Statistica Sinica **5**(1995), 421-440

ON EFFICIENT DESIGNING OF NONLINEAR EXPERIMENTS

Probal Chaudhuri and Per A. Mykland

Indian Statistical Institute and University of Chicago

Abstract. Adaptive designs that optimize the Fisher information associated with a nonlinear experiment are considered. Asymptotic properties of the maximum likelihood estimate and related statistical inference based on dependent data generated by sequentially designed adaptive nonlinear experiments are explored. Conditions on the experimental designs that ensure first order efficiency of the maximum likelihood estimate when the parametrization of the nonlinear model is sufficiently smooth and regular are derived. A few interesting open questions that arise naturally in course of the investigation are mentioned and briefly discussed.

Key words and phrases: Adaptive sequential designs, first order efficiency, Fisher information, local Φ-optimality, martingales, maximum likelihood, nonlinear models.

1. Introduction: Adaptive Sequential Design of Nonlinear Experiments

In a typical nonlinear set up involving a response Y and a regressor X, the dependence of the response on the regressor is modeled using a family of probability distributions, which involve an unknown Euclidean parameter (to be estimated from the data) in such a way that the parametrization is smooth and regular, and the associated Fisher information happens to be a nonlinear function of that parameter of interest. Standard examples of such nonlinear models include the usual nonlinear regression models (see e.g. Gallant (1987), Bates and Watts (1988), Seber and Wild (1989)), various generalized linear models (see e.g. McCullagh and Nelder (1989)), and many popular heteroscedastic regression models (see e.g. Box and Hill (1974), Bickel (1978), Jobson and Fuller (1980), Carroll and Ruppert (1988)). An awkward situation arises when one needs to design an experiment in order to ensure maximum efficiency of parameter estimates in such a nonlinear problem. Since the Fisher information associated with any such problem depends on the unknown parameter in a nonlinear way, an efficient designing of the experiment to guarantee optimal performance of the nonlinear least squares or the maximum likelihood estimate will require knowledge of that unknown parameter! (See Cochran (1973, pp. 771-772), Bates and Watts (1988,

p. 129) for various interesting remarks in connection with this). An extensive discussion on possible resolutions of this problem together with several examples of nonlinear experiments can be found in Ford, Titterington and Kitsos (1989), Myers, Khuri and Carter (1989) and Chaudhuri and Mykland (1993). A few more interesting examples are discussed below.

Example 1.1. Treloar (1974) investigated an enzymatic reaction in which the number of counts per minute of a radioactive product from the reaction was measured, and from these counts the initial velocity (Y) of the reaction was calculated. This velocity is related to the substrate concentration (X) chosen by the experimenter through the nonlinear regression equation (Michaelis-Menten model) $Y = \beta_0 X(\beta_1 + X)^{-1} + e$. Here $\theta = (\beta_0, \beta_1)$ is the unknown parameter of interest, and e is the random noise with zero mean as usual. The experiment was conducted once with the enzyme (Galactosyltransferase of Golgi Membranes) treated with Puromycin and once with the enzyme untreated. The main objective of the study was to investigate the effect of the introduction of Puromycin on the ultimate velocity parameter β_0 and the half velocity parameter β_1.

Example 1.2. Cox and Snell (1989, pp. 10-11) while discussing binary response regression models mentioned about an industrial experiment, where the number of ingots that were not ready for rolling were counted out of a fixed number of ingots tested. The number counted depends on two covariates, which are heating time and soaking time determined by the experimenter at the time of processing. This leads to a binomial regression model and the main interest is in getting insights into how variations in heating and soaking times influence the number of ingots that turn out to be not ready for rolling.

Example 1.3. Powsner (1935) collected data from an experiment that was conducted to determine the effect of experimental temperature on the developmental stages (i.e. embryonic, egg-larval, larval and pupal stages) of the fruit fly Drosophila Melanogaster. McCullagh and Nelder (1989) discussed the data and demonstrated how it can be analyzed using appropriate gamma regression models with the experimental temperature as the covariate.

Chaudhuri and Mykland (1993) investigated an adaptive sequential scheme for designing nonlinear experiments and established asymptotic D-optimality of the design as well as $n^{1/2}$-consistency, asymptotic normality and first order efficiency of the maximum likelihood estimate based on observations generated by their proposed scheme in a nonlinear set up satisfying suitable regularity conditions. Adaptive sequential and batch sequential designs for certain very specific types of nonlinear experiments related to some special models were explored earlier by Box and Hunter (1965), Ford and Silvey (1980), Abdelbasit and Plackett (1983), Ford, Titterington and Wu (1985), Wu (1985), Minkin (1987) and McCormick, Mallik and Reeves (1988) among others. Some of these authors made attempts to study the asymptotic behavior of their respective design schemes and resulting parameter estimates. The key idea lying at the root of such design strategies is to divide available resources into several groups and to split the entire experiment into a number of steps. At each step, the experiment is conducted using only a single portion of divided resources. At the end of each step, an analysis is carried out and parameter estimates are updated using the available data. The results emerging from this data analysis are used in efficient designing of the experiment at subsequent steps. As pointed out by several authors (see e.g. Chernoff (1975), Fedorov (1972), Silvey (1980)), the most attractive feature of adaptive sequential experiments is their ability to optimally utilize the dynamics of the learning process associated with experimentation, data analysis and inference. At the beginning, the scientist does not have much information, and hence an initial experiment is bound to be somewhat tentative in nature. As the experiment continues and more and more observations are obtained, the scientist is able to form a more precise impression of the underlying theory, and this more definite perception is used to design a more informative experiment in the future that is expected to give rise to more relevant and useful data.

The main purpose of this article is to explore some fundamental theoretical issues that are of critical importance in studying the asymptotics of sequentially designed adaptive nonlinear experiments. In Section 2, we will explore the asymptotic performance of maximum likelihood estimates based on dependent observations generated by adaptive and sequentially designed experiments in a very general nonlinear set up, which amply covers situations occurring in practice and models studied in the literature. Unlike Chaudhuri and Mykland (1993), who considered a very specific adaptive sequential scheme, our focus here will be on general adaptive procedures for sequentially designing experiments and some related asymptotics. In Section 3, we will discuss the limiting behavior of some adaptive design procedures and work out some useful conditions that guarantee their large sample optimality with respect to a broad class of optimal design criteria. Once again, unlike Chaudhuri and Mykland (1993), who concentrated on D-optimal designs only, we will consider a general family of optimal design criteria that includes D-optimality and many other optimality criteria as special cases. In the course of the development of the principal theoretical results in Sections 2 and 3, we observe certain intriguing facts, and some interesting questions, which arise naturally, remain unanswered at this moment. All technical proofs are presented in Appendix.

2. Maximum Likelihood and Related Statistical Inference

Suppose now that the conditional distribution of the response Y given $X = x$ has a probability density function or probability mass function $f(y|\theta, x)$. Here the form of the function f is assumed to be known, and θ is an unknown d-dimensional Euclidean parameter of interest that takes its values in Θ (the parameter space), which will be assumed to be a convex and open subset of R^d. The value of the regressor X is determined by the experimenter, who chooses it from a set Ω (the experiment space) before each trial of the experiment. In applications, Ω can be a finite set (e.g. factorial experiments with each factor having a finite number of levels) or some appropriate compact (i.e. closed and bounded) interval or region in an Euclidean space (e.g. when the regressor X is continuous in nature) or a mixture of the two (e.g. analysis of covariance type problems). As pointed out in Chaudhuri and Mykland (1993), the conditional distribution of Y given $X = x$ may involve an unknown nuisance parameter (e.g. the unknown error variance in a homoscedastic normal regression model) in addition to the parameter of principal interest θ. However, as argued in Chaudhuri and Mykland (1993, p. 539), for a number of models frequently used in practice and extensively discussed in the literature, we can ignore the presence of that nuisance parameter (and thereby keep our notations simple) as long as we are interested in the large sample behavior of the maximum likelihood estimate of θ and optimal designing of the experiment to ensure greatest efficiency in asymptotic inference based on that estimate. We assume that the parametrization in the model $f(y|\theta, x)$ is smooth and regular in the sense that the following Cramér type conditions will be satisfied. From now on all vectors in this article will be column vectors, the superscript T will be used to denote the transpose, and $|\cdot|$ will be used to denote the Euclidean norm of a vector or a matrix.

Condition 2.1. Suppose that the response Y takes its values in a set $\mathcal{R}$ (the response space, which can be finite, countably infinite, an interval or a region in an Euclidean space depending on the situation), which is equiped with a σ-field on it, and μ is a σ-finite measure (which can be the usual counting measure or the Lebesgue measure depending on the situation) on $\mathcal{R}$ such that $\int_{\mathcal{R}} f(y|\theta, x)\mu(dy) = 1$. Then the support of $f(y|\theta, x)$ is $\mathcal{R}$ for all possible values of θ and x. Further, for every fixed $x \in \Omega$ and $y \in \mathcal{R}$, $\log\{f(y|\theta, x)\}$ is thrice continuously differentiable in θ at any $\theta \in \Theta$.

Condition 2.2. Let $\nabla \log\{f(y|\theta, x)\} = G(y, \theta, x)$ be the d-dimensional gradient vector obtained by computing the first order partial derivatives of $\log\{f(y|\theta, x)\}$ with respect to θ. Then $\int_{\mathcal{R}} G(y, \theta, x) f(y|\theta, x)\mu(dy) = 0$ and $\sup_{x\in\Omega} \int_{\mathcal{R}} |G(y, \theta, x)|^{2+t} f(y|\theta, x)\mu(dy) < \infty$ for some $t > 0$.

Condition 2.3. Let $H(y,\theta,x)$ denote the $d\times d$ Hessian matrix of $\log\{f(y|\theta,x)\}$ obtained by computing its second order partial derivatives with respect to θ. Then we have

$$\int_{\mathcal{R}} H(y,\theta,x)f(y|\theta,x)\mu(dy)$$
$$= -\int_{\mathcal{R}}\{G(y,\theta,x)\}\{G(y,\theta,x)\}^T f(y|\theta,x)\mu(dy) = -I(\theta,x),$$

where $I(\theta,x)$ is the $d\times d$ Fisher information matrix associated with the model $f(y|\theta,x)$. Also, $\sup_{x\in\Omega}\int_{\mathcal{R}}|H(y,\theta,x)|^2 f(y|\theta,x)\mu(dy)<\infty$.

Condition 2.4. For every $\theta\in\Theta$, there exists an open neighborhood $N(\theta)\subseteq\Theta$ of θ and a non-negative function $K(y,\theta,x)$ that satisfies $\sup_{x\in\Omega}\int_{\mathcal{R}} K(y,\theta,x)\ f(y|\theta,x)\ \mu(dy)<\infty$, and each of the third order partial derivatives of $\log\{f(y|\theta',x)\}$ with respect to θ' is dominated by $K(y,\theta,x)$ for all $\theta'\in N(\theta)$.

It is obvious that for a typical model in the class of generalized linear models, Conditions 2.1 through 2.4 will hold. As a matter of fact, these regularity conditions are satisfied whenever the conditional distribution of the response given the regressor is modeled using standard exponential families (see e.g. Lehmann (1983)) or various curved exponential families (see Efron (1975, 1978)). Many important heteroscedastic linear regression models, where the response variable follows normal distribution with a variance that is assumed to be a nonlinear function of the mean (see Box and Hill (1974), Bickel (1978), Jobson and Fuller (1980), Carroll and Ruppert (1988), etc.) with a fixed and known form, are curved exponential models satisfying Conditions 2.1 through 2.4. For a standard nonlinear regression model with the error having normal distribution with zero mean and a fixed (but possibly unknown) variance, Conditions 2.1 through 2.4 translate into some regularity conditions on the regression function that depends on the unknown parameter of interest in a smooth and nonlinear way. It is easy to see that such regularity conditions on the regression function are closely related to the conditions used in the literature (see e.g. Jenrich (1969), Wu (1981), Gallant (1987), Seber and Wild (1989)) in order to establish $n^{1/2}$-consistency and asymptotic normality of the nonlinear least squares estimate based on independent observations.

Consider next an adaptive sequential experiment in which the ith design point X_i is allowed to depend on the past data $(Y_1,X_1),\ldots,(Y_{i-1},X_{i-1})$ for $i>1$. Clearly, when data will be generated from such an experiment, we will no longer have a sequence of independent observations, and the standard asymptotics for maximum likelihood estimates based on independent observations will not be applicable. However, the dependence (if any) of Y_i on $(Y_1,X_1),\ldots,(Y_{i-1},X_{i-1})$ will be through X_i. As a result, the likelihood based on $(Y_1,X_1),\ldots,(Y_i,X_i)$

will continue to remain in a product form $\prod_{r=1}^{i} f(Y_r|\theta, X_r)$ for any $i \geq 1$. It is then obvious that the log-likelihood and its various derivatives (with respect to the parameter θ) with suitable adjustments will give rise to sums of martingale difference sequences, which form the rows of certain triangular arrays (see also Chaudhuri and Mykland (1993)). As usual, define the maximum likelihood estimate based on $(Y_1, X_1), \ldots, (Y_n, X_n)$ as

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \prod_{r=1}^{n} f(Y_r|\theta, X_r).$$

Alternatively, $\hat{\theta}_n$ can be viewed as a root of the likelihood equation

$$\sum_{r=1}^{n} \nabla \log\{f(Y_r|\theta, X_r)\} = \sum_{r=1}^{n} G(Y_r, \theta, X_r) = 0.$$

In some cases, the likelihood may have multiple maxima, and in those cases the likelihood equation will have multiple roots. However, for generalized linear models, the concavity of the log-likelihood will guarantee the uniqueness of the maximum likelihood estimate for suitable sample sizes. For a standard nonlinear regression problem, appropriate conditions on the regression function will ensure the uniqueness (at least in large samples) of the nonlinear least squares estimate. The following results yield very general sufficient conditions, which are to be satisfied by the sequence of design points, for consistency, $n^{-1/2}$ convergence rate and asymptotic normality of the maximum likelihood estimate, when it is based on dependent data arising from an adaptive sequential experiment.

Result 2.5. Assume that Conditions 2.1 through 2.4 hold, and the sequence of observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ is generated from an adaptive sequential experiment. Let λ_n denote the smallest eigenvalue of the average Fisher information matrix $n^{-1}\sum_{r=1}^{n} I(\theta, X_r)$ up to the nth trial. Suppose that the design scheme is such that for some positive constant $\alpha < 1/4$, $n^{\alpha}\lambda_n$ remains bounded away from zero *in probability* as n tends to infinity. Then there exists a choice of the maximum likelihood estimate (i.e. a root of the likelihood equation) which will be *weakly* consistent.

A proof of Result 2.5, which makes use of certain asymptotic properties of the martingales associated with the likelihood and its derivatives, will be given in the Appendix. A natural question that arises at this point is how to ensure that $n^{\alpha}\lambda_n$ remains bounded away from zero *in probability* for some positive $\alpha < 1/4$ as n tends to infinity. We will now exhibit a simple way of designing the experiment so that this condition holds. First observe that for several models used in practice, there exists a probability measure ξ_0 on Ω such that $\int_{\Omega} I(\theta, x)\xi_0(dx)$ is a positive definite matrix for all $\theta \in \Theta$, and frequently ξ_0 can be taken to be the uniform probability measure (discrete or continuous depending on the nature of Ω) on Ω. Let $k_1 < \ldots < k_m < \ldots$ be an increasing sequence of positive integers such that $m^{-1/(1-\alpha)}k_m$ tends to one as m tends to infinity. Then choose the design points $X_{k_1}, \ldots, X_{k_m}, \ldots$ as independent and identically distributed random elements in Ω with ξ_0 as their common distribution. Note that these design points are chosen in a non-adaptive manner. Further, these design points can be chosen without depending on the rest of the design points (i.e. the X_i's for which $i \neq k_m$ for any m), the latter ones being chosen in some appropriate adaptive way. It is straight forward to verify that such a choice of the design points will guarantee the condition assumed about λ_n in the statement of Result 2.5.

Note at this point that the condition assumed in Result 2.5 on the minimum eigenvalue of the average Fisher information matrix is weaker than asymptotic positive definiteness of the average information. However, it is sufficient for *weak* consistency of the maximum likelihood estimate. In the next couple of results, we will consider sufficient conditions for $n^{1/2}$-consistency and asymptotic normality of the maximum likelihood estimate.

Result 2.6. Assume that Conditions 2.1 through 2.4 hold, and the observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ are generated by means of adaptive sequential trials. Suppose that the design scheme is such that the average Fisher information up to the nth trial $n^{-1}\sum_{r=1}^{n} I(\theta, X_r)$ converges *in probability* to a nonrandom positive definite matrix A as n tends to infinity. Then there exists a choice of the maximum likelihood estimate $\hat{\theta}_n$ such that as n tends to infinity, $n^{1/2}(\hat{\theta}_n - \theta)$ converges *weakly* to a d-dimensional normal random vector with zero mean and A^{-1} as the dispersion matrix.

A proof of this result, which utilizes a standard martingale central limit theorem, will be given in the Appendix. The following is an easy consequence of Result 2.6, and is quite useful in statistical applications.

Result 2.7. Assume that $I(\theta, x)$ is a continuous function of both of its arguments, where θ varies in Θ and x varies in the compact space Ω. Then under the conditions assumed in Result 2.6, $n^{-1}\sum_{r=1}^{n} I(\hat{\theta}_n, X_r)$ converges to A *in probability* as n tends to infinity for an appropriate choice of the maximum likelihood estimate $\hat{\theta}_n$, and consequently the *weak* limit of $\{\sum_{r=1}^{n} I(\hat{\theta}_n, X_r)\}^{1/2}(\hat{\theta}_n - \theta)$ will be d-dimensional normal with zero mean and the $d \times d$ identity matrix as the dispersion matrix. As a result, the confidence ellipsoid for the parameter θ, which can be constructed using the χ^2 distribution with d degrees of freedom and based on the maximum likelihood estimate $\hat{\theta}_n$ together with the estimated total Fisher information up to the nth trial $\sum_{r=1}^{n} I(\hat{\theta}_n, X_r)$, will asymptotically have the right coverage probability.

A particular adaptive procedure for designing nonlinear experiments was proposed in Chaudhuri and Mykland (1993), and it has been shown by these authors to satisfy the limiting condition on the design sequence assumed in Results 2.6 and 2.7. Results 2.5 through 2.7 appear to be the most general sufficient conditions that are available at this moment for consistency, $n^{-1/2}$ convergence rate and asymptotic normality of the maximum likelihood estimate constructed from dependent observations generated by a sequentially designed adaptive experiment in a nonlinear set up. The fundamental implication of Results 2.6 and 2.7 is that as long as the *weak* limit of the average Fisher information, which happens to be a random object in finite samples in view of the adaptive sequential nature of the design, is degenerate and positive definite, valid asymptotic inference is possible based on the maximum likelihood estimate just like the large sample frequentist inference in fixed design problems, where successive observations are independent. In other words, the adaptive sequential nature of the experiment can essentially be ignored for making large sample inference even though there will be randomness in selected design points and dependence in the sequence of observations in such an experiment. It has been established in the literature (see e.g. Johnson (1970), Johnson and Ladalla (1979), LeCam (1986), Prakasa Rao (1987)) that under some standard smoothness conditions (including some Cramér type conditions) on the likelihood and the prior distribution (i.e. a probability measure on the parameter space Θ) many desirable asymptotic properties (in the frequentist sense) hold for the Bayes estimate and related Bayesian inference based on independent data (e.g. $n^{1/2}$-consistency and asymptotic normality of the Bayes estimate, consistency of the posterior and its asymptotic expansion, asymptotic accuracy of the Bayesian credible region constructed through highest posterior distribution, etc.). A natural question, which arises at this point and is currently being investigated by the authors, is to what extent analogous asymptotic results will hold if Bayesian techniques are applied to dependent data arising from adaptive sequential trials. Recently Woodroofe (1989) has explored some issues closely related to this in the context of adaptive linear models that arise in some sequentially designed experiments.

The *weak* convergence of the average Fisher information to a degenerate positive definite limit is a crucial ergodicity condition related to the growth of the martingale processes intrinsically associated with the likelihood (see also Hall and Heyde (1980), Sweeting (1980, 1983), Prakasa Rao (1987), who considered maximum likelihood estimation in various dependent processes). It is desirable that the chosen design ensures optimality of that *weak* limit with respect to some suitable optimal design criterion. Then Results 2.6 and 2.7 together with such an asymptotic optimality (if it holds) of the generated design will guarantee first order efficiency of likelihood based parameter estimates and related inference in sequentially designed adaptive experiments. In the following section, we consider limiting properties of design sequences generated by some adaptive procedures in terms of the asymptotic behavior of the average Fisher information matrix $n^{-1}\sum_{r=1}^{n} I(\theta, X_r)$ and establish the large sample optimality of those procedures with respect to a broad and useful class of optimal design criteria.

3. Asymptotic Optimality of Designs

Consider the following real valued functions defined on the space of $d \times d$ positive definite matrices. They are frequently used to define standard optimal design criteria based on the Fisher information matrix associated with an experiment.

(1) $\Phi(A) = -\log\{\det(A)\}$ (D-optimality criterion).
(2) $\Phi(A) = \text{trace}(A^{-1})$ (A-optimality criterion).
(3) $\Phi(A) = \text{trace}(A^{-p})$ (Kiefer's Φ_p optimality criterion with p being a fixed positive integer).

An important property of each of these functions is that it is continuous, strictly convex and remains bounded below as A varies over any convex and bounded subset of the space of $d \times d$ positive definite matrices (see e.g. Kiefer (1974)). Suppose now that Ω is a compact metric space, and $I(\theta, x)$ is a continuous function of both of its arguments. Following Kiefer (1959, 1961) and Kiefer and Wolfowitz (1959), let the design space $\mathcal{D}(\Omega)$ be defined as the collection of all possible probability measures on Ω. We will assume that for every $\theta \in \Theta$, there exists a $\xi \in \mathcal{D}(\Omega)$ such that the matrix $\int_\Omega I(\theta, x)\xi(dx)$ is positive definite. Next we define a locally Φ-optimal design (see also Chernoff (1953)) $\xi^* \in \mathcal{D}(\Omega)$ at θ as

$$\Phi\left\{\int_\Omega I(\theta, x)\xi^*(dx)\right\} = \min_{\xi \in \mathcal{D}(\Omega)} \Phi\left\{\int_\Omega I(\theta, x)\xi(dx)\right\},$$

where Φ is as defined in (1) or (2) or (3) at the beginning of this section. Then, in view of the compactness of the metric space Ω and the continuity of $I(\theta, x)$, such a locally Φ-optimal ξ^*, which depends on θ, will always exist though it may not be unique. However, the locally Φ-optimal Fisher information $\int_\Omega I(\theta, x)\xi^*(dx)$ will be a uniquely defined positive definite matrix in view of the strict convexity of Φ. Therefore, it is meaningful to look for adaptive procedures generating the design sequence $X_1, \ldots, X_n$ in such a way that the average Fisher information $n^{-1}\sum_{r=1}^{n} I(\theta, X_r)$ converges *weakly* to $\int_\Omega I(\theta, x)\xi^*(dx)$ as n tends to infinity. From now on we will assume that the following condition holds.

Condition 3.1. Ω is a compact metric space, and $I(\theta, x)$ has the form $I(\theta, x) = \{V(\theta, x)\}\{V(\theta, x)\}^T$ for all $\theta \in \Theta$ and $x \in \Omega$, where $V(\theta, x)$ is a continuous R^d valued function of both of its arguments.

Observe that this condition will be trivially satisfied for typical experiment spaces used in practice (e.g. when Ω is a finite set or a compact set in an Euclidean space or some kind of a mixture of the two). Also, for Fisher information matrices associated with frequently occurring nonlinear models (e.g. standard nonlinear regression models, generalized linear models), it is easy to see that this condition holds (see also some of the remarks in Chaudhuri and Mykland (1993, p. 640)). In Section 3.1 that follows, we will explore the asymptotic performance of adaptive schemes for Φ-optimal sequential designs, which generalize the D-optimal design considered by Chaudhuri and Mykland (1993). We will establish some desirable asymptotic properties of the generated design sequence under suitable regularity conditions.

3.1. Adaptive schemes for regular optimal design criteria

We now concentrate on experiments, where the value of n is determined by the available resources, and is known to the experimenter at the beginning. We will carry out n_1 of these n trials at the initial stage when either very little or almost no information on the parameter of interest is available. Let us denote the first n_1 design points by $X_1, \ldots, X_{n_1}$, which are to be chosen in a static (i.e. non-dynamic or non-adaptive) fashion. In the absence of any prior knowledge about the parameter θ, these initial design points can be selected systematically to make them evenly distributed in Ω. Alternatively, the uniform probability distribution on Ω can be used to generate n_1 i.i.d random points in Ω. In some situations, reliable and adequate prior informations on θ may be available. In those cases, numerous theoretical and empirical results available in the literature (see e.g. Atkinson and Hunter (1968), Box (1968, 1970), Rasch (1990), Ford, Torsney and Wu (1992), Haines (1993)) on locally optimal designs for nonlinear experiments can provide useful guidelines for selecting the initial design points incorporating such prior informations. Next, for each i such that $n_1 < i \le n$, the ith design point X_i is to be chosen in a dynamic fashion using the adaptive sequential procedure that minimizes the quadratic form $\{V(\theta_i^*, X_i)\}^T \nabla\Phi \left\{(i-1)^{-1}\sum_{r=1}^{i-1} I(\theta_i^*, X_r)\right\} \{V(\theta_i^*, X_i)\}$. Here θ_i^* is an estimate of θ based on data available up to the current stage (i.e. the observations $(Y_1, X_1), \ldots, (Y_{i-1}, X_{i-1})$ obtained at the $(i-1)$th trial and before), and $\nabla\Phi$ denotes the derivative matrix of Φ as defined and discussed in Wu and Wynn (1978, pp. 1274-1275). For illuminating discussions on some related algorithms for sequential generation of design points for special as well as general optimality criteria, the reader is referred to Wynn (1970, 1972), Fedorov (1972), Atwood (1973, 1976), Tsay (1976), Pazman (1986), Kitsos (1989), Robertazzi and Schwartz (1989), etc.. Clearly, the selection rule for X_i, where $n_1 < i \le n$, is myopic in nature as it looks only one step into the future beyond present (note that since we are selecting only one design point at any particular stage of selection, the implementation of the algorithm becomes quite convenient). Nevertheless, the following result, which generalizes an earlier D-optimality result in Chaudhuri and Mykland (1993, Theorem 3.5), establishes a set of simple sufficient conditions that ensure asymptotic Φ-optimality of such a myopic scheme. A proof of this result, which utilizes some key technical results developed in Wu and Wynn (1978) in the course of their exploration of the convergence properties of general step-length algorithms for regular optimal design criteria, will be given in the Appendix.

Result 3.2. Assume that Condition 3.1 holds, and the initial static experiment is chosen in such a way that n_1 tends to infinity, n_1/n tends to zero, and the smallest eigenvalue of the matrix $n_1^{-1}\sum_{r=1}^{n_1} I(\theta, X_r)$ remains bounded away from zero as n goes to infinity. Further, assume that the estimates θ_i^*'s $(n_1 < i \leq n)$ used in the adaptive sequential stage of the experiment satisfy the following consistency and stability conditions. (a) $\max_{n_1<i\leq n} Pr\{|\theta_i^* - \theta| > \epsilon\}$ tends to zero as n tends to infinity for any $\epsilon > 0$.
(b) The sum $\sum_{i=n_1+1}^{n-1} \left|\Phi\left\{i^{-1}\sum_{r=1}^{i} I(\theta_{i+1}^*, X_r)\right\} - \Phi\left\{i^{-1}\sum_{r=1}^{i} I(\theta_i^*, X_r)\right\}\right|$ remains bounded *in probability* as n tends to infinity.
Then, if the design points X_i's with $n_1 < i \leq n$ are chosen following the scheme described above, the average Fisher information $n^{-1}\sum_{r=1}^{n} I(\theta, X_r)$ converges *in probability* to the locally Φ-optimal (at θ, the true value of the unknown parameter) Fisher information $\int_\Omega I(\theta, x)\xi^*(dx)$ as n tends to infinity.

It will be appropriate to note here that in view of some of the remarks in Chaudhuri and Mykland (1993, p. 543), it is not difficult to choose the initial design sequence properly so that the conditions mentioned at the beginning of the statement of Result 3.2 will be satisfied. Also, it is possible to show that there exist estimates θ_i^*'s $(n_1 < i \leq n)$ satisfying the required conditions (a) and (b) if we follow the principal ideas behind some of the explicit constructions given by Chaudhuri and Mykland (1993, p. 543), who restricted their attention to D-optimal designs and used some similar conditions. For an insight into some of the consequences and implications of these two technical conditions, the reader is referred to Chaudhuri and Mykland (1993, Section 3). An important issue that remains unresolved at this moment is the effect of the choices of the initial static experiment and the parameter estimates θ_i^*'s on the finite sample accuracy of the maximum likelihood estimate $\hat{\theta}_n$. The authors are currently looking into certain martingale expansions related to the log-likelihood and its derivatives in an attempt to explore beyond the asymptotics that is achievable through a martingale central limit theorem. It is hoped that such expansions will provide deeper insights into the problem leading towards more explicit guidelines for choosing the initial static experiment and a more effective construction of the estimates θ_i^*'s.

Recently McLeish and Tosh (1990) have reported an extensive simulation study on some specific sequential designs in bioassay. To investigate the finite sample performance of our adaptive sequential designs, we ran some simulations with the two parameter simple linear logistic regression model (i.e. $\Pr(Y = 1|X) = \exp(\beta_0 + \beta_1 X)\{1 + \exp(\beta_0 + \beta_1 X)\}^{-1}$). The real valued covariate X was allowed to vary in the unit interval $[0,1]$. It is possible to show in this case (see e.g Minkin (1987)) that the D-optimal design is uniform and supported on a two point subset of $[0,1]$. This two point subset varies depending on the parameter vector $\theta = (\beta_0, \beta_1)$. Specifically, when $\theta = (0.0, 0.0)$, the optimal design is the uniform distribution on $\{0.0, 1.0\}$, while for $\theta = (3.0, 5.0)$, it is uniform on $\{0.0, 0.41\}$. We decided to take $n = 55$ and the size of the initial static design was chosen to be $n_1 = 15$ with first 15 design points evenly placed in the interval $[0,1]$. After that for $n_1 < i \leq n$, the design points X_i's were determined one by one using the adaptive D-optimality criterion as described at the beginning of this Section 3.1. We chose θ_i^* = the maximum likelihood estimate based on $(Y_1, X_1), \ldots, (Y_j, X_j)$ if $j < i \leq j + 10$ for $j = 15, 25, 35, 45$. In other words, the estimate was updated after the 15th, the 25th, the 35th and the 45th trials. In the case $\theta = (0.0, 0.0)$, the values of the sequentially selected design points X_{16} through X_{55} were observed to alternate between 0.0 and 1.0, while for $\theta = (3.0, 5.0)$ they were alternating between 0.0 and 0.41. Even with several other choices of θ, we always observed the sequentially chosen values of X_{16} through X_{55} oscillating between two supporting points of the true optimal design. Even though the maximum likelihood estimates computed at various stages happened to be somewhat away from the true parameters due to sampling variations caused by the randomness in the data, the adaptive sequential procedure did not miss the optimal design in any of the simulations that we ran.

3.2. Some technical issues and concluding remarks

Result 3.2 asserts the existence of asymptotically Φ-optimal designs that will guarantee asymptotic normality as well as first order efficiency of the maximum likelihood estimate based on dependent data arising from adaptive and sequentially designed nonlinear experiments. However, it does not give the freedom to do it in an arbitrary way. All of the n_1 static design points have to be at the beginning of the experiment, and in the general case one may have to use a batch sequential scheme rather than a fully sequential approach (see also some of the remarks in Chaudhuri and Mykland (1993, Section 3, p. 543)). This raises the question : whether it is possible to have fully adaptive and sequential schemes that will lead to asymptotically Φ-optimal designs. A partial answer to this question is provided by the following Result.

Result 3.3. Assume that Ω is a compact metric space, $I(\theta, x)$ is a continuous function of both of its arguments, and there exists $\xi_0 \in \mathcal{D}(\Omega)$ such that $\int_\Omega I(\theta, x)\xi_0(dx)$ is positive definite for all $\theta \in \Theta$. With the exception of the static design points $X_{k_1}, \ldots, X_{k_m}, \ldots$ that are chosen following the idea described in the remark immediately after Result 2.5, choose the design point X_i ($i \neq k_m$ for any $m \geq 1$) using the probability distribution ξ_i^* on Ω. Here ξ_i^* denotes a locally Φ-optimal design at θ_i^*, and θ_i^*'s form a sequence of consistent estimates of θ such that θ_i^* is based on the data $(Y_1, X_1), \ldots, (Y_{i-1}, X_{i-1})$ obtained up to the $(i-1)$th trial. Then the average Fisher information $n^{-1}\sum_{r=1}^{n} I(\theta, X_r)$ converges *in probability* to the locally Φ-optimal (at θ, the true value of the unknown parameter) Fisher information $\int_\Omega I(\theta, x)\xi^*(dx)$ as n tends to infinity.

Note that in view of Result 2.5 and the remark following it, θ_i^* can be taken to be the maximum likelihood estimate based on data up to the $(i-1)$th trial. However, it is not at all clear how good or useful is the procedure outlined in Result 3.3. In practice, it may be difficult to determine ξ_i^* (i.e. a locally Φ-optimal design at θ_i^*) explicitly, and generating a random element from Ω using the probability measure ξ_i^* may be even harder. We therefore recommend the scheme described in Result 3.2 whenever possible. Nevertheless, it has substantial theoretical interest that an asymptotically Φ-optimal adaptive sequential design as described in Result 3.3 can be established under such minimal conditions. Note that for Result 3.3, we only need the continuity of the information matrix $I(\theta, x)$, and no special structure (like the one assumed in Condition 3.1) is necessary. In particular, this result is applicable to many heteroscedastic regression models, where the conditional distribution of the response given the regressor is assumed to be normal with a variance that depends on the mean in a nonlinear way. Result 3.3 does make one wonder if it is possible to find more convenient, fully adaptive and sequential designs that will ensure first order efficiency of the maximum likelihood estimate of a parameter in a smooth and regular nonlinear model.

Finally, we would like to close this section by pointing out that the sequence of non-adaptive design points $X_{k_1}, \ldots, X_{k_m}, \ldots$ are introduced here to ensure consistency of the maximum likelihood estimate. While this is a convenient strategy that can be used to achieve consistency in a very general set up, one should be able to avoid it if necessary, and use more natural and easier methods in specific situations. Note that in view of the fact that $m^{-1/(1-\alpha)}k_m$ tends to one as m tends to infinity, the sequence $X_{k_1}, \ldots, X_{k_m}, \ldots$ does not affect the first order asymptotic properties of the $n^{1/2}$-consistent maximum likelihood estimate $\hat{\theta}_n$.

Acknowledgements

The research of Probal Chaudhuri was partially supported by a grant from

Indian Statistical Institute. The research of Per A. Mykland was partially supported by NSF grant DMS 93-05601. The authors thank the Referee and the Associate Editor for several helpful comments.

Appendix : The Proofs

Proof of Result 2.5. Let us introduce an increasing sequence of σ-fields $\mathcal{F}_{ni}$'s with $1 \le i \le n$ such that $\mathcal{F}_{ni}$ is generated by $Y_1, \ldots, Y_i$ (in other words, $\sigma(Y_1, \ldots, Y_i) = \mathcal{F}_{ni}$). Then it follows from Condition 2.2 that $\{\sum_{r=1}^{i} G(Y_r, \theta, X_r); \mathcal{F}_{ni}\}_{1 \le i \le n}$ is a square integrable martingale. Further, we must have

$$E\left|\sum_{r=1}^{n} G(Y_r, \theta, X_r)\right|^2 = E\left\{\sum_{r=1}^{n} |G(Y_r, \theta, X_r)|^2\right\} = O(n) \quad \text{as} \quad n \longrightarrow \infty,$$

so that

$$\sum_{r=1}^{n} G(Y_r, \theta, X_r) = O_P(n^{1/2}) \quad \text{as} \quad n \longrightarrow \infty. \tag{A.1}$$

Next observe that $\{\sum_{r=1}^{i}[H(Y_r, \theta, X_r) + I(\theta, X_r)]; \mathcal{F}_{ni}\}_{1 \le i \le n}$ is another square integrable martingale, and therefore a similar argument as above implies that

$$\sum_{r=1}^{n}[H(Y_r, \theta, X_r) + I(\theta, X_r)] = O_P(n^{1/2}) \quad \text{as} \quad n \longrightarrow \infty. \tag{A.2}$$

For any $\delta > 0$, let $N_\delta(\theta)$ denote the neighborhood with θ as the center and the radius δ. It is now easy to see using Condition 2.4 and the condition assumed on λ_n in the statement of the result that if we choose $\delta_n = n^{-\beta}$, where $\alpha < \beta < (1/2) - \alpha$, the smallest eigenvalue of the Hessian matrix of $n^{\alpha-1} \sum_{r=1}^{n} \log\{f(Y_r|\theta', X_r)\}$ will remain negative and bounded away from zero *in probability* as n tends to infinity for all $\theta' \in N_{\delta_n}(\theta)$. In other words, we will have the following,

$$\lim_{n \to \infty} Pr\left\{\sum_{r=1}^{n} \log\{f(Y_r|\theta', X_r)\} \text{ is concave for } \theta' \in N_{\delta_n}(\theta)\right\} = 1. \tag{A.3}$$

Consider next the following third order Taylor expansion of the log-likelihood around the true parameter θ,

$$\sum_{r=1}^{n} \log\{f(Y_r|\theta', X_r)\} = \sum_{r=1}^{n} \log\{f(Y_r|\theta, X_r)\} + (\theta' - \theta)^T \left\{\sum_{r=1}^{n} G(Y_r, \theta, X_r)\right\} + (\theta' - \theta)^T \left\{\sum_{r=1}^{n} H(Y_r, \theta, X_r)\right\}(\theta' - \theta) + R_n(\theta', \theta'). \tag{A.4}$$

Here, in view of Condition 2.4, the remainder term in (A.4) will satisfy

$$\sup_{\theta':|\theta'-\theta|\le\delta_n} |R_n(\theta',\theta)| = O_P(n\delta_n^3) \quad \text{as} \quad n \longrightarrow \infty. \tag{A.5}$$

Results (A.1) through (A.5) now imply that the probability of the event that the likelihood equation has a root in the interior of $N_{\delta_n}(\theta)$ will tend to one as n tends to infinity. This completes the proof.

Proof of Result 2.6. Applying Result 2.5, let $\hat{\theta}_n$ be a *weakly* consistent solution of the likelihood equation. Consider the following first order Taylor expansion of $n^{-1}\sum_{r=1}^{n} G(Y_r, \hat{\theta}_n, X_r)$ around the true parameter θ.

$$\begin{aligned} 0 = n^{-1}\sum_{r=1}^{n} G(Y_r,\hat{\theta}_n,X_r) &= n^{-1}\sum_{r=1}^{n} G(Y_r,\theta,X_r) \\ &+ \left\{ n^{-1}\sum_{r=1}^{n} H(Y_r,\theta,X_r) + \Delta_n(\theta)\right\}(\hat{\theta}_n - \theta), \end{aligned} \tag{A.6}$$

where $\Delta_n(\theta)$ is a $d\times d$ random matrix such that $|\Delta_n(\theta)|$ tends to zero *in probability* as n tends to infinity in view of Condition 2.4 and the *weak* consistency of $\hat{\theta}_n$. Further, in view of Conditions 2.2 and 2.3, the design condition in the statement of the result and the martingale central limit result stated in Corollary 3.1 in Hall and Heyde (1980, pp. 58-59), we conclude that $n^{-1/2}\sum_{r=1}^{n} G(Y_r,\theta,X_r)$ converges *weakly* to a d-dimensional normal random vector with zero mean and A as the dispersion matrix. Note that Corollary 3.1 in Hall and Heyde (1980, pp. 58-59) is stated for real valued martingales, and we need to argue here via the well known Cramér-Wold device. The conditional Lindeberg condition that is needed for applying this Corollary is satisfied because of the moment restriction imposed in Condition 2.2. Also, the condition on the conditional variance process assumed in that corollary can be easily verified using Condition 2.3 and the design condition assumed in the statement of the result. Finally, using (A.2) we get that $n^{-1}\sum_{r=1}^{n} H(Y_r,\theta,X_r)$ must converge *in probability* to the non-random positive definite matrix A. The proof is now complete using (A.6).

Proof of Result 2.7. The *weak* consistency of $\hat{\theta}_n$ implies that for any $\delta > 0$, the probability of the event $\{\, \hat{\theta}_n \in N_\delta(\theta)\,\}$ will tend to one as n tends to infinity. The continuity of the Fisher information matrix in both of its arguments ensures its uniform continuity as the two arguments vary in compact subsets contained in their respective domains. In particular, we must have

$$\max_{1\le r\le n} \left| I(\hat{\theta}_n, X_r) - I(\theta, X_r)\right| \xrightarrow{P} 0 \quad \text{as} \quad n \longrightarrow \infty. \tag{A.7}$$

The *weak* convergence of the estimated average information $n^{-1}\sum_{r=1}^{n} I(\hat{\theta}_n, X_r)$ to A and subsequent assertions in the statement of the result now follow immediately from (A.7) and Result 2.6.

Proof of Result 3.2. We begin by introducing some notations. Suppose that the design points X_i's are generated following the strategy described before the statement of the result. For any $\theta' \in \Theta$ and $1 \le m \le n$, let us write $J_m(\theta') = m^{-1}\sum_{r=1}^{m} I(\theta', X_r)$, and set $J^* = \int_\Omega I(\theta, x)\xi^*(dx)$, where θ is the true parameter and ξ^* is a locally Φ-optimal design at θ. Also, following Wu and Wynn (1978), we define

$$\nabla\Phi(M_1, M_2) = \lim_{\rho\to 0+} \frac{\partial\phi}{\partial\rho}\{(1-\rho)M_1 + \rho M_2\}$$

and

$$\nabla^2\Phi(M_1, M_2) = \lim_{\rho\to 0+} \frac{\partial^2\Phi}{\partial\rho^2}\{(1-\rho)M_1 + \rho M_2\},$$

where M_1 and M_2 are positive definite matrices. Since each of the functions defined in (1), (2) and (3) at the beginning of Section 3 are nicely differentiable, we will have $\nabla\Phi(M_1, M_2) = \text{trace}\{\nabla\Phi(M_1)(M_2 - M_1)\}$. Here for $M = (m_{ij})$, the derivative $\nabla\Phi(M)$ of Φ evaluated at M is defined as $\nabla\Phi(M) = \left(\frac{\partial\Phi(M)}{\partial m_{ij}}\right)$. A second order Taylor expansion now yields the following for $n_1 \le i < n$,

$$\begin{aligned}
&\Phi\{J_{i+1}(\theta^*_{i+1})\}\\
&= \Phi\{J_i(\theta^*_{i+1})\} + (i+1)^{-1}\nabla\Phi\{J_i(\theta^*_{i+1}), I(\theta^*_{i+1}, X_{i+1})\}\\
&\quad + 2^{-1}(i+1)^{-2}\nabla^2\Phi\{(1-\rho_i)J_i(\theta^*_{i+1}) + \rho_i I(\theta^*_{i+1}, X_{i+1}), I(\theta^*_{i+1}, X_{i+1})\}, \quad (A.8)
\end{aligned}$$

where $0 \le \rho_i \le (i+1)^{-1}$. Next, it follows from some crucial observations in Wu and Wynn (1978, Section 3) and the conditions assumed in the statement of the result that $\Phi\{J_i(\theta^*_{i+1})\}$ remains bounded (uniformly in i) *in probability* as n tends to infinity and i varies between n_1 and $n-1$. Now, fix $\eta > 0$ and suppose that $\Phi\{J_i(\theta^*_{i+1})\} > \Phi(J^*) + \eta$ for all i with $n_1 \le i < n$. Then, using the convexity of Φ and the method of construction of X_{i+1}, we get (see also the arguments used in connection with (2.2) in Wu and Wynn (1978))

$$\begin{aligned}
0 < \eta &< \Phi\{J_i(\theta^*_{i+1})\} - \Phi(J^*)\\
&\le -\nabla\Phi\{J_i(\theta^*_{i+1}), J^*\} \le -\nabla\Phi\{J_i(\theta^*_{i+1}), I(\theta^*_{i+1}, X_{i+1})\}. \quad (A.9)
\end{aligned}$$

On the other hand, the assumptions in the statement of the result ensure that the $\nabla^2\Phi$ term (i.e. the third term on the right) in (A.8) remains bounded (again uniformly in i) *in probability* as n tends to infinity and i varies between n_1 and $n-1$. Therefore, using the condition that n_1 tends to infinity as n tends to infinity and (A.9), we have

$$\lim_{n\to\infty} \min_{n_1 \le i < n} Pr\left[\Phi\{J_{i+1}(\theta^*_{i+1})\} \le \Phi\{J_i(\theta^*_{i+1})\} - (i+1)^{-1}(\eta/2)\right] = 1. \quad (A.10)$$

Note at this point that $\sum_{i=n_1}^{n-1}(i+1)^{-1}$ tends to infinity as n goes to infinity in view of the condition n_1/n tends to zero as n tends to infinity. Condition (b) assumed in the statement of the result now implies the following,

$$\lim_{n\to\infty} Pr\,[\text{ There exists } i \text{ such that } n_1 < i < n \text{ and } \Phi\{J_i(\theta^*_{i+1})\} \leq \Phi(J^*) + \eta] \;=\; 1. \tag{A.11}$$

Further, it follows from (A.8) that

$$\lim_{n\to\infty} \min_{n_1\leq i<n} Pr\,[\Phi\{J_{i+1}(\theta^*_{i+1})\} \leq \Phi\{J_i(\theta^*_{i+1})\} + \eta] \;=\; 1. \tag{A.12}$$

One can conclude from (A.8) and the arguments used in connection with (A.9) and (A.10) (see also the arguments used in connection with (2.5) in Wu and Wynn (1978)) that

$$\lim_{n\to\infty} \min_{n_1\leq i<n} Pr\,[\Phi\{J_{i+1}(\theta^*_{i+1})\} \leq \Phi\{J_i(\theta^*_{i+1})\} \text{ whenever } \Phi\{J_i(\theta^*_{i+1})\} \;>\; \Phi(J^*) + \eta] = 1. \tag{A.13}$$

Finally, (A.11), (A.12), (A.13) and Condition (a) in the statement of the result yield

$$\lim_{n\to\infty} Pr\;[\Phi\{J_n(\theta)\} \quad \leq \quad \Phi(J^*) + 4\eta] = 1.$$

This completes the proof in view of the continuity and strict convexity of Φ.

Proof of Result 3.3. For $n \geq 1$, define the set of positive integers S_n as $S_n = \{r : 1 \leq r \leq n \text{ and } r \neq k_m \text{ for any } m \geq 1\}$. Then, in view of the way the sequence of positive integers k_m's is constructed, $n^{-1}\{\#(S_n)\}$ must tend to one as n tends to infinity. Recall now that Φ is a continuous and strictly convex function. It is easy to show using the continuity of $I(\theta,x)$, the compactness of Ω, the *weak* consistency of the sequence of estimates θ^*_i's and straight forward modifications of some of the arguments in the proof of Lemma A.2 in Chaudhuri and Mykland (1993) that

$$\{\#(S_n)\}^{-1}\sum_{r\in S_n}\int_\Omega I(\theta,x)\xi^*_r(dx) \stackrel{P}{\longrightarrow} \int_\Omega I(\theta,x)\xi^*(dx) \quad \text{as} \quad n\longrightarrow\infty. \tag{A.14}$$

Observe also that $\{\sum_{r\in S_t}[I(\theta,X_r) - \int_\Omega I(\theta,x)\xi^*_r(dx)];\mathcal{F}_t\}_{t\in S_n}$ is a square integrable martingale, where $\mathcal{F}_t$ is the σ-field generated by $(Y_1,X_1),\ldots,(Y_t,X_t)$ (in other words $\sigma\{(Y_1,X_1),\ldots,(Y_t,X_t)\} = \mathcal{F}_t$). Hence, we must have

$$\{\#(S_n)\}^{-1}\sum_{r\in S_n}[I(\theta,X_r) - \int_\Omega I(\theta,x)\xi^*_r(dx)] \stackrel{P}{\longrightarrow} 0 \quad \text{as} \quad n\longrightarrow\infty. \tag{A.15}$$

The desired result is now immediate from (A.14) and (A.15).

References

Abdelbasit, K. M. and Plackett, R. L. (1983). Experimental design for binary data. J. Amer. Statist. Assoc. **78**, 90-98.

Atkinson, A. C. and Hunter, W. G. (1968). The design of experiments for parameter estimation. Technometrics **10**, 271-289.

Atwood, C. L. (1973). Sequences converging to D-optimal designs of experiments. Ann. Statist. **1**, 342-352.

Atwood, C. L. (1976). Convergent design sequences, for sufficiently regular optimality criteria. Ann. Statist. **4**, 1124-1138.

Bates, D. M. and Watts, D. G. (1988). Nonlinear Regression : Analysis and Its Applications. John Wiley, New York.

Bickel, P. J. (1978). Using residuals robustly I : tests for heteroscedasticity, nonlinearity. Ann. Statist. **6**, 266-291.

Box, G. E. P. and Hill, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting. Technometrics **16**, 385-389.

Box, G. E. P. and Hunter, W. G. (1965). Sequential design of experiments for nonlinear models. In Proceedings of the IBM Scientific Computing Symposium on Statistics, 113-137, October 21-23, 1963.

Box, M. J. (1968). The occurrence of replications in optimal designs of experiments to estimate parameters in non-linear models. J. Roy. Statist. Soc. Ser.B **30**, 290-302.

Box, M. J. (1970). Some experiences with a nonlinear experimental design criterion. Technometrics **12**, 569-589.

Carroll, R. J. and Ruppert, D. (1988). Transformation and Weighting in Regression. Chapman and Hall, London.

Chaudhuri, P. and Mykland, P. A. (1993). Nonlinear experiments : Optimal design and inference based on likelihood. J. Amer. Statist. Assoc. **88**, 538-546.

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

Chernoff, H. (1975). Approaches in sequential design of experiments. In A Survey of Statistical Design and Linear Models (Edited by J. N. Srivastava), 67-90, North Holland, New York.

Cochran, W. G. (1973). Experiments for nonlinear functions (R. A. Fisher Memorial Lecture). J. Amer. Statist. Assoc. **68**, 771-781.

Cox, D. R. and Snell, E. J. (1989). Analysis of Binary Data. Chapman and Hall, London.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Statist. **3**, 1189-1242.

Efron, B. (1978). The geometry of exponential families. Ann. Statist. **6**, 362-376.

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

Ford, I. and Silvey, S. D. (1980). A sequentially constructed design for estimating a nonlinear parametric function. Biometrika **67**, 381-388.

Ford, I., Torsney, B. and Wu, C. F. J. (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. J. Roy. Statist. Soc. Ser.B **54**, 569-583.

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

Ford, I., Titterington, D. M. and Wu, C. F. J. (1985). Inference and sequential design. Biometrika **72**, 545-551.

Gallant, A. R. (1987). Nonlinear Statistical Models. John Wiley, New York.

Haines, L. M. (1993). Optimal design for nonlinear regression models. Comm. Statist. Theory Methods **22**, 1613-1627.

Hall, P. G. and Heyde, C. C. (1980). Martingale Limit Theory and Its Application. Academic Press, New York.

Jenrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. Ann. Math. Statist. **40**, 633-643.

Jobson, J. D. and Fuller, W. A. (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. J. Amer. Statist. Assoc. **75**, 176-181.

Johnson, R. A. (1970). Asymptotic expansions associated with posterior distributions. Ann. Math. Statist. **41**, 851-864.

Johnson, R. A. and Ladalla, J. N. (1979). The large sample behavior of posterior distributions when sampling from multiparameter exponential family models, and allied results. Sankhyā, Ser.B **41**, 196-215.

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

Kiefer, J. (1961). Optimum designs in regression problems, II. Ann. Math. Statist. **32**, 298-325.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). Ann. Statist. **2**, 849-879.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. Ann. Math. Statist. **30**, 271-294.

Kitsos, C. P. (1989). Fully-sequential procedures in nonlinear design problems. Comput. Statist. Data Anal. **8**, 13-19.

Le Cam, L. (1986). Asymptotic Methods in Statistical Decision Theory. Springer Verlag, New York.

Lehmann, E. L. (1983). Theory of Point Estimation. John Wiley, New York.

McCormick, W. P., Mallik, A. K. and Reeves, J. H. (1988). Strong consistency of the MLE for sequential design problems. Statist. Probab. Lett. **6**, 441-446.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models. Chapman and Hall, London.

McLeish, D. L. and Tosh, D. (1990). Sequential designs in bioassay. Biometrics **46**, 103-116.

Minkin, S. (1987). Optimal designs for binary data. J. Amer. Statist. Assoc. **82**, 1098-1103.

Myers, R. H., Khuri, A. I. and Carter, W. H. (1989). Response surface methodology : 1966-1988. Technometrics **31**, 137-157.

Pazman, A. (1986). Foundations of Optimum Experimental Design. D. Reidel, Boston.

Powsner, L. (1935). The effects of temperature on the durations of the developmental stages of Drosophila Melanogaster. Physiological Zoology **8**, 474-520.

Prakasa Rao, B. L. S. (1987). Asymptotic Theory of Statistical Inference. John Wiley, New York.

Rasch, D. (1990). Optimum experimental design in nonlinear regression. Comm. Statist. Theory Methods **19**, 4789-4806.

Robertazzi, T. G. and Schwartz, S. C. (1989). An accelerated sequential algorithm for producing D-optimal designs. SIAM J. Sci. Statist. Comput. **10**, 341-358.

Seber, G. A. F. and Wild, C. J. (1989). Nonlinear Regression. John Wiley, New York.

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

Sweeting, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. Ann. Statist. **8**, 1375-1381.

Sweeting, T. J. (1983). On estimator efficiency in stochastic processes. Stochastic Process. Appl. **15**, 93-98.

Treloar, M. A. (1974). Effects of Puromycin on Galactosyltransferase of Golgi Membranes. Unpublished Master's Thesis. University of Toronto.

Tsay, J. Y. (1976). On the sequential construction of D-optimal designs. J. Amer. Statist. Assoc. **71**, 671-674.

Woodroofe, M. (1989). Very weak expansions for sequentially designed experiments: Linear models. Ann. Statist. **17**, 1087-1102.

Wu, C. F. J. (1981). Asymptotic theory of nonlinear least squares estimation. Ann. Statist. **9**, 501-513.

Wu, C. F. J. (1985). Asymptotic inference from sequential design in a nonlinear situation. Biometrika **72**, 553-558.

Wu, C. F. J. and Wynn, H. P. (1978). The convergence of general step-length algorithms for regular optimum design criteria. Ann. Statist. **6**, 1273-1285.

Wynn, H. P. (1970). The sequential generation of D-optimum experimental designs. Ann. Math. Statist. **41**, 1655-1664.

Wynn, H. P. (1972). Results in the theory and construction of D-optimum experimental designs. J. Roy. Statist. Soc. Ser.B **34**, 133-147.

Indian Statistical Institute, Calcutte, India.

Department of Statistics, University of Chicago, IL 60637, U.S.A.

(Received December 1993; accepted March 1995)