Bernoulli **4**(1), 1998, 3–14

Optimal design for curve estimation by local linear smoothing

MING-YEN CHENG,[1,2] PETER HALL[1*] and D. MICHAEL TITTERINGTON[1,3]

[1]*Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*
[2]*Institute of Mathematical Statistics, National Chung Cheng University, Minghsiung, Chiayi, Taiwan Republic Of China*
[3]*Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK*

The integral of the mean squared error of an estimator of a regression function is used as a criterion for defining an optimal design measure in the context of local linear regression, when the bandwidth is chosen in a locally optimal manner. An algorithm is proposed that constructs a sequence of piecewise uniform designs with the help of current estimates of the integral of the mean squared error. These estimates do not require direct estimation of the second derivative of the regression function. Asymptotic properties of the algorithm are established and numerical results illustrate the gains that can be made, relative to a uniform design, by using the optimal design or sub-optimal, piecewise uniform designs. The behaviour of the algorithm in practice is also illustrated.

Keywords: bandwidth choice; local linear regression; mean squared error; nonlinear regression; optimal design; sequential design

1350–7265

1. Introduction

The problem of optimal design for linear regression models began with seminal work by Kiefer (1959), leading to research summarized in books by Fedorov (1972), Silvey (1980) and Pukelsheim (1993). In contrast to the linear case, in nonlinear problems the true values of parameters can strongly influence optimal designs; see, for instance, Ford *et al*. (1989) and Chaloner and Verdinelli (1995) for reviews of, respectively, non-Bayesian and Bayesian approaches. Recently, optimal design ideas have been applied to particular nonlinear models in the neural computation literature, where the concept of optimally slanted sequential design is described as 'active learning'; see, for example, Fedorov (1972), MacKay (1992), Cohn (1994) and the statistical introduction by Cheng and Titterington (1994).

Cohn (1994) mentions the possibility of extending these ideas to environments such as locally weighted regression. It is this direction that the present paper takes, by considering

*To whom correspondence should be addressed. e-mail: halpstat@durras.anu.edu.au

1350–7265

the application of optimal design ideas to so-called 'nonparametric' regression. We shall modify the usual criterion on which optimality is based, since in nonparametric regression it is central that the 'model' represented by the fitted curve is incorrect. Indeed, optimality is achieved by trading off model inaccuracy, expressed through bias, against model suitability, represented by purely stochastic error. As our optimality criterion we shall use the integral of the mean squared error over a compact design space, although, in principle, versions that reflect differential weighting across the design space could be employed instead.

Using this viewpoint of optimality, Section 2 will propose an empirical, asymptotically optimal sequential rule for selecting both the bandwidth and the design density when the estimator is based on local linear regression. For purposes of illustration we shall adopt as our goal the modification of uniform design obtained by putting less weight in regions of low curvature. Thus, the task becomes one of determining how to estimate the curve reasonably in places of high curvature, subject to a constant weight in the definition of mean squared error. Other options will be discussed in Section 2.3. Numerical properties of our procedure will be reported in Section 3, in terms of a simulation study. Section 4 will outline technical arguments behind the results in Section 2.

It is assumed that observations are generated by the model

$$Y = g(x) + \varepsilon, \tag{1.1}$$

where g is the function to be estimated, ε has mean zero and variance σ^2 (and a distribution not depending on x), and, conditional on design points $x = X_i$, the ordinates $Y = Y_i$ are independent with a distribution determined by the model at (1.1). The algorithm for selecting design points is at the disposal of the experimenter, and should be chosen to optimize performance. We shall suppose that the design is restricted to the interval $\mathscr{I} = [0, 1]$, although clearly other possibilities may be treated in a similar way.

2. An algorithm and its properties

2.1. Algorithm for computing design and bandwidth

The sequential rule that we propose is based on updating in a geometric sequence of steps. This represents a compromise between the fully sequential Anscombe-type algorithm, which involves adjusting the algorithm for datum-by-datum increments in sample size, n, and is not really appropriate in nonparametric regression; and the double sampling Stein-type approach, which 'guesses' the final order of magnitude of the desired sample size, and uses a single but sizeable subsample to refine the initial guess. A wide variety of approaches, involving, for example, polynomial increases in n instead of fully sequential methods, can be effective, depending on the 'cost' of each update. In the context to which we shall apply our techniques, cost would depend largely on computational complexity.

Given $r > 1$, let n_k denote the integer part of r^k. Estimation of g is conducted iteratively, with step k employing information derived from the previous n_{k-1} data pairs. Step k may be conveniently broken into two parts: (a) determining a design density $\hat{f}_k$ from which to draw the design points for the next $N_k = n_k - n_{k-1}$ pairs; and (b) drawing these new data,

adjoining them to the earlier n_{k-1} pairs to produce a new set $\mathscr{X}_k = \{(X_i, Y_i), 1 \leqslant i \leqslant n_k\}$, and using $\mathscr{X}_k$ to construct estimators $\hat{g}_k$ of g and $\hat{\sigma}_k^2$ of σ^2. We compute $\hat{f}_k$ as a histogram, define $\hat{g}_k$ using local linear smoothing, and construct $\hat{\sigma}_k$ using first-order differences.

Algorithm for completing part (a). (i) *Definition of histogram.* Given an integer $m_k \ll n_k$, and positive constants $a_1, \ldots, a_{m_k}$ satisfying $\sum a_i = m_k$, let $f = f(\cdot|a_1, \ldots, a_{m_k})$ denote the density on $\mathscr{I}$ that equals a_i on $((i-1)/m_k, i/m_k]$ for $1 \leqslant i \leqslant m_k$.

(ii) *Estimation of mean squared error.* Write $\hat{g}_{k-1}$ and $\hat{\sigma}_{k-1}^2$ for the estimators of g and σ^2 computed from the set $\mathscr{X}_{k-1}$ of the first n_{k-1} data pairs. Let $\nu_k \ll n_k$ be an integer, let K be the kernel that we shall employ to construct $\hat{g}_k$, and define $\kappa_1 = \int K^2$ and $h = h(\cdot|a_1, \ldots, a_{m_k}, b) = b\nu_k^{-1/5} f^{-2}$, where $b > 0$. (Thus, in contradistinction to near-neighbour methods which effectively take h inversely proportional to f, we ask that h be inversely proportional to the square of f.) Our estimator of mean integrated squared error is

$$\Delta(a_1, \ldots, a_{m_k}, b) = \kappa_1 \hat{\sigma}_{k-1}^2 \nu_k^{-4/5} b^{-1} + \Delta_1(a_1, \ldots, a_{m_k}, b),$$

where

$$\Delta_1(a_1, \ldots, a_{m_k}, b) = \int_{\mathscr{J}} \left(\int [\hat{g}_{k-1}\{x - h(x)y\} - \hat{g}_{k-1}(x)] K(y)\, \mathrm{d}y \right)^2 \mathrm{d}x,$$

$\mathscr{J} = (C\nu_k^{-1/5}, 1 - C\nu_k^{-1/5})$, and $C > 0$ is a constant such that the restriction $0 < h \leqslant C\nu_k^{-1/5}$ is imposed by constraints on the choice of $(a_1, \ldots, a_{m_k}, b)$.

(iii) *Definition of* $\hat{f}_k$. Estimate $a_1, \ldots, a_{m_k}, b$ as the values of those quantities that minimize Δ, subject to a restriction which implies that $0 < h \leqslant C\nu_k^{-1/5}$. (There are many examples of such restrictions. Simple ones will be considered in Section 2.2.) Note that minimizing Δ over $a_1, \ldots, a_{m_k}$, for fixed b, is equivalent to minimizing Δ_1 over $a_1, \ldots, a_{m_k}$. Let $\hat{f}_k$ be the version of f obtained on substituting the estimators for the values of $a_1, \ldots, a_{m_k}, b$. Write $\hat{b}_k$ for the estimator of b.

A key feature of our approach is that the term, Δ_1, that measures the contribution due to bias avoids the troublesome, direct estimation of the second derivative of the curve.

Algorithm for completing part (b). Conditional on $\mathscr{X}_{k-1}$, draw N_k independent and identically distributed random pairs $\mathscr{Y}_k = \{(X_{ki}, Y_{ki}), 1 \leqslant i \leqslant N_k\}$, generated by the model, with the X_{ki} having density $\hat{f}_k$. Compute the local linear estimator $\hat{g}_k$, based on the data $\mathscr{X}_k = \mathscr{X}_{k-1} \cup \mathscr{Y}_k$ and using locally adaptive bandwidth $h = \hat{b}_k n_k^{-1/5} \hat{f}_k^{-2}$ and kernel K. Define

$$\tilde{\sigma}_k^2 = \{2(N_k - 1)\}^{-1} \sum_{i=1}^{N_k - 1} (Y'_{k,i+1} - Y'_{ki})^2,$$

where $\{(X'_{ki}, Y'_{ki}), 1 \leqslant i \leqslant N_k\}$ denotes an ordering of the pairs in $\mathscr{Y}_k$ such that $X'_{k1} < \cdots < X'_{kN_k}$. Define $\hat{\sigma}_k^2$ by either $\hat{\sigma}_k^2 = \tilde{\sigma}_k^2$ or

$$\hat{\sigma}_k^2 = n_{k-1}(n_{k-1} + N_k)^{-1} \hat{\sigma}_{k-1}^2 + N_k(n_{k-1} + N_k)^{-1} \tilde{\sigma}_k^2.$$

2.2. Motivation for the algorithm

We shall motivate the algorithm in terms of its ability to minimize integrated squared error. This is often appropriate when the curve is to be used for calibration or prediction, for example. In other cases, where applications will be more interpretative, perhaps through analysis of unusual features or turning points, a different criterion would be employed. The difference in the criterion may be as simple as employing a weight function when defining mean squared error, perhaps with the weight chosen adaptively so as to select features of interest. Alternatively, a measure of risk allied to a geometric description of distance, such as the Haussdorff metric, might be employed instead of squared error loss. The broad approach to constructing the algorithm would be similar in such cases, with the aim still being to optimize an empirical measure of loss. But details will of course differ.

We shall describe motivation in a heuristic fashion, but it may be rigorously justified under the following conditions: (a) the target function g has two continuous derivatives on $\mathscr{I}$, and g'' vanishes only at a finite number of points; (b) the error distribution has all moments finite, mean zero and variance σ^2; and (c) the symmetric, non-negative kernel K is Hölder continuous and supported on a bounded interval, say $(-c, c)$. With these assumptions, Section 2.3 will state a theorem addressing the performance of the algorithm.

Suppose n independent observations are made of a pair (X, Y) generated as $Y = g(X) + \varepsilon$, in which the design variables X are distributed with a continuous density f, and the distribution of the error, ε, has mean zero and variance σ^2. An estimator of g based on local linear smoothing, using kernel K and bandwidth h, has its asymptotic mean squared error at $x \in \mathscr{I}$ given by

$$H_n(x,\, h|f) = (nh)^{-1}\kappa_1\sigma^2 f(x)^{-1} + \tfrac{1}{4}h^4\kappa_2 g''(x)^2, \tag{2.1}$$

where κ_1 is as in Section 2.1 and $\kappa_2 = \{\int y^2 K(y)\,\mathrm{d}y\}^2$; see, for example, Fan (1993) and Hastie and Loader (1993). For fixed x, which we now suppress, the quantity at (2.1) is minimized by taking $h = h_0 = (n\kappa_3 f g''^2)^{-1/5}$, where $\kappa_3 = \kappa_2/(\kappa_1\sigma^2)$. Substituting back into (2.1), we deduce that with an optimal local choice of bandwidth, mean squared error is given asymptotically by $n^{-4/5}\kappa_4(g''^2/f^4)^{1/5}$, where the constant κ_4 depends only on K and σ^2. The minimum mean integrated squared error is obtained by integrating this quantity over $\mathscr{I}$, producing a functional proportional to $A(f) = \int_{\mathscr{I}}(g''^2/f^4)^{1/5}$.

The optimal design density f is that function which minimizes $A(f)$ subject to $\int_{\mathscr{I}} f = 1$ and $f \geqslant 0$. A simple calculus of variations argument shows that this is given by $f_0 = c_0|g''|^{2/9}$, where the constant c_0 is chosen to ensure that $\int_{\mathscr{I}} f_0 = 1$. Note particularly that for this choice, the optimal bandwidth is inversely proportional to the square of f: $h_0 = c_1 f^{-2}$, where c_1 is a constant. This explains why, in the algorithm suggested in Section 2.1, we took the bandwidth for computing $\hat{g}_k$ to vary in proportion to $\hat{f}_k^{-2}$.

In (2.1), let the bandwidth h be $h_1 = bn^{-1/5}f^{-2}$, where b is an arbitrary positive constant and f is an arbitrary continuous density, bounded away from zero on $\mathscr{I}$. On integrating $H_n(\cdot,\, h_1|f)$ over $\mathscr{I}$ the contribution from the first term may be seen to equal

$$\int_{\mathscr{I}} (nh_1)^{-1}\kappa_1\sigma^2 f^{-1} = \kappa_1\sigma^2 n^{-4/5}b^{-1}.$$

Note particularly that the effect of f has disappeared. This explains the origin of the first term in the formula for Δ, and why it depends only on b, not on $a_1, \ldots, a_{m_k}$. The second term is an estimate of the integral of the second term of (2.1), again with h_1 substituted for h.

If the function g is at all 'interesting' then it will enjoy one or more points of inflection, where $g'' = 0$. There, the optimal design density determined by the argument two paragraphs above equals zero, and in such cases the approximate formula at (2.1) is not valid. We overcome this problem by introducing a threshold, η, below which the value of $\hat{f}_k$ is not allowed to fall. When discussing theoretical properties of our algorithm it is convenient to also impose a ceiling on f. For economy of notation we take this to be η^{-1}, although we could have used any large positive number. Given $\eta \in (0, 1)$, define

$$f_\eta = \begin{cases} c_\eta |g''|^{2/9} & \text{if } c_\eta |g''|^{2/9} \in (\eta, \eta^{-1}) \\ \eta & \text{if } c_\eta |g''|^{2/9} \leqslant \eta \\ \eta^{-1} & \text{if } c_\eta |g''|^{2/9} \geqslant \eta^{-1}, \end{cases} \tag{2.2}$$

where the positive constant c_η is uniquely defined by the requirement that $\int f_\eta = 1$, and where $f_\eta \to f_0$ and $c_\eta \to c_0$ as $\eta \to 0$. Let b_η denote the value of the constant b which minimizes

$$b^{-1}\kappa_1\sigma^2 + \tfrac{1}{4}b^4\kappa_2^2 \int_{\mathscr{I}} g''^2 f_\eta^{-8},$$

and let $\eta_1 \in (0, 1)$ be so small that $b_\eta \in (\eta_1, \eta_1^{-1})$.

If in part (a) of the algorithm we restrict attention to $(a_1, \ldots, a_{m_k}, b)$ such that

$$\begin{gathered} \eta \leqslant a_i \leqslant \eta^{-1} \qquad \text{for } 1 \leqslant i \leqslant m_k, \\ \eta_1 \leqslant b \leqslant \eta_1^{-1}, \end{gathered} \tag{2.3}$$

where $\eta \in (0, 1)$ and η_1 is as defined in the previous paragraph, then the histogram $\hat{f}_k$ is an estimator of f_η, $\hat{b}_k$ is an estimator of b_η, and the constant C in the definition of $\mathscr{J}$ may be taken equal to $c\eta^{-2}\eta_1^{-1}$. The consistency of $\hat{f}_k$ and $\hat{b}_k$ for f_η and b_η, respectively, will be shown during the proof of Theorem 2.1.

In problems involving high orders of estimation the optimal design density is broadly similar to that at (2.2). For example, if we are locally fitting a polynomial of degree $2\nu - 3$, where $\nu \geqslant 2$ is an integer, then the asymptotically optimal design density is proportional to $|g^{(\nu)}|^{2/(4\nu+1)}$. This generalizes the case $\nu = 2$ treated just above. (The case of fitting polynomials of even order is a little different; note for example the results of Ruppert and Wand 1994.)

2.3. Properties of the algorithm

Let (C) denote conditions (a)–(c) introduced in the second paragraph of Section 2.2, as well as condition (2.3) for sufficiently small η_1, and the assumption that for some $\delta \in (0, 1)$ we have $\nu_k = O(n_k^{1-\delta})$, $m_k = O(\nu_k^{1-\delta})$ and $m_k \to \infty$. Let f_η have the definition given in Section 2.2, with η as in (2.3). The minimum mean squared error derived by optimizing over all design densities, subject to the constraint at (2.3), is $H_n(x, h|f_\eta)$. Our main theorem states that $\hat{g}_k$ achieves this optimum.

Theorem 2.1. *Assume conditions* (C)*, with* η_1 *in* (2.3) *chosen as suggested in Section 2.2. Then*

$$\left\{\int_{\mathscr{I}} (\hat{g}_k - g)^2\right\} \Big/ \left\{\int_{\mathscr{I}} \inf_h H_n(\cdot,\, h|f_\eta)\right\} \to 1 \tag{2.4}$$

with probability 1 as $k \to \infty$.

Remark 2.1 *Use of ridge methods*. The local linear estimator $\hat{g}_k$ is defined without recourse to adjustments, such as ridging, which are designed to induce more stable numerical properties. For a discussion of stabilization methods in non-sequential contexts, see, for example, Seifert and Gasser (1996). There is no difficulty in developing an analogue of Theorem 2.1 for the case where $\hat{g}_k$ is defined by ridged or interpolated local linear smoothing.

Remark 2.2 *Efficiency.* The efficiency of employing an optimal design density may be expressed as the ratio of two mean integrated squared errors, the first using the optimal density and the second using another density, f, say. In view of Theorem 2.1, efficiency may also be expressed in purely empirical terms. Indeed, the ratio of $\int_{\mathscr{I}}(\hat{g}_k - g)^2$ to $\int_{\mathscr{I}} \inf_{f,h} \mathrm{E}_f(\hat{g} - g)^2$, where $\hat{g}$ is a standard local linear estimator using kernel K (with ridging employed to ensure that mean squared error is well defined), and where the infimum is taken over all choices of f and of a non-stochastic, locally adaptive bandwidth h, converges (as first $k \to \infty$ and then $\eta \to 0$) to 1. In this sense, the sequential estimator $\hat{g}_k$ is fully efficient.

Remark 2.3 *Alternative regression estimators*. It is of interest to consider alternative approaches to regression, not least because they can produce results of a different character from those described earlier. We shall treat two, the classical Nadaraya–Watson method (see, for example Härdle 1990, Section 3.1; and Wand and Jones 1995, p. 119) and a 'transformation approach' (Hart 1991). By judicious choice of the design density, depending on the target function g, both these techniques permit the asymptotic bias of a second-order kernel estimator $\hat{g}$ to vanish, and hence the mean squared error to be an order of magnitude smaller than in the case of local linear smoothing. This result may seem to contradict the known minimax optimality of local linear smoothing (Fan 1993), but it should be noted that such optimality results pertain only to the case of a fixed design density that is bounded away from zero and infinity, and hence do not apply to sequentially chosen design. While reduced mean squared error is an attractive feature, it is achieved only at the expense of significant practical difficulties in constructing the sequential version of the estimator. Therefore, we do not develop the methods here beyond outlining theoretical properties and discussing their implications.

For a Nadaraya–Watson kernel estimator the asymptotic mean squared error formula, the analogue of (2.1), may be written as

$$(nh)^{-1}\kappa_1\sigma^2 f(x)^{-1} + \kappa h^4\{g''(x) + 2g'(x)f'(x)f(x)^{-1}\}^2, \tag{2.5}$$

where the constant κ depends only on the kernel. The second term here represents the squared bias contribution to the mean squared error, and may be rendered equal to zero (except at zeros of g') by defining $f = f_0 = c_0|g'|^{-1/2}$, where $c_0^{-1} = \int |g'|^{-1/2}$. Hart's transformation estimator has a mean squared error which enjoys a similar expansion, this time with the second term in (2.5) replaced by

$$\kappa h^4 \{g''(x) f(x)^{-2} - g'(x) f'(x) f(x)^{-3}\}^2.$$

This quantity is rendered equal to zero by using the design density $f_0 = c_0|g'|$, with $c_0^{-1} = \int |g'|$.

For both the Nadaraya–Watson and transformation definitions of $\hat{g}$ we may estimate $|g'|$ either explicitly or implicitly, and employ the estimator in the obvious way to construct an estimator of f_0. In principle, if g has four bounded derivatives then this technique can produce an order of squared bias equal to h^8, and hence an order of mean squared error equal to $n^{-8/9}$ (rather than the $n^{-4/5}$ typically associated with second-order methods), away from points where $g' = 0$. In practice, however, there are significant obstacles to achieving such performance. The first problem is that these methods demand an estimator of g' whose high-order derivatives accurately estimate those of g'. While this is feasible for large samples, it is not really practicable with smaller data sets. Since a sequential procedure would usually start with relatively small samples, this is a drawback. The inherent numerical instability of estimators of ratios of functions, such as appear in formulae for bias terms of Nadaraya–Watson or transformation estimators, makes it even more difficult to ensure good performance in small samples.

Secondly, attaining optimal performance at the level $n^{-8/9}$ demands a particularly complex bandwidth formula, depending on the fourth derivative of g. The theoretical difficulties are easily overcome, but an attractive numerical algorithm seems to be out of reach. In particular, there does not appear to exist a high-order analogue of the simple algorithm suggested in Section 2.1, which simultaneously produces an estimator of the optimal bandwidth and optimal design density. Therefore, selection of the appropriate bandwidth in a high-order sequential procedure is a particularly awkward problem. Thirdly, this high-order performance seems to be achievable only away from points where g' vanishes; at those points the optimal design density is either infinite (in the case of the Nadaraya–Watson estimator) or zero (for the transformation estimator). Different bandwidth selection procedures seem to be necessary at those points, producing different convergence rates. This makes for a cumbersome approach to inference.

3. Numerical results

3.1. Efficiencies of some suboptimal designs

Recall, from Section 2.2, that the mean integrated squared error associated with a design f and an optimally chosen bandwidth, is proportional to $A(f) = \int_{\mathcal{I}} (g''^2/f^4)^{1/5}$, and that the optimal design is given by $f_0 = c_0|g''|^{2/9}$. Thus, $A(f_0) = (\int_{\mathcal{I}} |g''|^{2/9})^{9/5}$ is the minimum achievable value of $A(f)$, and it is natural to define the *efficiency* of design f, relative to the

optimal design, by $A(f_0)/A(f)$. We evaluated this in the context of 15 true curves given by the class of Gaussian mixtures proposed by Marron and Wand (1992). These vary from the standard Gaussian density function through a range of Gaussian mixtures of varying complexity. The design space is the interval $(-3, 3)$. We describe our results here only in words. A more detailed account, including tabular results, is available in a technical report (Cheng *et al*. 1995).

We computed Monte Carlo approximations to efficiencies of designs that are themselves mixtures of the optimal design, f_0, and the uniform design on $(-3, 3)$, with mixing weights θ and $1 - \theta$, respectively. The range of θ was between 0 and 1 in steps of 0.1. For many of the 15 curves, there was not much to be gained from using the optimal design rather than the uniform, but in some cases the gains were considerable. Particular instances were curves 3–5 and 10–14, in the nomenclature of Marron and Wand (1992).

Bearing the above robustness in mind, as well as the fact that our algorithm for sequential design involves a sequence of piecewise uniform designs, we computed the efficiencies of such designs. We addressed cases corresponding to $m = 2^r$, for $r = 0, 1, \ldots, 4$, and with the a_i chosen at their optimal values. If such a design is denoted by f_{m0} then it is straightforward to show that $A(f_{m0}) = (L/m)^{4/5}(\sum A_i^{5/9})^{9/5}$, in which L denotes the length of the design space (an interval), and $A_i = \int |g''|^{2/5}$, where the range of integration is the interval $((i-1)L/m_k, iL/m_k]$ for $1 \leqslant i \leqslant m_k$. We obtained good gains in efficiency by $m = 8$. For example, in the case of curves 3–5, efficiencies were respectively 69%, 74% and 52% for $m = 1$, but had increased to 93%, 90% and 82% by $m = 8$.

Since zeros of the optimal design density correspond to points of inflexion of g (see (2.2)), the departure of optimal design from uniformity will tend to be in proportion to the 'wiggliness' of g. The numerical analyses described above confirm this informal theoretical conclusion. For example, functions 10–15 in Marron and Wand (1992) exhibit the greatest number of points of inflexion, and include a class of functions for which we observed substantial gains in performance when attempting to optimize design.

3.2. An illustration of the algorithm

In the small study reported here we concentrated on the mutual similarity of the formula for the asymptotic integrated mean squared error and the corresponding estimator thereof, defined by Δ, and on how closely the design created after one iteration of the algorithm approximates the optimal design. Although our theory is asymptotic in k, in practice only a small number of iterations of the algorithm would be carried out, bearing in mind the relationships between successive sample sizes. The design procedure is best described as *batch-sequential*, with large batch sizes, and anything approaching genuinely sequential design, in which the design is updated one observation at a time, does not make sense in the context of nonparametric regression. To help make the conclusions clear, the exemplar curve chosen was the continuous, but not continuously differentiable, piecewise quadratic curve given by

$$g(x) = \begin{cases} x(1-2x)/4 & \text{if } 0 \leqslant x \leqslant 1/2 \\ 2(1-2x)(1-x) & \text{if } 1/2 \leqslant x \leqslant 1. \end{cases}$$

For this curve, therefore, the optimal design is piecewise uniform, corresponding to $m = 2$, and with $a_1 = 2/(1 + 64^{1/9}) = 0.773$ and $a_2 = 2 - a_1$. The Epanechnikov kernel was used in the local linear fitting. Thus $\kappa_1 = 3/5$ and $\kappa_2 = 1/25$, so that the optimal value of b was $(2^9 \cdot 15 \cdot \sigma^2)^{1/5}/(1 + 64^{1/9})^{6/5}$. Therefore, the experiment involved choosing an initial sample of size n_0, taking $m_1 = 2$ and investigating the fruits of the algorithm in terms of the resulting estimates of a_1 and a_2. A variety of values were chosen for n_0 and ν_1, and care was taken to choose the range of integration $\mathscr{I}$ suitably when calculating Δ_1. In fact Δ_1 included a small gap near the discontinuity in $g'(x)$ at $x = \frac{1}{2}$, where the calculations were unstable. The bandwidth used for calculating the local linear curve estimates $\hat{g}_0$ was the asymptotically optimal value, constant in x, based on the estimate of σ calculated from first-order differences and on an assumption of constant curvature, for g, of magnitude 4. (Note that the true curvature has magnitude 1 on the first half of the interval, and 8 thereafter.)

We report here on ten replicates of each of the cases $\sigma = 0.01$, 0.05 and $n_0 = 400$, 1000. Throughout, $\nu_1 = 200$. For each replicate, Δ was evaluated on a grid of values of (b, a_1), and the optimal combination of values was found, correct to two decimal places in each of the two variables. This would seem to be adequate from a practical point of view. The optimal values for b were 0.171 and 0.326, for $\sigma = 0.01$ and 0.05 respectively, and the algorithm achieved these values closely, especially for $\sigma = 0.01$. In this case, the optimal value of a_1 (0.773) was slightly overestimated, because of the smoothing involved in calculating the bias term Δ_1, but the errors were not great. The Δ surface always had a nice single minimum in (b, a_1), at least in the region investigated in the experiment. The values of Δ were typically within a few per cent of those from the asymptotic formula. Occasionally, the difference went into double figures, in percentage terms, but was fairly constant as (b, a_1) varied, so that, as reported above, the positions of the minima on the two surfaces were very similar.

4. Outline proof of Theorem 2.1

The proof is given only in barest outline. Details are given in Cheng *et al.* (1995).

The first step is to establish a relatively crude upper bound to $|\hat{g}_k - g|$,

$$\sup_{\mathscr{I}} |\hat{g}_k - g| = O(n_k^{-(2/5)+\eta}) \tag{4.1}$$

with probability 1, as $k \to \infty$, for all $\eta > 0$. As a prelude to establishing (4.1), define $\mathscr{N}(l) = \{1, \ldots, N_l\}$, let $\{X_{li}, i \in \mathscr{N}(l)\}$ denote the set of design points constructed in step l of the algorithm, write $Y_{ki} = g(X_{ki}) + \varepsilon_{ki}$ for the associated measurements of Y, and let $h = \hat{b}_k n_k^{-1/5} \hat{f}_k^{-2}$. Let $s_{kj}(x)$ denote the sum of $(x - X_{li})^j K\{(x - X_{li})/h\}$ over $i \in \mathscr{N}(l)$ and $1 \leqslant l \leqslant k$, put $w_{li}(x) = \{s_{k2}(x) - s_{k1}(x)(x - X_{li})\}K\{(x - X_{li})/h\}$ for $i \in \mathscr{N}(l)$, and define W_k to equal the sum of w_{li} over $i \in \mathscr{N}(l)$ and $1 \leqslant l \leqslant k$. In this notation,

$$\hat{g}_k = g + (A_k + B_k)W_k^{-1}, \tag{4.2}$$

where

$$A_k = \sum_{l=1}^{k} A(l), \qquad A(l) = \sum_{i \in \mathcal{N}(l)} w_{li} \varepsilon_{li},$$

$$B_k = \sum_{l=1}^{k} B(l), \qquad B(l) = \sum_{i \in \mathcal{N}(l)} w_{li} \{ g(X_{ki}) - g \}.$$

Defining $K_1 = K$, $K_2(y) = yK(y)$ and

$$V_{lj}(x) = \sum_{i \in \mathcal{N}(l)} K_j\{(x - X_{li})/h\} \varepsilon_{li},$$

we have

$$|A_k| \leqslant \sum_{l=1}^{k} |A(l)| \leqslant \sum_{l=1}^{k} \sum_{j=1}^{2} |s_{k,3-j}| h^{j-1} |V_{lj}|$$

$$\leqslant (s_{k2} + |s_{k1}| h) \sum_{l=1}^{k} (|V_{l1}| + |V_{l2}|).$$

Also, $|s_{kj}| \leqslant (ch)^j U_k$, where $c > 0$ is chosen so that $(-c, c)$ contains the support of K, and

$$U_k = \sum_{l=1}^{k} \sum_{i \in \mathcal{N}(l)} K\{(x - X_{li})/h\}.$$

Therefore,

$$|A_k| \leqslant C_1 h^2 U_k \sum_{l=1}^{k} (|V_{l1}| + |V_{l2}|),$$

where, here and below, C_1 and C_2 are positive constants. Similarly, an upper bound may be established for B_k. In consequence, it may be shown from (4.2) that

$$|\hat{g}_k - g| \leqslant C_2 W_k^{-1} \left\{ h^2 U_k \sum_{l=1}^{k} (|V_{l1}| + |V_{l2}|) + h^4 U_k^2 \right\}. \tag{4.3}$$

Computations based on large-deviation bounds for the variables V_{l1} and V_{l2} show that $\sup_{\mathcal{J}} |V_{lj}| = O_p(n_k^{(2/5)+\eta})$ for all $\eta > 0$. Hence, by (4.3),

$$\sup_{\mathcal{J}} |\hat{g}_k - g| = O\{ W_{k0}^{-1} (h_0^2 n_k^{(2/5)+\eta} U_{k0} + h_0^4 U_{k0}^2) \}, \tag{4.4}$$

with probability 1, where $U_{k0} = \sup_{\mathcal{J}} U_k$, $W_{k0} = \inf_{\mathcal{J}} W_k$ and $h_0 = \sup_{\mathcal{J}} h$. It may be shown that $U_{k0} = O(n_k^{4/5})$ and $W_{k0}^{-1} = O(n_k^{-6/5})$; and, by definition of h, that $h_0 = O(n_k^{-1/5})$, all results holding with probability 1. The desired result (4.1) follows from these bounds and (4.4).

The next step is to show that

$$\sup_{\mathscr{I}} |\hat{f}_k - f_\eta| \to 0, \qquad \hat{b}_k \to b_\eta, \tag{4.5}$$

with probability 1. First, using (4.1), we may prove that

$$\Delta_1(a_1, \ldots, a_{m_k}, b) = \tfrac{1}{4}\nu_k^{-4/5} b^4 \kappa_2^2 \psi(f) + o(\nu_k^{-4/5}),$$

$$\Delta(a_1, \ldots, a_{m_k}, b) = \nu_k^{-4/5} \phi(f, b) + o(\nu_k^{-4/5}), \tag{4.6}$$

uniformly in $a_1, \ldots, a_{m_k}$, b satisfying (2.3), where $\psi(f) = \int_{\mathscr{I}} (g''/f^4)^2$ and $\phi(f, b) = \kappa_1 \sigma^2 b^{-1} + \frac{1}{4} b^4 \kappa_2^2 \psi(f)$. Results (4.5) may be derived from (4.6).

Using (4.5), it may be shown that

$$s_{kj} = n_k^{(4-j)/5} t^{(j+1)/5} \rho_j f_\eta + o(n_k^{(4-j)/5}) \tag{4.7}$$

uniformly on $\mathscr{I}$, with probability 1, where $\rho_j = \int y^j K(y)\, \mathrm{d}y$ and t is defined by $h = n_k^{-1/5} t$. Now, $\rho_0 = 1$, $\rho_1 = 0$, $\rho_2 = \kappa_2$ and $W_k(x) = s_{k2} s_{k0} - s_{k1}^2$, and so by (4.7),

$$\sup_{\mathscr{I}} |W_k (n_k^2 h_\eta^4 \kappa_2 f_\eta^2)^{-1} - 1| \to 0, \tag{4.8}$$

where $h_\eta = n_k^{-1/5} f_\eta^{-2} b_\eta$. Similarly, it may be shown that

$$\sup_{\mathscr{I}} |B_k (\tfrac{1}{2} n_k^2 h_\eta^6 \kappa_2^2 f_\eta^2)^{-1} - g''| \to 0. \tag{4.9}$$

Combining (4.8) and (4.9), we conclude that

$$\sup_{\mathscr{I}} |B_k W_k^{-1} - \tfrac{1}{2} h_\eta^2 \kappa_2 g''| = o(n_k^{-2/5}). \tag{4.10}$$

Combining (4.2), (4.8) and (4.10), we deduce that

$$\begin{aligned} \int_{\mathscr{I}} (\hat{g}_k - g)^2 &= (1 + \xi_1) \int_{\mathscr{I}} \{A_k (n_k^2 h_\eta^4 \kappa_2 f_\eta^2)^{-1} + (1 + \delta) \tfrac{1}{2} h_\eta^2 \kappa_2 g''\}^2 \\ &= (1 + \xi_2) \int_{\mathscr{I}} \{A_k^2 (n_k^2 h_\eta^4 \kappa_2 f_\eta^2)^{-2} + \tfrac{1}{4} (h_\eta^2 \kappa_2 g'')^2\} \\ &\quad + 2 \int_{\mathscr{I}} A_k (n_k^2 h_\eta^4 \kappa_2 f_\eta^2)^{-1} (\tfrac{1}{2} h_\eta^2 \kappa_2 g''), \end{aligned} \tag{4.11}$$

where the function δ satisfies $\sup_{\mathscr{I}} |\delta| \to 0$ with probability 1, and the random variables ξ_j converge to 0 with probability 1. After some simplification of the right-hand side of (4.11), we obtain (2.4).

Acknowledgement

The authors are grateful to David Cohn for access to unpublished material which stimulated

the work in this paper. The helpful comments of a referee and editor led to this shortened version of the original manuscript.

References

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: a review. *Preprint*.

Cheng, B. and Titterington, D.M. (1994) Neural networks: a review from a statistical perspective (with discussion). *Statist. Sci.*, **9**, 2–54.

Cheng, M.-Y., Hall, P. and Titterington, D.M. (1995) Optimal design for curve estimation by local linear smoothing. Research Report No. SRR046-95, Centre for Mathematics and its Applications, Australian National University.

Cohn, D.A. (1994) Neural network exploration using optimal experimental design. MIT AI Laboratory Memo No. 1491.

Fan, J. (1993) Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, **21**, 196–216.

Fedorov, V.V. (1972) *Theory of Optimal Experiments*. New York: Academic Press.

Ford, I., Kitsos, C.P. and Titterington, D.M. (1989) Recent advances in nonlinear experimental designs. *Technometrics*, **31**, 49–60.

Härdle, W. (1990) *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hart, J.D. (1991) Contribution to discussion of Chu and Marron (1991). *Statist. Sci.*, **6**, 425–527.

Hastie, T. and Loader, C. (1993) Local regression: automatic kernel carpentry. *Statist. Sci*., **8**, 120–143.

Kiefer, J. (1959) Optimal experimental designs (with discussion). *J. Roy. Statist. Soc. Ser. B*, **21**, 272–319.

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

Marron, J.S. and Wand, M. (1992) Exact mean integrated squared error. *Ann. Statist.*, **20**, 712–736.

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

Ruppert, D. and Wand, M.P. (1994) Multivariate locally weighted least squares regression. *Ann. Statist.*, **22**, 1346–1370.

Seifert, B. and Gasser, T. (1996) Finite sample analysis of local polynomials: analysis and solutions. *J. Amer. Statist. Assoc.*, **91**, 267–275.

Silvey, S.D. (1980) *Optimal Design*. London: Chapman & Hall.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman & Hall.

Received October 1995 and revised February 1997.