has to involve the treatment of very large-scale problems, that whatever develops from current neural-network research should not be cursorily ignored and that statisticians have a contribution to make.

## ADDITIONAL REFERENCES

AMARI, S. (1967). Theory of adaptive pattern classifiers. *IEEE Trans.* **16** 299–307.

AMARI, S. (1972). Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Trans. Comput.* **21** 1197–1206.

AMARI, S. (1977). Neural theory of association and concept-formation. *Biol. Cybernet.* **26** 175–185.

AMARI, S. (1985). *Differential-Geometrical Methods of Statistics. Lecture Notes in Statist.* **28**. Springer, Berlin.

AMARI, S. (1993a). Mathematical methods of neurocomputing, In *Chaos and Networks — Statistical and Probabilistic Aspects* (O.E. Barndorff-Nielsen et al., eds.) Chapman and Hall, London.

AMARI, S. (1993b). A universal theorem on learning curves. *Neural Networks* **6** 161–166.

AMARI, S. and MURATA, N. (1993). Statistical theory of learning curves under entropic loss criterion. *Neural Computation* **5** 140–153.

AMARI, S., YOSHIDA, K. and KANATANI, K. (1977). A mathematical foundation for statistical neurodynamics. *SIAM J. Appl. Math.* **33** 95–126.

ANDERSON, J. A. (1972). A simple neural network generating interactive memory. *Math. Biosci.* **14** 197–220.

ANTHONY, M. and BIGGS, N. L. (1992). *Computational Learning Theory: An Introduction.* Cambridge Univ. Press.

BARRON, A. R. and COVER, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory* **37** 1034–1054.

BERTSIMAS, D. and TSITSIKLIS, J. (1993). Simulated annealing. *Statist. Sci.* **8** 10–15.

BISHOP, C. (1992). Exact calculation of the Hessian matrix for the multilayer perceptron. *Neural Computation* **4** 494–501.

BLUMER, A., EHRENFEUCHT, A., HAUSSLER, A. and WARMUTH, M. (1989). Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.* **36** 926–965.

BOSER, B., GUYON, I. and VAPNIK, V. (1992). A training algorithm for optimal margin classifiers, *Digest Fifth ACM Workshop on Computational Learning Theory*, 144–152. ACM, New York.

BUNTINE, W. L. and WEIGEND, A. S. (1993). Calculating second derivatives on feed-forward networks. *IEEE Trans. Neural Networks*. To appear.

CHERKASSKY, V., FRIEDMAN, J. H. and WECHSLER, H., eds. (1994). *From Statistics to Neural Networks. Theory and Pattern Recognition Applications.* Springer, New York.

COLTHEART, M., CURTIS, B., ATKINS, P., and HALLER, M. (1994). Models of reading aloud: Dual-route and parallel distributed processing approaches. *Psychological Review.* To appear.

CYBENKO, G. (1988). Continuous valued neural networks with two hidden layers are sufficient. Technical Report, Dept. Computer Science, Tufts Univ.

DENBY, B. (1993). The use of neural networks in high-energy physics. *Neural Computation* **5** 505–549.

DEVIJVER, P. A. and KITTLER, J. V. (1982). *Pattern Recognition. A Statistical Approach.* Prentice-Hall, Englewood Cliffs, NJ.

DIACONIS, P. and SHAHSHAHANI, M (1984). On non-linear functions of linear combinations. *SIAM J. Sci. Statist. Comput.* **5** 175–191.

DIETTERICH, T. G. and BAKIRI, G. (1991). Error-correcting output codes: a general method for improving multiclass inductive learning programs. In *Proceedings of the Ninth National Conference on Artificial Intelligence*. AAAI Press.

FARLOW, S. J. (1984). *Self-Organizing Methods in Modeling: GMDH Type Algorithms*. Dekker, New York.

FUKUNAGA, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic, London.

GALLANT, S. L. (1993). *Neural Network Learning and Expert Systems.* MIT Press.

GORMAN, R. P. and SEJNOWSKI, T. J. (1988a). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks* **1** 75–89.

GORMAN, R. P. and SEJNOWSKI, T. J. (1988b). Learned classification of sonar targets using a massively parallel network. *IEEE Transactions on Acoustics, Speech and Signal Processing* **36** 1135–1140.

GREEN, D. M. and SWETS, J. A. (1966). *Signal Detection Theory and Psychophysics.* Wiley, New York.

GROSSBERG, S. (1976). Adaptive pattern classification and universal recording. *Biol. Cybernet.* **23** 121–134.

GROTHER, P. J. and CANDELA, G. T. (1993). Comparison of handprinted digit classifiers. U.S. National Institute of Standards and Technology report NISTIR 5209.

HAUSSLER, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. and Comput.* **100** 78–150.

HWANG, J.-N., LAY, S.-R., MAECHLER, M., MARTIN, D. and SCHIMERT, J. (1993). Regression modeling in backpropagation and projection pursuit learning *IEEE Trans. Neural Networks*. To appear.

HWANG, J.-N., LI, D., MAECHLER, M., MARTIN, D. and SCHIMERT, J. (1992a). A comparison of projection pursuit and neural network regression modeling. In *Advances in Neural Information Processing Systems 4* (J. E. Moody, S. J. Hanson and R. P. Lippmann, eds.), 1159–1166. Morgan Kaufmann, San Mateo, CA.

HWANG, J.-N., LI, D., MAECHLER, M., MARTIN, D. and SCHIMERT, J. (1992b). Projection pursuit learning networks for regression. *Engineering Applications Artificial Intelligence* **5** 193–204.

JERVIS, T. T. and FITZGERALD, W. J. (1993). Optimization schemes for neural networks. Cambridge Univ. Engineering Dept. Report CUED/F-INFENG/TR144.

JORDAN, M. I. and JACOBS, R. A. (1993). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*. To appear.

KNERR, S., PERSONNAZ, L. and DREYFUS, G. (1992). Handwritten digit recognition by neural networks with single-layer training. *IEEE Trans. Neural Networks* **3** 962–968.

KOHONEN, T. (1972). Correlation matrix memories. *IEEE Trans. Comput.* **C-21** 353–359.

KOHONEN, T. (1984). Self-organized formation of topologically correct feature maps. *Biol. Cybernet.* **43** 59–69.

KRUSCHKE, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* **99** 22–44.

MACWHINNEY, B. and LEINBACH, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition* **40** 121–158.

MARTIN, G. L. and PITMAN, J. A. (1990). Recognizing handprinted letters and digits. In *Advances in Neural Information Processing Systems 2* (D.S.Touretzky, ed.) 405–414. Morgan Kaufmann, San Mateo, CA.

MARTIN, G. L. and PITMAN, J. A. (1991). Recognizing handprinted letters and digits using backpropagation learning. *Neural Computation* **3** 258–267.

MASSARO, D. W. (1975). *Experimental Psychology and Information Processing*. Rand-McNally, Chicago.

MASSARO, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology* **21** 398–421.

MCCLELLAND, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology* **23** 1–44.

MCCLELLAND, J. L. and ELMAN, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology* **18** 1–86.

MCCLELLAND, J. L. and RUMELHART, D. E. (1981). An interactive activation model of context effects in letter perception, Part I: An account of basic findings. *Psychological Review* **88** 375–407.

MCCLELLAND, J. L., ST. JOHN. M. and TARABAN, R. (1989). Sentence comprehension: A parallel distribution processing approach. *Language and Cognitive Processes* **4** 287–335.

MEDIN, D. L. and SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review* **85** 207–238.

MILLER, W. T. III, SUTTON, R. S. and WERBOS, P. J., eds. (1990). *Neural Networks for Control*. MIT Press.

MØLLER, M (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* **6** 525–533.

NAKANO, K. (1972). Association—A model of associative memory, *IEEE Trans. Systems Man Cybernet.* **SMC-2** 381–388.

PEARLMUTTER, B. A. (1994). Fast exact multiplication by the Hessian. *Neural Computation*. **6** 147–160.

PINKER, S. and PRINCE, A. (1988). On language and connectionism: Analysis of a parallel distribution processing model of language acquistion. *Cognition* **28** 73–194.

PLAUT, D. C. and MCCLELLAND, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* 824–829. Erlbaum, Hillsdale, NJ.

QUINLAN, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

RIEDMILLER, M. and BRAUN, H. (1992). RPROP—a fast adaptive learning algorithm. Technical Report, Univ. Kahlsruhe.

RIPLEY, B. D. (1993b). Neural networks and flexible regression and discrimination. In *Statistics and Images* (K.V. Mardia, ed.) *Advances in Applied Statistics* **1**. Carfax, Abingdon.

RIPLEY, B. D. (1994b). Flexible non-linear approaches to classification. In *From Statistics to Neural Networks. Theory and Pattern Recognition Applications* (V. Cherkassky, J. H. Friedman and H. Wechsler, eds.). Springer, New York.

RIPLEY, B. D. and HJORT, N. L. (1994). *Pattern Recognition and Neural Networks—A Statistical Approach*. Cambridge Univ. Press.

ROBERTS, S. and TARASSENKO, L. (1993). Automated sleep EEG analysis using an RBF network. In *Neural Network Applications* (A.F. Murray, ed.). Kluwer, Boston.

ROBERTS, S. and TARASSENKO, L. (1994). A probabilistic resource allocating network for novelty detection. *Neural Computation*. To appear.

RUMELHART, D. E. (1977). Understanding and summarizing brief stories. In *Basic Processes in Reading: Perception and Comprehension* (D. LaBerge and S. J. Samuels, eds.) 265–303. Erlbaum, Hillsdale, NJ.

RUMELHART, D. E. and MCCLELLAND, J. L. (1986). On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume II* (J. L. McClelland, D. E. Rumelhart and the PDP Research Group, eds.) 216–271. MIT Press.

SCHIFFMANN, W., JOOST, M. and WERNER, R. (1992). Optimization of the backpropagation algorithm for training multilayer perceptrons. Preprint, Institute of Physics, Univ. Koblenz.

SEGRE, A. and GORDON, G. (1993). Book review: *Computer Systems that Learn* by S.M. Weiss and C.A. Kulikowski. *Artif. Intell.* **62** 363–378.

SIMARD, Y. S., LE CUN, Y. and DENKER, J. (1993). Efficient pattern recognition using a new transformation distance. *Neural Information Processing Systems*, **5**. Morgan Kaufman, San Mateo, CA.

TAKEUCHI, A. and AMARI, S. (1979). Formation of topographic maps and columnar microstructures. *Biol. Cybernet.* **35** 63–72.

THORNTON, C. J. (1992). *Techniques in Computational Learning. An Introduction*. Chapman and Hall, London.

TITTERINGTON, D. M. and ANDERSON, N. H. (1994). Boltzmann machines. In *Probability, Statistics and Optimization: A Tribute to Peter Whittle* (F. P. Kelly, ed.). Wiley, Chichester. To appear.

TSYPKIN, Y. Z. (1973). Foundations of the theory of learning systems. Academic, New York.

VAPNIK, V. (1992). Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems 4* (J. E. Moody, S. J. Hanson and R. P. Lippmann, eds.) 831–838. Morgan Kaufmann, San Mateo, CA.

VARDI, Y. and LEE, D. (1993). From image deblurring to optimal investments: maximum likelihood estimation for positive linear inverse problems. *J. Roy. Statist. Soc. Ser. B* **55** 569–612.

VON DER MALSBURG, Ch. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* **14** 85–100.

WASSERMAN, P. D. (1993). *Advanced Methods in Neural Computing.* Van Nostrand Reinhold, New York.

WEIGEND, A. S. and GERSHENFELD, N. A., eds. (1993). *Times Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley, Reading, MA.

WEISS, S. M. and KULIKOWSKI, C.A. (1991). *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems.* Morgan Kaufmann, San Mateo, CA.

WOLPERT, D. H. (1990). Constructing a generalizer superior to NETtalk via a mathematical theory of generalization. *Neural Networks* **3** 445–452.

ZHAO, Y. and ATKESON, C. G. (1992). Some approximation properties of projection pursuit learning networks. In *Advances in Neural Information Processing Systems 4* (J. E. Moody, S. J. Hanson and R. P. Lippmann, eds.) 936–943. Morgan Kaufmann, San Mateo, CA.