New Developments and Applications in Experimental Design
IMS Lecture Notes – Monograph Series (1998) Volume 34

OPTIMAL DESIGN FOR NEURAL NETWORKS

BY LINDA M. HAINES

University of Natal

In this paper the statistical principles underlying hidden-layer feed-forward neural networks are introduced and are invoked to develop strategies for the construction of appropriate optimal experimental designs. The ideas are illustrated by means of a simple network involving single input and output neurons and two neurons in the hidden layer. Locally and Bayesian optimal designs are obtained for the underlying nonlinear model and in particular it is shown that the relevant Bayesian criteria can be estimated from samples generated using Markov chain Monte Carlo methods.

1. Introduction. Neural networks are models abstracted from certain functions of the brain, and are proving to be valuable and exciting tools for solving problems in a diversity of areas such as economics, medicine, and psychology. The focus of the present paper is on hidden-layer feed-forward neural networks which are used extensively to model regression and classification data and, in particular, on the issue of choosing experimental data for these networks so that the fitted curve or surface is in some sense optimal. This problem of optimal design, also referred to within the neural network literature as "active data selection" and "query-based learning", is of some current interest. For example, Baum (1991) and Hwang, Choi, Oh and Marks (1991) provide heuristic procedures for sequentially selecting data, while MacKay (1992a, b), Plutowski and White (1993), Williams, Qazaz, Bishop and Zhu (1995) and Cohn (1996) draw closely on statistical notions to develop optimal strategies in which points are added "one-at-a-time" to the existing data. In addition Sollich (1994) provides a broad and fascinating framework for design within the context of statistical physics.

The aim of the present study is to construct optimal designs for nonlinear regression models describing hidden-layer feed-forward neural networks. Some necessary statistical insights are provided in Section 2 and designs for a specific example which are optimal in a classical and in a Bayesian sense are presented in Sections 3 and 4 respectively. Some broad conclusions and pointers for future research are given in Section 5.

Received September 1997; revised January 1998.

AMS 1991 subject classifications. Primary 62K05, 62J02; secondary 62G07, 62L05, 65C05.

Key words and phrases. Hidden-layer feed-forward neural networks, nonlinear regression, nonparametric regression, optimal design, Markov chain Monte Carlo.

2. Preliminaries. Consider a set of simple regression data, $(x_i, y_i), i = 1, \ldots, n$, with x-values taken from some design space, $\mathcal{X}$. Suppose that a hidden-layer feed-forward neural network, with an input layer comprising a neuron to accommodate the explanatory variable, x, and a bias or constant term, a single hidden-layer comprising two neurons with logistic activation functions together with a bias term, and a single neuron with a linear activation function in the output layer, is invoked to fit a curve to these data. The network, together with its connecting weights, is shown in Figure 1 and it follows immediately that the output, o, corresponding to a given input, x, can be expressed as

$$o = \beta_0 + \frac{\beta_1}{1 + e^{-(\gamma_{01} + \gamma_{11} x)}} + \frac{\beta_2}{1 + e^{-(\gamma_{02} + \gamma_{12} x)}}. \tag{2.1}$$

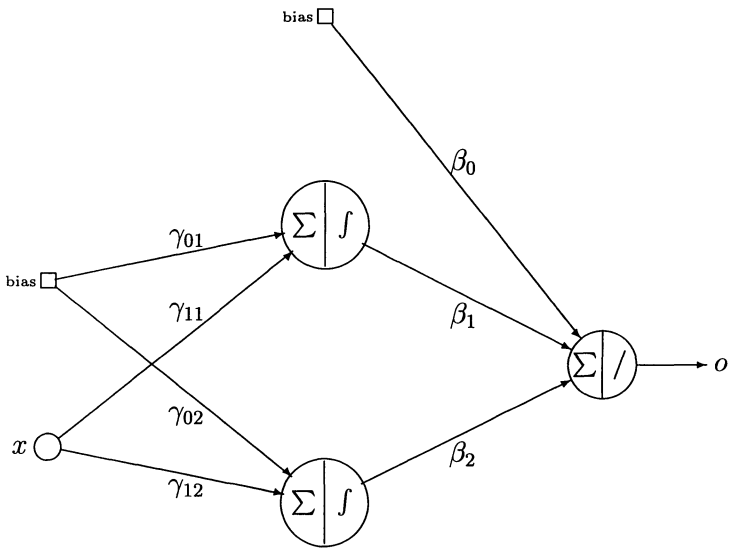

FIG. 1. *An example of a hidden-layer feed-forward neural network. (The neurons are displayed as circles and the connecting weights are indicated on the appropriate edges linking the neurons. The symbol, Σ, represents summation of the inputs into a neuron and the symbols, $\int$ and /, logistic and linear activation functions respectively.)*

Suppose further that the network is "trained" or optimized by minimizing the error sum-of-squares, $\sum_{i=1}^{n}(y_i - o_i)^2$, with respect to the unknown connecting weights. Then it is clear that this process involves, in effect, the fitting of a nonlinear model, $y_i = o_i + \epsilon_i, i = 1, \ldots, n$, with independent error terms, ϵ_i, having mean, 0, and constant variance, σ^2, to the data using the principle of least squares, and hence that the underlying mechanism of the neural network is that of nonlinear regression. It is also clear however that this nonlinear model is not in anyway meaningful, but rather that two sigmoidal curves are scaled and located, and an appropriate constant term is added to them, in such a way as to best fit the data. The overall process is therefore,

essentially, one of nonparametric regression. These notions are well-known and have been extensively discussed and developed in the literature [Bishop (1995), Ripley (1996) and Brittain and Haines (1997)].

Attention here is focused on the nonlinear regression models describing hidden-layer feed-forward neural networks, and in particular, and for illustrative purposes, on the model associated with the network shown in Figure 1. It is firstly important to observe that the same model function (2.1) can be obtained by permuting the two hidden-layer neurons and also by changing the signs of the slope and location parameters describing the logistic functions with a concomitant and appropriate adjustment of the scaling and constant terms. As a consequence, the parameters of the model are not identifiable and, specifically, there is a total of $2 \times 2^2 = 8$ redundancies [see Bishop (1995), p. 133 and Ripley (1996), p. 159]. It is sensible therefore to introduce a reparametrization of the model which involves hyperbolic tangent rather than logistic activation functions as

$$o = \eta(x, \theta) = \theta_5 + \theta_6 \tanh\{\theta_2(x - \theta_1)\} + \theta_7 \tanh\{\theta_4(x - \theta_3)\}, \tag{2.2}$$

where $\theta = (\theta_1, \ldots, \theta_7)$ and, in addition, to assume that $\theta_1 < \theta_3$, thereby ordering the location of the hidden neurons, and to take the scale parameters, θ_6 and θ_7, to be positive. The problem of parameter non-identifiability can then be removed by adopting this reparametrized and restricted model.

The information matrix for the parameters, θ, at a design point, $x \in \mathcal{X}$, for the nonlinear regression model specified by (2.2) and having error terms independently distributed as $N(0, 1)$ is given by

$$M(x, \theta) = \left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)\left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)^T,$$

where $\partial \eta(x, \theta)/\partial \theta$ is the gradient of $\eta(x, \theta)$ with respect to θ and where the variance σ^2 is assumed without loss of generality to be 1. It thus follows that the information associated with an approximate design, μ, which is a measure on the space, $\mathcal{X}$, placing probabilities, μ_i, at the distinct design points, $x_i, i = 1, \ldots, n$ respectively, can be succinctly expressed as

$$M(\mu, \theta) = \sum_{i=1}^{n} \mu_i M(x_i, \theta).$$

In constructing optimal designs for nonlinear models it is usual to adopt criteria which are based on the information matrix and which are, at the same time, practically meaningful. In the present context, the parameters of the model are essentially artifacts in the fitting process and interest centers on the predicted response at a design point, x, and, more particularly, on the asymptotic variance of such a prediction given by

$$d(x, \mu, \theta) = \left[\frac{\partial \eta(x, \theta)}{\partial \theta}\right]^T [M(\mu, \theta)]^{-1} \left[\frac{\partial \eta(x, \theta)}{\partial \theta}\right].$$

There are many criteria for optimal design which incorporate this variance, and attention here is restricted primarily to the criterion of G-optimality, for which the maximum

value of $d(x, \mu, \theta)$ over the design space, $\mathcal{X}$, is minimized, and to its Bayesian analogue. The associated criterion of D-optimality, for which the determinant of the information matrix is maximized and which thus, in a sense, provides the best fitting curve to the data, is also considered.

Finally it must be emphasized that data emanating from the nonlinear regression model describing the network, and not from an unknown model function, are discussed in this study. At the same time the nonparametric nature of hidden-layer feed-forward neural networks is captured in the choice of model and of optimal design criteria.

3. Classical design approach. Suppose that designs for the nonlinear model specified by (2.2) which are G-optimal, i.e. designs for which the maximum value of the asymptotic variance of a prediction, $d(x, \mu, \theta)$, over all $x \in \mathcal{X}$ is minimized, are sought. Clearly the optimality criterion so invoked depends upon the unknown parameters of the model and it is therefore sensible, following the classical approach of Chernoff (1953), to adopt a best guess, say θ_o, for the parameter values and to construct "locally" G-optimal designs based on those values. The following Equivalence Theorem is an important tool in the construction of such designs.

THEOREM 3.1. *A design measure, $\mu^\star$, for a nonlinear model with p parameters can be equivalently characterized by any one of the three conditions:*

1. *$\mu^\star$ minimizes $\sup_{x \in \mathcal{X}} d(x, \mu, \theta_o)$*
2. *$\mu^\star$ maximizes $\mid M(\mu, \theta_o) \mid$*
3. *$\sup_{x \in \mathcal{X}} d(x, \mu^\star, \theta_o) = p$.*

Furthermore, the support of $\mu^\star$ is contained in the set of values of x for which $d(x, \mu^\star, \theta_o) = p$.

In particular, this theorem relates the precision of the predictions to those of the parameter estimates, and thus specifically G-optimality to D-optimality, and, in addition, provides a mechanism for confirming the global optimality or otherwise of a candidate design by means of an examination of the asymptotic prediction variances [see Silvey (1980), p. 54 and Atkinson and Donev (1992), Section 18.2].

It is assumed throughout the present study that data from the nonlinear regression model describing a particular neural network are readily available, and it thus follows that the maximum likelihood estimator of the parameters, written $\hat{\theta}$ and obtained by minimizing the error sum-of-squares function, $S(\theta) = \sum_{i=1}^{n} \{y_i - \eta(x_i, \theta)\}^2$, is a natural choice for the best guess, θ_o. This choice is not, however, an entirely straightforward and unambiguous one. Typically, the model function (2.2), or equivalently (2.1), is highly over-parametrized and the associated error sum-of-squares surface can, as a consequence, be fraught with local minima other than those emanating from the non-identifiability of the parameters [Ripley (1996), p. 159]. Thus there may well be a number of distinct parameter values corresponding to values of $S(\theta)$ close to the global minimum, $S(\hat{\theta})$, which can be considered as possible choices for θ_o. The following

example is now introduced to illustrate the ideas developed in this and the ensuing sections.

EXAMPLE 3.1. Data were generated from a nonlinear regression model having deterministic component (2.2) and normally distributed error terms. The parameter values were taken to be $\theta = (-1, 0.25, 1, -0.5, -1.35, 0.5, 0.75)$ and $\sigma = 0.1$, and 25 x-values, equally spaced between -12 and 12, were selected. For the error sum-of-squares function, $S(\theta)$, local and global minima were identified by using a quasi-Newton nonlinear optimization routine with a range of initial starting values. In particular, the maximum likelihood estimate of the parameters, corresponding to the global minimum of $S(\theta)$, was found to be $\hat{\theta}$ = (-1.829, 0.678, 0.485, -0.731, 1.360, 0.321, 0.607) with $S(\hat{\theta}) = 0.08411$, and local minima were observed with $S(\theta)$ values of 0.09033 and of 0.08451. The parameter values corresponding to these latter minima were, however, extremely insensitive to changes in θ_2 and in θ_6 or θ_7 respectively, and, not surprisingly therefore, the associated information matrices were found to be severely ill-conditioned. The locally G- or, equivalently, D-optimal design for this example based on a best guess for the parameters of $\hat{\theta}$ was constructed numerically by maximizing the determinant $| M(\mu, \hat{\theta}) |$ and comprises 7 equally weighted support points at -12, -3.126, -1.805, -0.680, 0.413, 1.648, and 12. This design is intuitively appealing in that it places points at the extremes of the design space, $[-12, 12]$, thereby anchoring the corresponding responses, and otherwise concentrates experimental effort around the peak of the fitted curve where high precision for the predictions is expected to be difficult to attain. An attempt was also made to construct G-optimal designs at parameter values corresponding to the local minima of $S(\theta)$, but the associated information matrices were too close to singularity for this to be achieved.

An interesting variant on the above approach to optimal design is to proceed sequentially and, in particular, to choose a set of design points so as to suitably augment the existing data. Suppose that the available data emanate from a design specified by the design measure, μ_o, and comprising n points of support, and suppose further that a single design point is to be added to these data. Then it would seem sensible, following Silvey (1980, p. 63), to choose this additional point to maximize the determinant of the information matrix of the augmented design evaluated at $\hat{\theta}$, i.e. to maximize

$$| nM(\mu_o, \hat{\theta}) + M(x, \hat{\theta}) | . \tag{3.1}$$

In fact, as is well-known, this single design point corresponds to the value of x for which the asymptotic variance of a prediction, $d(x, \mu_o, \hat{\theta})$, is a maximum, and can thus be regarded, in some sense, as G-optimal. It is interesting to note here that a similar design strategy for hidden-layer feed-forward neural networks was developed, essentially from first principles, by MacKay (1992a) and by Cohn (1996), and that their results are in accord with those of Dykstra (1971) and Ford and Silvey (1980).

In the present study this sequential procedure was also extended to include augmented designs for which the added design point is represented by a measure, μ_{add}, on the design space, $\mathcal{X}$, and for which the information matrix is therefore given by

$$nM(\mu_o, \hat{\theta}) + M(\mu_{add}, \hat{\theta}).$$

Specifically, it is natural to consider designs for which the determinant of this matrix is a maximum and to further observe that there is an Equivalence Theorem pertaining to such designs which can be gleaned from the Bayesian design literature [Pukelsheim (1993), Sections 11.5 to 11.8], and which is analogous to Theorem 3.1. This theorem is particularly valuable firstly in that it relates designs for which $\mid nM(\mu_o, \hat{\theta}) + M(\mu_{add}, \hat{\theta}) \mid$ is a maximum to those for which the maximum of the associated directional derivative,

$$
\begin{aligned}
(3.2) \qquad & tr\{nM(\mu_o, \hat{\theta}) + M(\mu_{add}, \hat{\theta})\}^{-1}\{M(x, \hat{\theta}) - M(\mu_{add}, \hat{\theta})\} \\
& \qquad = tr\{nM(\mu_o, \hat{\theta}) + M(\mu_{add}, \hat{\theta})\}^{-1}\{nM(\mu_o, \hat{\theta}) + M(x, \hat{\theta})\} - 7,
\end{aligned}
$$

taken over all $x \in \mathcal{X}$, is a minimum, and secondly in that it provides a tool for confirming the global optimality or otherwise of a candidate design. It is interesting to note that the derivative (3.2) can be re-expressed as an appropriately weighted sum of asymptotic variances for predictions at the support points of the existing design, μ_o, and at the point, $x \in \mathcal{X}$, and hence that the associated minimax designs can be regarded, in a certain sense, as G-optimal.

For Example 3.1, the original design, μ_o, comprises 25 x-values equally-spaced between -12 and 12. The single design point which maximizes the determinant (3.1) was obtained by maximizing the asymptotic prediction variance, $d(x, \mu_o, \hat{\theta})$, over the region, $[-12, 12]$, and is located at $x = -0.677$, while the corresponding optimal augmented design measure maximizing (3.2) was found to comprise the three support points, $-1.818, -0.675$ and 0.443, with attendant weights, $0.2878, 0.3726$ and 0.3396 respectively. For both of these sequential designs, the added points correspond to predicted values close to the peak of the fitted curve.

4. Bayesian designs. The Bayesian paradigm for hidden-layer feed-forward neural networks is proving to be both an appealing and a powerful one [Bishop (1995), Chapter 10, Ripley (1996) Section 5.5 and Neal (1996)]. Furthermore it offers important advantages within the present context of optimal design firstly in that the dependency of design criteria on the unknown parameters can be countered by placing a prior distribution on those parameters [Chaloner and Larntz (1989)] and secondly in that the notions of sequential design are accommodated in a very natural way [Tsutakawa (1972), Zacks (1977) and Chaloner (1989)].

For neural networks described by the model function (2.2) it is interesting to consider Bayesian D-optimal designs, i.e. designs which maximize the criterion,

$$
(4.1) \qquad \int_{\theta \in \Theta} \ln \mid M(\mu, \theta) \mid q(\theta) d\theta,
$$

where $q(\theta)$ represents the probability distribution function for the parameters, θ, over some parameter space, Θ. It then follows immediately from the Equivalence Theorem formulated by Chaloner and Larntz (1989) as a Bayesian analogue to Theorem 3.1 that such designs also minimize the criterion,

$$
\max_{x \in \mathcal{X}} \int_{\theta \in \Theta} d(x, \mu, \theta)\ q(\theta) d(\theta).
$$

It is tempting to regard this latter criterion as Bayesian G-optimality, but with the reservation expressed by Chaloner and Verdinelli (1995) that it is does not emanate from a utility function and is therefore not strictly Bayesian.

It is assumed in this study that observations, y_i, corresponding to the design points, $x_i, i = 1, \ldots, n$, are available and are normally distributed, and it thus follows that a natural choice for the distribution of the parameters, θ and τ, is the posterior distribution based on a noninformative prior. In particular, it is sensible to consider a prior distribution in which θ is uniformly distributed on a region of parameter space, $\mathcal{D}$, with $\theta_1 < \theta_3$, $\theta_6 > 0$, and $\theta_7 > 0$, thereby ensuring identifiability of the parameters, and in which the precision, $\tau = 1/\sigma^2$, follows a gamma distribution with parameters α and β, independently of θ. This choice of prior contrasts with that of normality for the parameters, θ, usually adopted within the neural network setting [Bishop (1995) Section 10.1 and Neal (1996)], but would, in justification, seem to be a particularly pragmatic one. In summary therefore, the joint posterior distribution for θ and τ is assumed to have probability density function,

$$\begin{aligned} h(\theta, \tau) &\propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum_{i=1}^{n}\{y_i - \eta(x_i,\theta)\}^2} \tau^{\alpha-1} e^{-\beta\tau} && \text{for } \theta \in \mathcal{D} \text{ and } \tau > 0, \\ &= 0 && \text{elsewhere,} \end{aligned}$$

and the required distribution, $q(\theta)$, is thus the marginal specified by $\int_0^\infty h(\theta, \tau) d\tau$.

The Bayesian criterion, (4.1), is a multi-dimensional integral over a restricted parameter space, $\mathcal{D}$, and as such is extremely awkward to compute and, more particularly, to maximize with respect to the design measure, μ. An approximate solution to this problem can be obtained by invoking Monte Carlo maximization, a technique pioneered by Geyer and Thompson (1992) and Chen, Geisser, and Geyer (1993), and used independently in the context of optimal design by Atkinson, Demetrio and Zocchi (1995) and more recently by Atkinson and Bogacka (1997). In the present case, the approach is implemented broadly as follows.

1. Generate a sample, S_Q, from the distribution specified by $q(\theta)$ using Markov chain Monte Carlo techniques.

2. Maximize the sum, $\sum_{\theta \in S_Q} \ln \mid M(\mu, \theta) \mid$, which approximates the required integral, with respect to the design measure, μ, and confirm the global optimality or otherwise of the resultant design, $\mu^\star$, from the maxima of the sum of asymptotic prediction variances, $\sum_{\theta \in S_Q} d(x, \mu^\star, \theta)$.

For Example 3.1, the parameter space, $\mathcal{D}$, was chosen, somewhat conservatively, to comprise θ values satisfying the constraints, $-6 < \theta_1 < \theta_3 < 6$, $0 < \theta_2 < 3$, $-2 < \theta_4 < 0$, $1.3 < \theta_5 < 1.44$, $0 < \theta_6 < 2.3$ and $0 < \theta_7 < 2.6$, and the parameters of the gamma prior for τ were taken to be $\alpha = \beta = 0.0001$. A chain of 510,000 pairs of θ and τ values from the joint distribution specified by $h(\theta, \tau)$ was generated by sampling alternately from the full conditional distributions,

$$\tau \mid \theta \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{S(\theta)}{2}\right)$$

and

$$h(\theta \mid \tau) \propto e^{-\tau S(\theta)/2} \quad \text{for} \quad \theta \in \mathcal{D}.$$

Sampling from the latter distribution was achieved by invoking the random walk Metropolis algorithm of Bennett, Racine-Poon and Wakefield (1996). Specifically, a multivariate normal proposal for the parameter, θ, centered on the current value and having a variance matrix equal to a constant, c, times the inverse of the information matrix evaluated at $\hat{\theta}$ and the current value of τ, was adopted, and the value of c was chosen so as to ensure an acceptance rate of candidate θ values close to 30%. It should be noted that a long chain of (θ, τ) values was generated here in an attempt to achieve a thorough sampling of the appropriate parameter space. The required sample, S_Q, was obtained from this chain by discarding the first 10,000 values in order to accommodate "burn-in", and by then subsampling every 500th value of θ in the chain, and thus comprises 1000 near independent realizations from the distribution, $q(\theta)$. The frequency distributions obtained from S_Q for each of the parameters, $\theta_1, \ldots, \theta_7$, indicate that the corresponding marginal distributions, with the exception of that for θ_5, are severely skewed and non-normal.

The design maximizing the criterion, $\sum_{\theta \in S_Q} \ln \mid M(\mu, \theta) \mid$, was found to comprise the 17 support points, -12.0, -4.532, -3.937, -3.586, -3.245, -2.993, -2.762, -2.417, -2.103, -1.833, -1.405, -0.637, 0.389, 0.900, 1.461, 3.091, and 12.0 with corresponding weights, 0.122, 0.011, 0.016, 0.028, 0.047, 0.023, 0.043, 0.045, 0.026, 0.044, 0.057, 0.101, 0.120, 0.029, 0.122, 0.050, and 0.116 respectively, and its global optimality was confirmed numerically from the values of the appropriate sum of asymptotic prediction variances. It is interesting to observe that this approximate Bayesian D-optimal design concentrates points at x-values corresponding to predictions in the vicinity of the peak of the fitted curve, and also that, in accordance with other findings [Chaloner and Larntz (1989)], it comprises many more design points than its classical D-optimal counterpart.

A possibly more satisfactory approach to optimal Bayesian design than that described above is to draw the existing data more directly into design construction through the information matrix, $M(\mu_o, \theta)$, and, following Chaloner (1989), to proceed sequentially by maximizing the criterion,

$$\int_{\theta \in \Theta} \ln \mid nM(\mu_o, \theta) + M(\mu_{add}, \theta) \mid q(\theta) d(\theta),$$

with respect to the augmented design, μ_{add}. A strategy analogous to that just described for Bayesian D-optimality was adopted and, in particular, designs maximizing the approximate criterion,

$$\sum_{\theta \in S_Q} \ln \mid nM(\mu_o, \theta) + M(\mu_{add}, \theta) \mid \tag{4.2}$$

were found, and their global optimality confirmed from the values of the corresponding directional derivative,

$$\sum_{\theta \in S_Q} tr\{nM(\mu_o, \theta) + M(\mu_{add}, \theta)\}^{-1}\{M(x, \theta) - M(\mu_{add}, \theta)\}.$$

Thus for Example 3.1, the design maximizing the sum, (4.2), over the sample, S_Q, of 1000 θ values was found to be based on the 6 support points, -3.370, -2.540, -1.588, -0.655, 0.495, and 1.386 with attendant weights, 0.122, 0.385, 0.113, 0.007, 0.367, and 0.006 respectively. It is interesting to observe that this design is again more diffuse than its locally optimal counterpart.

5. Conclusions. The main thrust of this paper has been to develop a framework for constructing designs for nonlinear models describing hidden-layer feed-forward neural networks by invoking classical and Bayesian optimality criteria. A number of points of specific interest emerge immediately from the study. In particular, certain of the criteria involving augmented designs depend upon the number of observations in the original data set, and it would thus be worthwhile examining how the form of the associated optimal designs changes as this number of observations changes. In addition, it is common, within the classical framework for hidden-layer feed-forward neural networks, to incorporate regularization techniques into the modelling process in order to smooth the fitted curve [Bishop (1995), Section 9.2], and it would thus be of some interest to construct optimal designs to accommodate this methodology. The issue of overriding concern in the context of optimal Bayesian design is that of maximizing criteria which are multi-dimensional integrals. The approach of the present study relies upon an efficient and effective Markov chain Monte Carlo procedure for generating samples from the posterior distribution of the parameters, and it is not clear that the random-walk Metropolis algorithm used is satisfactory in that respect. It would thus be of interest to consider, for example, simulating Markov chains by using mixture proposal distributions of the type described by Tierney (1995), or implementing the dynamically-based hybrid Monte Carlo algorithm invoked recently within the neural network setting by Neal (1996, Chapter 3). More generally, alternative methods for maximizing integrals to that used in the present study are available, and the procedure of Müller and Parmigiani (1995), which involves maximizing a smoothed estimate of the integral, would seem to be a particularly relevant and promising one.

Finally, and in a broader context, it is clear that the use of a sequential approach to optimal design for nonlinear models in situations in which data are available is practically very appealing. Results relating to this are however difficult to derive, and in fact there are few studies, both in the classical and in the Bayesian setting, upon which the practitioner can draw [Ford, Titterington and Kitsos (1989)]. Also, it should be reiterated that hidden-layer feed-forward neural networks model data in an essentially nonparametric manner, and that, strictly, optimal designs which accommodate this implied model misspecification should be constructed [Chang and Notz (1996)]. In summary therefore the present study highlights the need for further research into two broad, and to some extent interwoven, areas, those of sequential design and of optimal design for nonparametric regression.

Acknowledgments. The author wishes to thank the University of Natal and the Foundation for Research Development, South Africa, for funding this work and the referees for helpful and insightful comments.

REFERENCES

ATKINSON, A. C. AND BOGACKA, B. (1997). Compound, D- and D_s-optimum designs for determining the order of a chemical reaction. *Technometrics* **39** 347–356.

ATKINSON, A.C., DEMETRIO, C. G. B. AND ZOCCHI, S. S. (1995). Optimum dose levels when males and females differ in response. *Appl. Statist.* **44** 213–226.

ATKINSON, A. C. AND DONEV, A. N. (1992). *Optimum Experimental Designs.* Clarendon Press, Oxford.

BAUM, E. (1991). Neural network algorithms that learn in polynomial time from examples and queries. *IEEE Trans. Neural Networks* **2** 5–19.

BENNETT, J. E., RACINE-POON, A. AND WAKEFIELD, J. C. (1996). MCMC for nonlinear hierarchical models. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.), 339-357. Chapman and Hall, London.

BISHOP, C. M. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

BRITTAIN, S. AND HAINES, L. M. (1997). Nonlinear models for neural networks. In *Mathematics of Neural Networks : Models, Algorithms and Applications* (S. W. Ellacott, J. C. Mason and I. J. Anderson, eds.), 129-133. Kluver, Boston.

CHALONER, K. (1989). Bayesian design for estimating the turning point of a quadratic regression. *Commun. Statist. - Theor. Meth.* **18** 1385–1400.

CHALONER, K. AND LARNTZ, K. (1989). Optimal Bayesian designs applied to logistic regression experiments. *J. Statist. Plann. Inference* **21** 191–208.

CHALONER, K. AND VERDINELLI, I. (1995). Bayesian experimental design: a review. *Statist. Sci.* **10** 273–304.

CHANG, Y-J. AND NOTZ, W. I. (1996). Model robust designs. In *Handbook of Statistics, Volume 13* (S. Ghosh and C. R. Rao, eds.), 1055–1098. Elsevier, Amsterdam.

CHEN, L.-S., GEISSER, S. AND GEYER, C. J. (1993). Monte Carlo minimization for sequential control. Technical Report 591, School of Statistics, Univ. Minnesota.

CHERNOFF, H. (1953). Locally optimal designs for estimating parameters. *Ann. Math. Statist.* **24** 586–602.

COHN, D. A. (1996). Neural network exploration using optimal experiment design. *Neural Networks* **9** 1071–1083.

DYKSTRA, O. (1971). The augmentation of experimental data to maximize $|X'X|$. *Technometrics* **13** 682–688.

FORD, I. AND SILVEY, S. D. (1980). A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika* **67** 381–388.

FORD, I., TITTERINGTON D. M. AND KITSOS, C. P. (1989). Recent advances in nonlinear experimental design. *Technometrics* **31** 49–60.

GEYER, C. J. AND THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. B* **54** 657–699.

HWANG, J.-N., CHOI, J.J., OH, S. AND MARKS, R.J. (1991). Query-based learning applied to partially trained multilayer perceptrons. *IEEE Trans. Neural Networks* **2** 131–136.

MACKAY, D. J. C. (1992a). Information-based objective functions for active data selection. *Neural Computation* **4** 590–604.

MACKAY, D. J. C. (1992b). The evidence framework applied to classification networks. *Neural Computation* **4** 720–736.

MÜLLER, P. AND PARMIGIANI, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *J. Amer. Statist. Assoc.* **90** 1322–1330.

NEAL, R. M. (1996). *Bayesian Learning for Neural Networks.* Springer-Verlag, New York.

PLUTOWSKI, M. AND WHITE, H. (1993). Selecting concise training sets from clean data. *IEEE Trans. Neural Networks* **4** 305–318.

PUKELSHEIM, F. (1993). *Optimal Design of Experiments.* Wiley, New York.

RIPLEY, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge Univ. Press, Cambridge.

SILVEY, S. D. (1980). *Optimal Design.* Chapman and Hall, London.

SOLLICH, P. (1994). Query construction, entropy, and generalization in neural-network models. *Physical Review E* **49** 4637–4651.

TIERNEY, L. (1995). A note on Metropolis-Hastings kernels for general state spaces. Technical Report 606, School of Statistics, Univ. Minnesota.

TSUTAKAWA, R. K. (1972). Design of experiment for bioassay. *J. Amer. Statist. Assoc.* **67** 584–590.

WILLIAMS, C.K.I., QAZAZ, C., BISHOP, C.M. AND ZHU, H. (1995). On the relationship between Bayesian error bars and the input data density. Technical Report NCRG/95/024, Neural Computing Research Group, Aston Univ.

ZACKS, S. (1977). Problems and approaches in design of experiments for estimation and testing in non-linear models. In *Multivariate Analysis 4* (P. R. Krishnaiah, eds.), 209–223. North-Holland, Amsterdam.

DEPARTMENT OF STATISTICS AND BIOMETRY
UNIVERSITY OF NATAL PIETERMARITZBURG
PRIVATE BAG X01
SCOTTSVILLE 3209
SOUTH AFRICA