Comment

Andrew R. Barron

Relationships between topics in statistics and artificial neural networks are clarified by Cheng and Titterington. There are fruitful concepts in artificial neural networks that are worthwhile for the statistical community to absorb. These networks provide a rich collection of statistical models, some of which are ripe for both mathematical analysis and practical applications. Many aspects of artificial neural networks are in need of further investigation. Here, I comment on approximation and computation issues and their impact on statistical estimation of functions.

APPROXIMATION

Attention is focussed on the most commonly studied feedforward networks (perceptrons) which have one or two "hidden" layers defined by composition of units of the form $\phi(wx + w_0)$, where ϕ is a hardlimiter or sigmoidal activation function and w_0, w denote the parameters (internal weights) that adjust the orientation, location and scale of the unit functions (Rosenblatt, 1962; Rumelhart, Hinton and Williams, 1986a). In the one hidden layer case, a linear combination of such units is taken with the internal weights adjusted so that the result approximates a target function. These networks may be regarded as an adjustable basis function expansion of ridge form similar to projection pursuit (Friedman and Stuetzle, 1981) and similar to sparse trigonometric series with adjustable frequency vectors. Linear combination of such adjustable basis functions can provide an accurate approximation with far fewer units than by linear combination of any fixed basis functions for certain classes of target functions when the number of input variables is greater than or equal to three (Barron, 1993). A consequence is that more accurate statistical function estimation is possible for such target functions (Barron, 1994).

These conclusions for one hidden layer networks are based, in part, on the following result developed in Jones (1992) and Barron (1993). Suppose a function $f(x)$ is such that $f(x)/V$ is in the closure of the convex hull of the set of units $\{\pm\phi(wx + w_0) : (w_0, w) \in R^{d+1}\}$, where ϕ is bounded by 1 for some positive number V. The closure is in the $L_2(\mu)$ norm, where μ is any given probability measure μ with bounded support on R^d. Then there are M such units with choices of weights depending on f and μ, such that their linear combination $f_M(x)$ (a single hidden layer network) achieves approximation error

$$||f - f_M|| \leq \frac{V}{\sqrt{M}},$$

where the norm is taken in $L_2(\mu)$. The surprising aspect is that the approximation rate as a function of M is independent of the dimension d. A subclass of functions that satisfy the condition are those that possess a bound on the first moment of the Fourier magnitude distribution. (This class includes all smooth positive definite functions and convex combinations of translates of such functions.) In contrast, approximation using any fixed M basis functions cannot achieve approximation error uniformly better than order $1/M^{1/d}$ for the same class of functions f, taking μ to be the uniform distribution on a d-cube (Barron, 1993).

It is of interest to characterize what classes of functions can be more parsimoniously approximated using two rather than one hidden layer in the network. Some functions such as the indicator of a cube or a ball are not accurately approximated by the ridge expansions represented by one-layer networks without resorting to a number of units exponentially large in the dimension. In these cases the network capabilities may be improved by inclusion of a second layer of threshold nonlinearities. Units on the second layer can provide indicators of the level sets of linear combinations of the first layer units. These level sets can be arranged to take arbitrary polygon shapes (Lippman, 1987). The linear combination of the outputs of the second layer then give piecewise constant approximations of a rather general form. One conclusion of the same flavor as above is that if a function f is such that $f(x)/V$ is in the closure of the convex hull of the set of signed indicators of K-sided polygons for some positive V, then there is a two hidden layer network function $f_{K,M}(x)$ with KM units on the first layer and M units on the second

Andrew R. Barron is Professor, Department of Statistics, Yale University, Box 2179, Yale Station, New Haven, Connecticut 06520.

layer such that $||f - f_{K,M}|| \leq V/\sqrt{M}$. It is not clear yet how much more general a class of functions this is than those in the convex hull of signed indicators of half-spaces. Another approach to examining approximation by two hidden layer networks is in Cybenko (1988). He shows that by using sigmoidal activation functions the second layer units can be arranged to implement localized kernel functions that are then linearly combined to provide the function approximation. He shows that the approximation error tends to zero but does not give a bound on the rate. It is not clear that localized basis expansions will be effective in high dimensions. Nevertheless, two hidden layer networks may provide one way to combine the positive benefits of global ridge approximations and local kernel approximations.

ESTIMATION

These multiunit perceptrons are nonlinearly parameterized models incorporated into least squares regression, classification and likelihood maximization. By combining results on network approximation with analysis of statistical risk, it is possible to bound the accuracy of neural network estimators in certain cases.

Frameworks exist for the analysis of the total risk of function estimation using neural networks or other nonlinear models for various choices of loss function. Analogous to the bias-variance decomposition of the mean squared error, the problem decomposes into separate consideration of the approximation error and the additional error due to estimation of the function from a finite sample (see, for instance, Haussler, 1992; Barron, 1991). With squared error loss, the estimation error can be bounded by the ratio of the number of parameters to the sample size times a logarithmic factor. The best rate of convergence for a network estimator occurs when the size of the network (indexed by the number of parameters) is chosen so that the estimation error is of the same order as the approximation error. In particular, the general risk bounds are applied to the case of one hidden layer networks in Barron (1994). There conditions are given such that the risk is bounded by

$$E||f - \hat{f}||^2 \leq O\left(\frac{V^2}{M} + \frac{Md}{N}\log N\right),$$

where M is the number of units, d is the input dimension, N is the sample size and V is as discussed above. This risk bound is of the order $V^2((d/N)\log N)^{1/2}$ with $M \sim (N/(d\log N))^{1/2}$. Thus, a satisfactorily small statistical risk is possible without requiring an exponentially large sample size. The estimator $\hat{f}$ that achieves these bounds is assumed to correspond to a global optimum of the empirical squared error loss, among one hidden layer networks with M units subject to certain constraints on the parameter values. It can be shown, under similar conditions, that the same risk bounds hold for any estimator that achieves an empirical squared error not larger than a prescribed value determined by the bound on the approximation error.

Since, in general, the network approximation error is not known in practice, data-based model selection criteria are useful to select a size of network that achieves approximately the best convergence rate permitted by the class of models. Such risk bounds are available for networks selected by certain complexity based criteria (Barron and Cover, 1991; Barron, 1991). It is an open problem whether risk bounds can be developed for networks selected by other criteria such as Akaike's AIC; such bounds would be analogous to the results available for linear models by Shibata (1981) and Li (1987).

COMPUTATION

In some cases, optimization of the appropriate objective function is proven to provide accurate estimators in the sense of statistical risk, as discussed above. However, there is no known algorithm for network estimation that is proven to produce accurate estimates of functions in a feasible amount of computation time. At the least, we should avoid having an average computation time that is exponential in the input dimension d. Ideally, the computation time should be bounded by small degree polynomial in N and d while achieving a satisfactory statistical risk bound (e.g., a fractional power of d/N) for a sensible class of target functions, where N is the sample size. It is not known whether such a feasible algorithm exists. Because of its potential practical implications, I regard the resolution of problems of this type as the most important task for theoretical research concerning neural networks.

Various algorithms have been suggested or used in practice that may or may not be appropriate for the function estimation task. Here, some of the standard approaches and associated problems are briefly mentioned. Many of the methods involve numerical search for an optimum of an empirical objective function. Unfortunately, this error surface for multiunit perceptrons is extremely multimodal as a function of the parameters (weights).

Gradient search and many of its variants, such as back-propagation, produce a local optimum of dubious scientific merit. The use of multiple starting points may rescue local search strategies, but it should be mathematically determined whether or

not the number of restarts needed on the average is exponential in the size of the problem. The objective function may be regularized by the addition of a large enough convex penalty term (weight decay term) to reduce multimodality, but can it be demonstrated whether the function estimates remain statistically accurate in that case? A concern is that if a penalty term is multiplied by a constant large enough to guarantee convexity of the objective function, then the effect of the empirical loss term may be washed out.

Stochastic search strategies such as simulated annealing or guided random search can avoid traps of local optima to converge to a global optimum, but it needs to be proven whether an accurate estimate is reached in feasible time for perceptrons. See Bertsimas and Tsitsiklis (1993) for some of the issues associated with proving a computation rate for simulated annealing. Convergence theory for random search should reveal what advantage, if any, the search strategy has over exponential time algorithms such as exhaustive search over a suitable grid.

Likelihood maximization can be replaced by averaging with respect to a Bayesian posterior distribution using importance sampling or Metropolis algorithms, but it is not proven whether these algorithms will provide suitable solutions in feasible time for highly multimodal surfaces. Indeed, suppose it were not feasible to find points of high likelihood that provide an accurate estimator. It would then be surprising (but not necessarily impossible) for an averaging technique to produce an accurate estimator.

The computational task is simplified by certain estimation strategies that build up a network one unit at a time. At each stage, the parameters of a new unit are to be determined given that the smaller network has been estimated. In some cases, convex objective functions can be defined that are readily optimized at each stage. One such class of network methods use compositions of small polynomial units, each of which is linearly parameterized and optimized by least squares (Farlow, 1984; Barron and Barron, 1988). Another approach involves logistic sigmoidal units optimized by a relative entropy criterion; see below. It needs to be determined under what conditions functions can be accurately approximated by such iteratively constructed networks.

Some progress has been made in the case of a single hidden layer network with a squared error criterion. Optimizing such networks one node at a time provides a lower dimensional multimodal search task while still permitting an accurate approximation (Jones, 1992; Barron, 1993). In particular, suppose a function f is such that $f(x)/V$ is in the closure of the convex hull of the set of functions $\phi(wx + w_0)$ (and for simplicity, assume odd symmetry $\phi(-z) = -\phi(z)$). Let $f_0(x) = 0$ and for $M = 1, 2, \ldots$ iteratively define $f_M(x) = v_1 f_{M-1}(x) + v_2\phi(wx + w_0)$, where the internal weights w_0, w of the Mth unit are found to maximize the inner product of the function $r_{M-1}(x)$ and $\phi(wx+w_0)$, where $r_{M-1}(x) = f(x) - f_{M-1}(x)$ and then the external linear weights v_1, v_2 are optimized by ordinary least squares. Then $\|f - f_M\| \leq 2V/\sqrt{M}$ which is the same order bound as stated above for noniterative approximation. Thus, the search has been reduced from $M(d + 2)$ dimensions down to $d + 1$ dimensions, but the objective function still may have multiple modes for each M. It remains to determine whether it is possible to provide approximate solutions to this simpler optimization (perhaps by a stochastic search or multistart algorithm) in a time that is not exponentially large in d.

An interesting approach worthy of further study is to choose w_0, w for unit M to minimize the average binary relative entropy $D(g, \phi) = g \log g/\phi + (1-g)\log(1-g)/(1-\phi)$ between the functions $g(x) = 1/2 + r_{M-1}(x)/2V$ and $\phi(wx + w_0)$, with ϕ chosen to be the logistic sigmoid $\phi(z) = e^z/(1 + e^z)$ and $r_M(x) = f(x) - f_M(x)$. With this choice, the objective function is strictly convex in w_0, w and an approximate minimizer is readily computed for each M by gradient or Gauss-Newton search as in logistic regression. Now $r_M(x) = 0$ is a fixed point of these iterations. It may be possible to prove that $f - f_M$ tends to zero as $M \to \infty$. Does it have the same $1/\sqrt{M}$ approximation rate? The problem of computational feasibility of accurate network estimation would be solved by the positive resolution of this approximation question.

SUMMARY

I concur with the conclusions of Cheng and Titterington that research in statistics and artificial neural networks is mutually beneficial and that increased awareness of work in the respective disciplines should be encouraged. It should be important to each field not only to acknowledge existing work from both fields but also to put it to use to advance the state of the art. Combined use of approximation theory, mathematical statistics and computation theory are essential to the treatment of fundamental problems of function estimation and neural networks.