



Convergence properties of sequential Bayesian D -optimal designs

Anindya Roy^{a,*}, Subhashis Ghosal^b, William F. Rosenberger^c

^aDepartment of Mathematics and Statistics, University of Maryland, Baltimore Co., Baltimore, MD 21250, USA

^bDepartment of Statistics, North Carolina State University, Raleigh, NC 27695, USA

^cDepartment of Statistics, George Mason University, Fairfax, VA 22030, USA

ARTICLE INFO

Article history:

Received 13 February 2007

Received in revised form

11 April 2008

Accepted 25 April 2008

Available online 8 May 2008

Keywords:

Adaptive designs

Asymptotic normality

Discrete optimal design

Dose–response

Posterior convergence

ABSTRACT

We establish convergence properties of sequential Bayesian optimal designs. In particular, for sequential D -optimality under a general nonlinear location-scale model for binary experiments, we establish posterior consistency, consistency of the design measure, and the asymptotic normality of posterior following the design. We illustrate our results in the context of a particular application in the design of phase I clinical trials, namely a sequential design of Haines et al. [2003. Bayesian optimal designs for phase I clinical trials. *Biometrics* 59, 591–600] that incorporates an ethical constraint on overdosing.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Statement of the problem

We motivate our problem with an example from clinical trials. Let $\Omega = \{d_1, d_2, \dots, d_K\}$ be a discrete space of K treatments and let $X_1, X_2, \dots, X_n \in \Omega$ be the treatments assigned to a sequence of n patients. Let Y_1, Y_2, \dots, Y_n be patient responses. The responses are assumed to follow a location-scale model:

$$\begin{aligned} Y_i | X_i = d_k &\sim \text{Bernoulli}(F_\theta(d_k)), & k = 1, 2, \dots, K, \\ F_\theta(d_k) &= F((d_k - \alpha)/\beta), \end{aligned} \quad (1)$$

where $F(\cdot)$ is a known distribution function with density $f(\cdot)$ with respect to Lebesgue measure, $0 < f(x) < M$ for all $x \in \mathbb{R}$ and some positive constant M and the parameter $\theta = (\alpha, \beta)$ belongs to $\mathbb{R} \times \mathbb{R}^+$. A design for this model is a probability measure $\xi = (\xi_1, \xi_2, \dots, \xi_K)'$ belonging to the K -dimensional simplex

$$\Xi = \left\{ \xi : 0 \leq \xi_k \leq 1 \text{ and } \sum_{k=1}^K \xi_k = 1 \right\},$$

which puts weight ξ_k on treatment d_k . An optimal design is one which optimizes a function of the Fisher information matrix for the parameters of the model over all possible choices of designs.

* Corresponding author.

E-mail addresses: anindya@math.umbc.edu (A. Roy), ghosal@stat.ncsu.edu (S. Ghosal), wrosenbe@gmu.edu (W.F. Rosenberger).

More specifically, suppose $I(\theta; d_k)$ is the information at a single dose level $d_k \in \Omega$. For model (1) the information (cf.: [Silvey, 1980](#)) is given by

$$I(\theta; d_k) = a_k(\theta) \begin{pmatrix} 1 & z_k \\ z_k & z_k^2 \end{pmatrix}, \quad (2)$$

where $a_k(\theta) = f^2(z_k)/[\beta^2 F(z_k)(1 - F(z_k))]$, $z_k = (d_k - \alpha)/\beta$ and $\theta = (\alpha, \beta)'$. The information matrix is given by

$$M_{\xi}(\theta) = \sum_{k=1}^K \xi_k I(\theta; d_k).$$

An optimal design will then maximize an appropriately chosen concave function of the information matrix $M_{\xi}(\theta)$. However, for such nonlinear problems one needs to have knowledge of the unknown parameter θ . One solution is to substitute a best guess for θ in the optimization criterion. This leads to a *locally optimal design*. Another choice is to optimize an expectation of the criterion function with respect to some prior probability distribution of the parameters. This leads to a *Bayesian optimal design*. Let $\pi(\theta)$ be a compactly supported prior probability distribution on $\mathbb{R} \times \mathbb{R}^+$, i.e., $\pi(\Theta) = 1$, where Θ is a compact subset of $\mathbb{R} \times \mathbb{R}^+$. Let the true value of the parameter, θ_0 , be an interior point of Θ . The criterion we will consider in this paper is *D-optimality*, where the design is found as a solution to

$$\xi^* = \arg \max_{\xi \in \Xi} \int_{\Theta} [\log \det(M_{\xi}(\theta))] \pi(\theta) d\theta.$$

Now consider a sequential experiment where patients arrive sequentially in time and each is assigned one of the K treatments, conditional on all previous treatment assignments and responses. Let

$$\mathcal{D}_r = \sigma\{X_1, X_2, \dots, X_r, Y_1, Y_2, \dots, Y_r\}$$

be the σ -algebra generated by the first r treatments and responses. Also let $\mathbf{N}(r) = (N_1(r), N_2(r), \dots, N_K(r))'$, where $N_k(r)$ is the number of patients assigned to dose level d_k after r treatments have been allocated and let $\xi_{r,k} = N_k(r)/r$ be the observed allocation proportion for dose level d_k . Then after r patients we can define

$$M_r(\theta) = \frac{1}{r} \sum_{k=1}^K N_k(r) I(\theta; d_k). \quad (3)$$

[Silvey \(1980, Chapter 7\)](#) expressed unease about what $M_r(\theta)$ actually represents. It is technically not the Fisher information because of the dependence structure of the sequential procedure. We consider that treatments are allocated through a sequential Bayesian *D-optimal* design by maximizing the posterior expectation of the log determinant of the information at each stage of allocation. Let $\pi(\theta|\mathcal{D}_r)$ denote the posterior distribution at stage r for $r = 1, 2, \dots, n$. The $(r + 1)$ th patient is then assigned to the dose level X_{r+1} where

$$X_{r+1} = \arg \max_{d \in \Omega} \int_{\Theta} [\log \det(rM_r(\theta) + I(\theta; d))] \pi(\theta|\mathcal{D}_r) d\theta. \quad (4)$$

At the end of the experiment, we are interested in computing Bayes estimators $\tilde{\theta}_n = E\pi(\theta|\mathcal{D}_n) = \int_{\Theta} \theta \pi(\theta|\mathcal{D}_n)$.

Before we pose the main questions that are answered in this paper, we introduce the following notation. Let $Q_{\theta}(\xi)$ denote the determinant of $M_{\xi}(\theta)$. Also let

$$Q(\xi) := \det(M_{\xi}(\theta_0)) = Q_{\theta_0}(\xi). \quad (5)$$

Simple algebra shows that $Q_{\theta}(\xi) = \xi' \mathbf{Q} \xi$ where elements of \mathbf{Q} are defined as

$$\mathbf{Q} = ((q_{ij})) = 0.5\beta^{-2}(d_i - d_j)^2 a_i(\theta_0) a_j(\theta_0), \quad (6)$$

and $a_i(\theta)$ are defined following (2).

In this paper, we answer the following questions:

1. Does the sequential Bayesian *D-optimal* design converge to the local *D-optimal* design at the true parameter value θ_0 , given by

$$\xi^0 = \arg \max_{\xi \in \Xi} \log(Q(\xi)) \quad (7)$$

We refer to this problem as *convergence of the design measure*.

2. Are the Bayes estimators consistent and asymptotically normal, and if so, what is the correct asymptotic variance?
3. How do we characterize the limiting design (a problem which takes on additional subtlety since Ω is discrete)?

Optimal designs in nonlinear experiments have a rich literature. Earlier work include those by Box and Hunter (1965), Fedorov (1972), Ford and Silvey (1980), Abdelbasit and Plackett (1983), Ford et al. (1985), Minkin (1987) and many others. Ford et al. (1989) provide a comprehensive review of the optimal design literature in nonlinear experiments. Also see Chaudhuri and Mykland (1993, 1995) for other references.

Sequential Bayesian optimal designs stem from earlier work by Tsutakawa (1972, 1980), Zacks (1977), Leonard (1982) and Chaloner (1989). The form of the allocation function in (4) is in Haines (1998) and Haines et al. (2003) use the form for a particular application to phase I clinical trials. The allocation function induces a *response-adaptive allocation procedure*, in which the response sequences are dependent random variables. Convergence and estimation must then be viewed in this context.

Sequential D -optimal designs have a long history in the non-Bayesian context, where at each stage the local optimal design is computed at the current maximum likelihood estimates of the parameters. The problem was explored by White (1975) in her doctoral thesis. Silvey (1980, p. 64) notes the difficulty of proving convergence of the sequentially constructed design to the locally optimal design measure determined at θ_0 . He suggests that techniques used to prove convergence of standard algorithms to sequentially compute D -optimal designs (cf. Wynn, 1970; Fedorov, 1972) could be useful in proving convergence in the sequential design context also. In the frequentist case the problem is largely solved by Wu (1985), Lai (1994), and Chaudhuri and Mykland (1993, 1995) who prove convergence of the sequentially computed maximum likelihood estimators (MLEs) or least squares estimators, give their asymptotic distribution, and show convergence of the sequentially computed information matrices under various assumptions about the regularity of the problem, the asymptotic of the eigenvalues of the observed information matrix and the sample size of the pilot design.

Our goals here are to show similar asymptotic properties of the posterior distribution of the parameter θ and estimators in a Bayesian context and also show convergence of the information matrices. Most of our assumptions are basic regularity assumptions which are similar to those in Wu (1985), Lai (1994) and Chaudhuri and Mykland (1993, 1995). However, to apply the results of Chaudhuri and Mykland (1993, 1995) we would have to assume that the size of the initial sample goes to infinity. This is not the case in our problem, specifically for the clinical trial example considered here. We prove our results when the pilot sample size is fixed. Due to our specific binary structure with compact parameter space we are able to show boundedness of the eigenvalues of the observed information rather than imposing conditions on them. This leads to stronger form of convergence of the optimality criterion.

A critical component in the analysis of sequential Bayesian designs is proving posterior consistency. Hu (1997) proves posterior consistency under the product measure, which is weaker than consistency under the prior measure, and therefore cannot be extended to derive any further asymptotic properties of estimators. Schwartz (1965) studies regular posterior consistency in a general context. Her result has been extended in various ways; see Ghosh and Ramamoorthi (2003) for a detailed account. Although our case is parametric, it is nontrivial because of the dependence structure of the sequentially computed posterior distributions. Indeed, we shall use some techniques that are normally used to establish Bayesian consistency in nonparametric problems. We also show asymptotic normality of the posterior distribution and of the Bayes estimators. We now discuss our particular application to phase I clinical trials and review the literature on convergence of sequential designs.

1.2. Application to phase I clinical trials

In phase I clinical trials, patients enter sequentially in time, each is assigned one of K predetermined dose levels, and the patient is observed for either a toxic or nontoxic response to the dose level. One of the goals of a phase I study is to find a "maximum tolerated dose," which in the parametric case is defined as a quantile of the dose–response curve. Rosenberger and Haines (2002) present a comprehensive review of the literature on phase I clinical trial designs. There has been some controversy in the literature as to whether accurate identification of that quantile through a stochastic approximation–type approach or efficient estimation of the quantile through an optimal design–type approach is more appropriate. The former stochastic approximation–type approach is the basis for the *continual reassessment method* (CRM; O’Quigley et al., 1990) and *escalation with overdose control* (EWOC; Babb et al., 1998). There has been some theoretical work dealing with asymptotic properties of these procedures (see Shen and O’Quigley, 1996 for the CRM; Zacks et al., 1998 for the EWOC procedure). The estimation and optimal design–type approach has been described by Whitehead and Brunier (1995) and generalized by Haines et al. (2003). In this paper, we also concentrate on estimation issues in an optimal design setup.

When response is toxicity and experimentation is on human beings, it is unethical to assign patients to highly toxic doses. Thus the procedure in (4) may not be appropriate. Haines et al. (2003) introduced an *overdosing constraint* (similar to one described by Babb et al., 1998), given by

$$\Pr(\mu_R(\theta) > d) \leq \varepsilon \quad (8)$$

for $\varepsilon > 0$ small, where μ_R is some quantile corresponding to a probability of toxicity Γ_R ; that is, $\mu_R(\theta) = \alpha + \beta F^{-1}(\Gamma_R)$. This leads to their suggested procedure that solves the optimization given in (4), subject to the constraint (8). Note that the constraint in (8) is computed with respect to the distribution of $\mu_R(\theta)$ induced by π . In order to ensure nonsingularity of the design resulting

from application of the sequential scheme under the constraint (8), we need the following assumption. Suppose the constraint $\mu_R(\theta_0) < d$ is satisfied by $K^*(\leq K)$ doses, i.e.,

$$K^* = \#\{d_i \in \Omega : \mu_R(\theta_0) < d_i\}, \tag{9}$$

where # denotes the cardinality of the set. Without loss of generality let the doses be d_1, \dots, d_{K^*} . We will assume throughout that $K^* \geq 2$. Any sensible choice of the dose levels and the value of ε in the constraint (8) will satisfy the condition $K^* \geq 2$.

Practical problems that need to be addressed in a sequential design procedure include initial patient assignments based solely on the prior distribution and appropriate prior elicitation. These and computational issues are addressed in [Rosenberger et al. \(2005\)](#).

1.3. Organization of the paper

In Section 2, we prove consistency of the posterior distributions computed under the sequential design framework. We also comment that all results hold for the clinical trials problem with an overdosing constraint. In Section 3, we prove convergence of the design measures. In Section 4, we derive posterior asymptotic normality. Because we are operating on a discrete design space, characterization of the limiting design measure is different from the usual consideration of continuous designs. In Section 5, we outline a characterization of the limiting design measure for the clinical trial application. We draw conclusions in Section 6.

2. Consistency of the posterior

As given in [Silvey \(1980, Section 7.3\)](#) the general heuristics in deriving the asymptotic justification for a sequential procedure is the following. First, one shows strong consistency of the estimator, then, for large enough samples, the stochastic procedure for choosing the next design point can be essentially replaced by the corresponding deterministic algorithm for finding the optimal design for a known value of the parameter. Then one needs to show convergence of the deterministic algorithm to the optimal design and hence show the convergence of the information corresponding to a design measure based on n observations to the Fisher information for the optimal design. This in turn will guarantee asymptotic normality of the estimators. [Wu \(1985\)](#) followed this approach in the non-Bayesian case, and we will also follow this approach. In this section we discuss strong consistency of the estimator. First we give conditions for strong consistency under the unconstrained sequential Bayesian design. Remark 1 argues how strong consistency continues to hold when the sequential Bayesian design is performed with the overdosing constraint (8). The proof of the main theorem in this section relies on the following lemma which is a variant of a theorem by [Schwartz \(1965\)](#).

Lemma 1. *Let X_1, \dots, X_n be random variables with arbitrary joint distribution $P_{\theta}^n, \theta \in \Theta \subset \mathbb{R}^d$. Assume that every P_{θ}^n admits a joint density $l_n(x_1, \dots, x_n; \theta)$ with respect to a σ -finite measure. Let Π be a prior on θ and $\theta_0 \in \Theta$ be the true value of θ . Suppose that*

1. *for every $\eta > 0$, there exists a $\delta > 0$ such that whenever $\|\theta - \theta_0\| < \delta, \theta \in \Theta$, we have*

$$\liminf_{n \rightarrow \infty} n^{-1} \log \frac{l_n(X_1, \dots, X_n; \theta)}{l_n(X_1, \dots, X_n; \theta_0)} \geq -\eta \quad \text{a.s. } P_{\theta_0}^{\infty},$$

2. *for every $\varepsilon > 0$, there exists a test function $\Phi_n = \Phi_n(X_1, \dots, X_n)$ for testing $H_0 : \theta = \theta_0$ against $H : \|\theta - \theta_0\| > \varepsilon$ such that for some $B, b > 0$,*

$$P_{\theta_0} \Phi_n \leq B e^{-bn}, \quad \sup_{\theta \in \Theta : \|\theta - \theta_0\| > \varepsilon} P_{\theta}(1 - \Phi_n) \leq B e^{-bn},$$

3. *$\Pi(\theta \in \Theta : \|\theta - \theta_0\| < \delta) > 0$ for all $\delta > 0$.*

Then for every $\varepsilon > 0, \Pi(\theta \in \Theta : \|\theta - \theta_0\| > \varepsilon | X_1, \dots, X_n) \rightarrow \text{a.s. } P_{\theta_0}^{\infty}.$

We now state our main theorem.

Theorem 1. *Let \mathcal{D}_n be the data generated by the sequential Bayesian scheme (4). Let $\theta_0 \in \Theta$ and U be an arbitrary neighborhood of θ_0 in Θ where Θ is a compact subset of $\mathbb{R} \times \mathbb{R}^+$. Let θ have prior π such that $\pi(\Theta) = 1$ and π has positive density for some neighborhood of θ_0 . Suppose there exists $M_1 > 0$ such that $\sup_{\theta \in \Theta} \max_{x \in \Omega} \|\partial F_{\theta}(x) / \partial \theta\| = M_1 < \infty$, where $\partial F_{\theta}(x) / \partial \theta$ is the vector of partial derivatives $\partial F_{\theta}(x) / \partial \theta_i$. Then, the posterior probability $\pi(\theta \in U | \mathcal{D}_n) \rightarrow 1$ a.s.*

Remark 1. Theorem 1 also holds under an overdosing constraint, given in (8). It will be clear from the proof of Theorem 1, consistency of the posterior is achieved as long as there are at least two design points that have positive allocation proportion. The following proposition establishes that even under the overdosing constraint, the allocation proportions of at least two design points stay bounded away from zero.

Proposition 1. Let $\xi_n = (\xi_{1,n}, \xi_{2,n}, \dots, \xi_{K,n})'$ be the sequence of observed proportion vectors for design points d_1, d_2, \dots, d_K for the sequential Bayesian design with an overdosing constraint (8). Let the initial allocation be an interior point of the K -dimensional simplex Ξ . Then there exist a positive constant $0 < \eta < 1$ and a positive integer N , possibly depending on η , such that if $n > N$ then for all $\theta \in \Theta$ we have

$$\|\xi_n - e_k\| > \eta/2, \quad k = 1, 2, \dots, K, \tag{10}$$

where e_k is the vertex of Ξ with all zeros except a one at the k th place.

The following corollary of Theorem 1 provides insight into the limiting structure of the design vector for the sequential Bayesian design with an overdosing constraint.

Corollary 1. Suppose $2 \leq K^* < K$, where K^* is defined in (9). Then, under the sequential Bayesian optimal design with overdosing constraint (8), the allocation proportions $\xi_{n,i}$, $i = K^* + 1, \dots, K$, converge to zero almost surely.

3. Consistency of the design measure

Given that the posterior is consistent under the Bayesian sequential design, the limiting properties of the design can be studied under deterministic allocation, which is essentially a quadratic optimization problem. In this section we show that the optimality criterion under the sequential design converges to a limiting value and the corresponding design measure has limit points that are all locally D -optimal. In view of Corollary 1, the limiting values of the design measure are all in a K^* dimensional simplex, i.e., at least $K - K^*$ components are zero.

We will prove the result in slightly more generality as it may be of interest as an optimization result. Suppose the goal is to maximize a continuous function $Q(\xi)$ which admits bounded second partial derivative over the simplex Ξ , i.e., $\|\partial^2 Q(\xi)/\partial \xi \partial \xi'\| < M''$ for all $\xi \in \Xi$, for all $\theta \in \Theta$ and for some constant M'' . In our special case, $Q(\xi) = \xi' Q \xi$ over the K -dimensional simplex where Q is a symmetric matrix with zeros on the diagonal and arbitrary positive off-diagonal entries. Suppose the algorithm chooses to move from a point ξ_n at the n th iteration to

$$\xi_{n+1}^i = (1 - \lambda_n)\xi_n + \lambda_n e_i, \tag{11}$$

where e_i is a vertex of the simplex which is chosen in the direction in which the $Q(\xi)$ is maximized. Suppose we have diminishing stepsize λ_n satisfying the following assumption.

Assumption (SS). $\sum_{n=1}^{\infty} \lambda_n = \infty$ and $\sum_{n=1}^{\infty} \lambda_n^2 < \infty$.

Then the following theorem shows that the iterations ξ_n converge to a point in the simplex.

Theorem 2. Let $\{\xi_n\}$ be the sequence of iterations defined by (11) and let λ_n satisfy Assumption (SS). Then there exists $\xi^* \in \Xi$ such that

$$\lim_{n \rightarrow \infty} Q(\xi_n) = Q(\xi^*),$$

and $\lim_{n \rightarrow \infty} d(\xi_n, A(\xi^*)) = 0$, where $A(\xi^*) = \{\xi : Q(\xi) = Q(\xi^*)\}$ where for any set A and a point x , $d(x, A) = \inf\{\|x - y\| : y \in A\}$.

Remark 2. We have proven convergence of the design criterion to an optimal value. The design measure need not converge if there are multiple optimum values, i.e., if the set of optimal values $A(\xi^*)$ is not a singleton set. If there is a ridge in the design space whose height is $Q(\xi^*)$, then the iterations can travel along the ridge without converging. Thus, the design measure need not converge to a single optimal value in the presence of such ridges.

4. Asymptotic normality of the posterior distribution

Next we derive asymptotic normality of the posterior and that of the Bayes estimators. Let $l_n(\theta)$ be the log-likelihood obtained from the data $(X_1, Y_1), \dots, (X_n, Y_n)$. Let the observed information matrix be defined as

$$\begin{aligned} M_n^*(\theta) &= -n^{-1} \frac{\partial^2}{\partial \theta \partial \theta'} l_n(\theta) \\ &= n^{-1} \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta'} [Y_i \log F_\theta(X_i) + (1 - Y_i) \log(1 - F_\theta(X_i))], \end{aligned} \tag{12}$$

where $l_n(\theta)$ is the log-likelihood function. Let $\hat{\theta}_n$ be the MLE, i.e.,

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} l_n(\theta). \tag{13}$$

Note that there is a direct relationship between the "observed" information matrix $M_n^*(\theta)$ and the matrix $M_n(\theta)$ in (3). If we write the expression (12) as $M_n^*(\theta) = n^{-1} \sum_{i=1}^K I_i^*(\theta)$, then

$$M_n(\theta) = n^{-1} \sum_{i=1}^n E(I_i^*(\theta) | \mathcal{D}_{i-1}).$$

Thus, the matrix $M_n(\theta)$ is not the actual Fisher information matrix, but rather an intermediate quantity between the observed information matrix and the Fisher information matrix. See [Silvey \(1980, Chapter 7\)](#) for more discussion about the matrices $M_n(\theta)$, $M_n^*(\theta)$ and the Fisher information matrix.

The proof of asymptotic normality of posterior uses the fact that the MLE is asymptotically normal. There are many results proving asymptotic normality of the MLE in similar adaptive sequential designs. The most relevant are by [Wu \(1985\)](#) and [Chaudhuri and Mykland \(1995\)](#). Also, [Lai \(1994\)](#) provides conditions for asymptotic normality of least squares estimators in nonlinear regression models with adaptive sequential designs. [Chaudhuri and Mykland \(1995\)](#) show that the MLE is asymptotically normal if the data are generated by an adaptive sequential design in a nonlinear experiment. In order to use their result we need to assume that the initial allocation size n_0 tends to infinity as the sample size increases. This is not directly applicable to the particular application of phase I clinical trial considered here. Also, Chaudhuri and Mykland show a weaker form of consistency for the MLE by assuming a weaker condition on the eigenvalues of the observed information. Due to the binary form of our experiment we are able to show boundedness of the eigenvalues of the observed information directly in Proposition 2. A key step toward the proof of asymptotic normality of the MLE is proving consistency of it. Consistency of some suitable solution of the likelihood equation may be shown using [Hall and Heyde \(1980, Section 6.2\)](#). However, unless the solution is unique, the convergent solution is unknown, a well-known problem associated with the solution of the likelihood equation (cf. [Serfling, 1981](#)). To avoid this difficulty, we use a global maxima. Our approach is useful even in the absence of a martingale structure. Due to the differences in the set of assumptions from those considered in the literature and due to the general appeal of our proof in situations where martingale structure is not applicable, we present the proof of consistency of MLE as a separate theorem in the appendix. Once consistency is established, asymptotic normality follows relatively easily by Taylor's expansion, a martingale central limit theorem and Proposition 2.

[Hu et al. \(2005\)](#) show asymptotic normality of a sequence of solutions of the likelihood equations in a response-adaptive randomization procedure under an exponential family model and the assumption that the design measure converges almost surely to an optimal value. Most proofs require the convergence of the design measure to a unique value. The convergence of the design measure is not necessarily guaranteed for our problem (cf. Remark 2). We circumvent the problem of convergence of the design measure by normalizing the MLE sequence with the observed information. First we prove the boundedness of the eigenvalues of the observed information matrix.

Proposition 2. *Let $M_n^*(\theta)$ be the observed information matrix as defined in (12). Let $\lambda_{\min,n}(\theta)$ and $\lambda_{\max,n}(\theta)$ be the smallest and the largest eigenvalues of $M_n^*(\theta)$, respectively. Then there exist constants $0 < L_{\min} < L_{\max} < \infty$ and $\delta > 0$ such that $L_{\min} < \lambda_{\min,n}(\theta) < \lambda_{\max,n}(\theta) < L_{\max}$ for all n and for all $\theta \in N_\delta(\theta_0)$ where $N_\delta(\theta_0) = \{\theta : \|\theta - \theta_0\| < \delta\}$.*

Define

$$w_n(\theta) = l(\theta) - l(\hat{\theta}_n). \tag{14}$$

Then by Taylor expansion

$$w_n(\hat{\theta}_n + (nM_n^*)^{-1/2}u) = -\frac{1}{2}u'u + R_n(u), \tag{15}$$

where $u = \sqrt{n}M_n^*{}^{1/2}(\theta - \hat{\theta}_n)$, $M_n^* = M_n^*(\hat{\theta}_n)$ is the observed Fisher information matrix (12) evaluated at the MLE $\hat{\theta}_n$, and $R_n(u)$ is the remainder term. We shall show that the L_1 -distance between the posterior density of u

$$\pi_n^*(u) = \pi(u | \mathcal{D}_n) = \frac{e^{w_n(u)} \pi(\hat{\theta}_n + (nM_n^*)^{-1/2}u)}{\int_{\mathbb{R}^2} e^{w_n(t)} \pi(\hat{\theta}_n + (nM_n^*)^{-1/2}t) dt}, \tag{16}$$

and the multivariate normal density $\phi(u) = e^{-(1/2)u'u} / \int_{\mathbb{R}^2} e^{-(1/2)t't} dt$ converges to zero a.s. $[P_{\theta_0}]$.

Theorem 3. *If \mathcal{D}_n satisfy all the assumptions of Theorem 1 and Lemma 1 and if, in addition, Assumption (SS) holds, then*

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}^2} |\pi_n^*(u) - \phi(u)| du = 0 \quad \text{a.s. } [P_{\theta_0}]. \tag{17}$$

The asymptotic normality of the Bayes estimator follows directly from the asymptotic normality of the posterior and the asymptotic normality of the MLE (cf. [Bickel and Yahav, 1969](#)).

Corollary 2. Let $\tilde{\theta}_n = E_{\pi}(\theta | \mathcal{D}_n)$ be the Bayes estimator of θ with respect to a prior π . Then

$$(nM_n^*)^{1/2}(\tilde{\theta}_n - \theta_0) \rightarrow N(0, I), \tag{18}$$

where $M_n^* = M_n^*(\hat{\theta}_n)$ is defined in (12).

5. Characterization of the limiting design

In this section we characterize the limiting design obtained following the sequential Bayesian optimal procedure. Kiefer and Wolfowitz (1960) give the general equivalence theorem for finding the global D -optimal design. For example, one can show that if the design space is large enough, the D -optimal design for the logistic link is a 2-point design, with equal weights at two symmetric quantiles in the z scale. If the design space is not wide enough, the design is still a 2-point design with at least one of the points falling on the boundary of the design space. However, when the design space is discrete, the optimal designs for the continuous design space are no longer optimal. In order to find the design that is D -optimal over the design space $\Omega = \{d_1, d_2, \dots, d_K\}$ one needs to solve the following optimization problem:

$$\max_{\xi \in \Xi} Q(\xi) := \max_{\xi \in \Xi} \xi' \mathbf{Q} \xi, \tag{19}$$

where \mathbf{Q} is defined in (6). Then, the Lagrangian of the problem can be written as

$$\mathcal{L}(\xi, \lambda, \mu) := -\xi' \mathbf{Q} \xi - \lambda c(\xi) - \sum_{i=1}^K \mu_i c_i(\xi), \tag{20}$$

where $c(\xi) = \lambda(\mathbf{J}'\xi - 1)$, $c_i(\xi) = \xi_i$, $i = 1, 2, \dots, K$, and \mathbf{J} is the vector of ones. Because \mathbf{Q} is a matrix with arbitrary positive off diagonal entries, (19) is not a convex optimization problem. Thus, the Karush–Kuhn–Tucker (KKT) conditions are only necessary provided the linear independence constraint qualification (LICQ) conditions are satisfied. However, clearly not all ξ_i can be in the inactive set (inequality constraints) as $\mathbf{J}'\xi = 1$. Suppose, without loss of generality, the first $l (< K)$ components of ξ are in the active set. Then, the LICQ conditions require the vectors in $\{\nabla c(\xi), \nabla c_1(\xi), \dots, \nabla c_l(\xi)\} = \{\mathbf{J}, \mathbf{e}_1, \dots, \mathbf{e}_l\}$ to be linearly independent. Thus, the LICQ are clearly satisfied. Then the KKT first order necessary conditions can be written as

$$\nabla_{\xi} \mathcal{L}(\xi, \lambda, \mu) = \mathbf{0}, \quad \mathbf{J}'\xi = 1, \quad \xi \geq 0, \quad \lambda \geq 0, \quad \mu \geq 0, \quad \mu \cdot \xi = 0, \tag{21}$$

where $\mathbf{a} \cdot \mathbf{b}$ is the Hadamard (componentwise) product between two vectors \mathbf{a} and \mathbf{b} and $\mathbf{a} \geq 0$ means that all components of \mathbf{a} are nonnegative. The last condition in (21) is the complementary slackness condition which implies that at the solution $(\xi^*, \lambda^*, \mu^*)$, for each i , either $\mu_i^* = 0$ or $\xi_i^* = 0$ or both are zero. This in turn implies that there exists a permutation matrix \mathbf{P} such that

$$\mathbf{P}\xi^* = \begin{pmatrix} \xi_1^* \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{P}\mu^* = \begin{pmatrix} \mathbf{0} \\ \mu_2^* \end{pmatrix},$$

where μ_2^* could be identically zero. From the first condition in (21) we see that a solution must necessarily satisfy $\mathbf{PQP}'\mathbf{P}\xi = \lambda^*\mathbf{J}$, or

$$\mathbf{Q}_1 \xi_1^* = \lambda^* \mathbf{J}, \tag{22}$$

where \mathbf{Q}_1 is the upper left block of \mathbf{PQP}' corresponding to ξ_1^* . Then, the set of points satisfying the necessary conditions is

$$\mathcal{A} = \left\{ \xi \in \Xi : \exists \text{ permutation } \mathbf{P} \text{ such that } \xi = \begin{pmatrix} \xi_1 \\ \mathbf{0} \end{pmatrix} \text{ and } \mathbf{PQ}\xi = \begin{pmatrix} \lambda \mathbf{J} \\ \mathbf{x} \end{pmatrix} \text{ for } \lambda > 0 \right\}.$$

Let \mathcal{M} be the set of all $r \times r$ principal minors of \mathbf{Q} for $r \geq 2$. Since diagonal elements of \mathbf{Q} are zero, the principal minors for $r = 1$ are all zero. In order to look at the solution more explicitly we make the following assumption:

Assumption (C). All $r \times r$ ($r \geq 2$) principal minors of \mathbf{Q} are nonsingular.

Consider the set $\mathcal{Q} = \{\mathbf{M} : \mathbf{M} \in \mathcal{M}; \mathbf{M}^{-1}\mathbf{J} > 0\}$. Then from (22) we have

$$\xi^* = (\mathbf{J}'\mathbf{M}^*-1\mathbf{J})^{-1}\mathbf{M}^*-1\mathbf{J}, \tag{23}$$

where $\mathbf{M}^* = \arg \min\{\mathbf{J}'\mathbf{M}^{-1}\mathbf{J} : \mathbf{M} \in \mathcal{Q}\}$. Thus the D -optimal design has positive weights for the components corresponding to the principal minor $\mathbf{M} \in \mathcal{Q}$ for which the sum of all the entries in the inverse is minimum among all minors in \mathcal{Q} . Note that \mathcal{Q} has at the most $2^K - K - 1$ elements. Of course even for moderate K this could be excessively large for the characterization (23) to be useful. However, for some simple cases it can be insightful.

Proposition 3. Suppose $K = 3$. Suppose the entries of \mathbf{Q} , q_{12} , q_{13} and q_{23} are all distinct. Then the D -optimal design is a 3-point design if q_{12} , q_{13} and q_{23} are the three sides of a triangle, i.e., the sum of any two is greater than the remaining third. Otherwise, the D -optimal design is a 2-point design with equal weights at d_i and d_j where $q_{ij} = \max\{q_{12}, q_{13}, q_{23}\}$.

Proof. For simplicity, let $q_{12} = a, q_{13} = b$ and $q_{23} = c$. Then

$$\mathbf{Q}^{-1}\mathbf{J} = (2abc)^{-1} \begin{pmatrix} c(a+b-c) \\ b(c+a-b) \\ a(b+c-a) \end{pmatrix}.$$

Clearly if a, b, c are the sides of a triangle then $\mathbf{Q} \in \mathcal{Q}$. The inverses of the 2×2 minors are $a^{-1}\mathbf{R}, b^{-1}\mathbf{R}$ and $c^{-1}\mathbf{R}$ where $\mathbf{R} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. All 2×2 minors belong to \mathcal{Q} . Then the design will be a 3-point design if

$$\mathbf{J}'\mathbf{Q}^{-1}\mathbf{J} < 2 \min(a^{-1}, b^{-1}, c^{-1}).$$

Equivalently, the D -optimal design is a 3-point design if

$$\frac{2(ab+bc+ca) - a^2 - b^2 - c^2}{2abc} < \frac{2}{\max\{a, b, c\}}.$$

Without loss of generality, if $a = \max\{a, b, c\}$, the above claim is true because $(b+c-a)^2 > 0$. If the triangle inequality is not satisfied with $a = \max\{a, b, c\}$, then the maximum value of the criterion is $a/2$ corresponding to the $(\frac{1}{2}, \frac{1}{2}, 0)$. \square

Proposition 4. Let \mathbf{Q} be as defined in (6). Then $\text{rank}(\mathbf{Q}) = \min(K, 3)$ and all 2×2 and 3×3 principal minors of \mathbf{Q} are nonsingular.

Proof. Note that $\mathbf{Q} = \mathbf{\Lambda}\mathbf{D}\mathbf{\Lambda}$ where $\mathbf{\Lambda}$ is a diagonal matrix with i th diagonal entry $\lambda_i = 2^{-1/2}\beta^{-1}a_i(\theta)$, \mathbf{D} is the matrix with (i, j) th entry $d_{ij} = (d_i - d_j)^2, i, j = 1, 2, \dots, K$, and $a_i(\theta)$ are defined in (6). Because $\lambda_i > 0$ for all i , $\text{rank}(\mathbf{Q}) = \text{rank}(\mathbf{D})$. Now,

$$\mathbf{D} = \mathbf{d}^2\mathbf{J}' + \mathbf{J}\mathbf{d}'^2 - 2\mathbf{d}\mathbf{d}', \tag{24}$$

where $\mathbf{d}^r = (d_1^r, d_2^r, \dots, d_K^r)'$. Because all matrices on the right side of (24) are of rank one, we have $\text{rank}(\mathbf{Q}) \leq 3$. Algebraic computation shows

$$\det(\mathbf{D}) = \begin{cases} -d_{12}^2 & \text{if } K = 2, \\ 2d_{12}d_{23}d_{13} & \text{if } K = 3. \end{cases}$$

Since the design points are all distinct, we have the result. \square

Proposition 5. The D -optimal design is either a 2-point or a 3-point design, where the nonzero weights ξ_1 are of the form $\xi_1 = (\mathbf{J}'\mathbf{M}^{-1}\mathbf{J})^{-1}\mathbf{M}^{-1}\mathbf{J}$ and \mathbf{M} is the principal minor of \mathbf{Q} corresponding to the components with nonzero weights.

Proof. By the KKT conditions, the D -optimal design is necessarily of the form $\xi = (\xi_1' : \mathbf{0}')'$ and $\mathbf{Q}\xi = (\lambda\mathbf{J}' : \mathbf{x}')'$ (after a rearrangement of the doses to put the doses with nonzero weights at the beginning) for some positive constant λ and some vector \mathbf{x} . But by (4), if \mathbf{Q}_{11} is the upper left 3×3 principle minor of \mathbf{Q}_1 where

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \\ \mathbf{Q}_2' & \mathbf{Q}_3 \end{pmatrix}$$

then $\mathbf{Q}_1 = \begin{pmatrix} \mathbf{I} \\ \mathbf{B}' \end{pmatrix} \mathbf{Q}_{11} (\mathbf{I} : \mathbf{B})$ for some matrix \mathbf{B} . Then solving $\mathbf{Q}_1 \xi_1 = \lambda\mathbf{J}$, we obtain

$$\xi_{11} + \mathbf{B}\xi_{12} = \lambda\mathbf{Q}_{11}^{-1}\mathbf{J}, \quad \mathbf{B}'\mathbf{Q}_{11}(\xi_{11} + \mathbf{B}\xi_{12}) = \lambda\mathbf{J}. \tag{25}$$

Therefore, $\mathbf{J}'\mathbf{B} = \mathbf{J}'$. This implies $\lambda\mathbf{J}'\mathbf{Q}_{11}^{-1}\mathbf{J} = \mathbf{J}'\xi_{11} + \mathbf{J}'\mathbf{B}\xi_{12} = \mathbf{J}'\xi_{11} + \mathbf{J}'\xi_{12} = 1$. Thus, the maximum value over a r -point design can be achieved over a 3-point design. However, because the choice of the 3-doses for \mathbf{Q}_{11} within \mathbf{Q}_1 was rather arbitrary, (25) implies that $\lambda = (\mathbf{J}'\mathbf{Q}_{11}\mathbf{J})^{-1}$ for any choice of 3×3 principal minor of \mathbf{Q}_1 . Simple algebra shows that this is not possible unless all elements of \mathbf{Q}_1 are zero. Hence, there are no $r > 3$ point design that satisfy the KKT necessary conditions for local maxima. By Proposition 4, if for some 3×3 principle minor, the elements above the diagonal form three sides of a triangle, then the design with positive weights corresponding to the rows in the principal minor is a 3-point candidate for the D -optimal design and one can eliminate the three 2-point subsets of the 3-point design. The form of the design by the KKT conditions is $\xi_1 = (\mathbf{J}'\mathbf{M}^{-1}\mathbf{J})^{-1}\mathbf{M}^{-1}\mathbf{J}$ where ξ_1 is the subvector of the design vector with positive weights and \mathbf{M} is the corresponding principle minor of \mathbf{Q} . \square

Proposition 3 characterizes the candidate set of D -optimal designs. In the case of design space with K design points, one needs to search within only $\binom{K}{3}$ possible designs. Typically the search will be over an even smaller set as the optimal 2-point design among all possible 2-point design is the one that puts equal weights to the two design points corresponding to the indices of the maximum element in \mathbf{Q} . Because the link function is quite arbitrary, there are cases when a 3-point design gives a solution with a higher value of the optimality criterion even though no principal minor involving the maximum element of \mathbf{Q} satisfies the criterion for a 3-point optimal design.

6. Discussion

We have shown that the sequential Bayesian D -optimal design approach generates data that allows parameters of the model to be estimated consistently. Also the design criterion converges over the iterations. A pertinent question is what is the rate of posterior convergence. Given the finite dimensional parametric structure one can expect to have classical \sqrt{n} rates for convergence. However, the link distribution function is often unknown. Consistency and rate questions become more challenging if one assumes only such a semiparametric model. Specifically, if the distribution function F is assumed to be arbitrary except for some conditions on the quantiles to make the model identifiable, then one can mimic the approach of Haines et al. (2003) by adding some process prior on F . Also error rates for estimation of specific quantiles will be of interest. It seems that in the semiparametric case, the design space must be continuous to be able to retrieve full information about F and the discrete design space at each stage must become gradually dense everywhere in the overall design space. Haines et al. (2003) demonstrate by simulation that the procedure works quite well with $n = 35$ patients, even with a misspecified prior distribution.

Acknowledgments

Anindya Roy’s and William Rosenberger’s research was partially supported by Grant R01-CA87746 from the US National Cancer Institute. Subhashis Ghosal’s research was supported by Grant DMS-0349111 from the US National Science Foundation. We would like to thank the referees for their thorough and helpful comments on this paper.

Appendix

Proof of Lemma 1. The above result follows by proceeding as in the proof of Schwartz’s theorem (cf. Schwartz, 1965) by observing that for any $b > 0$,

$$\begin{aligned} \liminf_{n \rightarrow \infty} e^{bn} \int \frac{l_n(X_1, \dots, X_n; \theta)}{l_n(X_1, \dots, X_n; \theta_0)} d\Pi(\theta) &\geq \liminf_{n \rightarrow \infty} e^{bn} \int_{\|\theta - \theta_0\| < \delta} \frac{l_n(X_1, \dots, X_n; \theta)}{l_n(X_1, \dots, X_n; \theta_0)} d\Pi(\theta) \\ &\geq \int_{\|\theta - \theta_0\| < \delta} \liminf_{n \rightarrow \infty} e^{bn} \frac{l_n(X_1, \dots, X_n; \theta)}{l_n(X_1, \dots, X_n; \theta_0)} d\Pi(\theta) \\ &= \infty \end{aligned}$$

for δ sufficiently small, by Fatou’s lemma, since

$$n \left(b + n^{-1} \log \frac{l_n(X_1, \dots, X_n; \theta)}{l_n(X_1, \dots, X_n; \theta_0)} \right) \rightarrow \infty$$

for all θ sufficiently close to θ_0 by Condition 1. \square

We also require the following lemma in the proof of Theorem 1. The proof of the lemma follows immediately from Taylor’s expansion.

Lemma 2. Let $0 < \varepsilon_0 < \frac{1}{2}$ and $\varepsilon_0 < \gamma, \psi < 1 - \varepsilon_0$. Then there exist a constant L depending only on ε_0 such that

$$\gamma \left(\log \frac{\gamma}{\psi} \right)^m + (1 - \gamma) \left(\log \frac{1 - \gamma}{1 - \psi} \right)^m \leq L(\gamma - \psi)^2, \quad m = 1, 2.$$

Proof of Theorem 1. First, we show the existence of a test function Φ_n satisfying Condition 2 of Lemma 1 in the appendix. Consider a pair of hypotheses $H_0 : \theta = \theta_0$ versus $H_A : \theta = \theta_1$. Let the design space be partitioned as $\Omega_0 = \Omega_0(\theta_1) = \{x \in \Omega : F_{\theta_0}(x) \geq F_{\theta_1}(x)\}$ and $\Omega_1 = \Omega_1(\theta_1) = \Omega \setminus \Omega_0(\theta_1)$. Consider the test $\Phi_n = 1\{T_n > n\varepsilon^*\}$, where

$$T_n = T_n(\theta_1) = \sum_{i: X_i \in \Omega_0} [Y_i - E_{\theta_0}(Y_i | \mathcal{D}_{i-1})] - \sum_{i: X_i \in \Omega_1} [Y_i - E_{\theta_0}(Y_i | \mathcal{D}_{i-1})], \tag{A.1}$$

$\varepsilon_* = \inf\{\delta(\theta) : \|\theta - \theta_0\| > \varepsilon\}$, $\delta(\theta) = \min\{|F_{\theta}(x) - F_{\theta_0}(x)| : x \in \Omega\}$ and $1\{A\}$ denotes the indicator of the set A . Note that $\varepsilon_* > 0$ by the identifiability of the family F_{θ} , continuity of $\theta \rightarrow F_{\theta}$ and the compactness of Θ .

Let

$$Z_i(\theta^*, \theta_1) = \begin{cases} Y_i - E_{\theta^*}(Y_i | \mathcal{D}_{i-1}), & X_i \in \Omega_0(\theta_1), \\ (1 - Y_i) - E_{\theta^*}(1 - Y_i | \mathcal{D}_{i-1}), & X_i \in \Omega_1(\theta_1). \end{cases}$$

Then Z_i is a martingale difference sequence for P_{θ^*} and $-1 \leq Z_i \leq 1$. By Azuma's inequality (cf. Ross, 1996, p. 307),

$$P_{\theta_0}(T_n > n\varepsilon_*) \leq e^{-n\varepsilon_*^2/2}.$$

To bound $P_{\theta}(T_n \leq n\varepsilon_*)$, note that

$$T_n = \sum_{i=1}^n Z_i(\theta_1, \theta_1) + \sum_{i: X_i \in \Omega_0} (F_{\theta_0}(X_i) - F_{\theta_1}(X_i)) - \sum_{i: X_i \in \Omega_1} (F_{\theta_1}(X_i) - F_{\theta_0}(X_i)).$$

Therefore

$$\begin{aligned} P_{\theta_1}(T_n \leq n\varepsilon_*) &= P_{\theta_1} \left(\sum_{i=1}^n Z_i(\theta_1, \theta_1) \leq n\varepsilon_* - \sum_{i: X_i \in \Omega_0} (F_{\theta_0}(X_i) - F_{\theta_1}(X_i)) + \sum_{i: X_i \in \Omega_1} (F_{\theta_1}(X_i) - F_{\theta_0}(X_i)) \right) \\ &\leq P_{\theta_1} \left(\sum_{i=1}^n Z_i(\theta_1, \theta_1) < -n\varepsilon_* \right) \leq e^{-n\varepsilon_*^2/2}. \end{aligned}$$

In order to remove the dependence of θ_1 on the test, where $\|\theta_1 - \theta_0\| > \varepsilon$, consider θ^* such that $\|\theta^* - \theta_1\| < \eta$. Then $\sum_{x \in \Omega} |F_{\theta^*}(x) - F_{\theta_1}(x)| < MK\eta$, and thus

$$\begin{aligned} T_n(\theta_1) &= \sum_{i=1}^n Z_i(\theta^*, \theta_1) + \sum_{i: X_i \in \Omega_0} (F_{\theta_0}(X_i) - F_{\theta^*}(X_i)) - \sum_{i: X_i \in \Omega_1} (F_{\theta^*}(X_i) - F_{\theta_0}(X_i)) \\ &\geq \sum_{i=1}^n Z_i(\theta^*, \theta_1) + \sum_{i: X_i \in \Omega_0} (F_{\theta_1}(X_i) - F_{\theta^*}(X_i)) - \sum_{i: X_i \in \Omega_1} (F_{\theta^*}(X_i) - F_{\theta_1}(X_i)) \\ &\geq \sum_{i=1}^n Z_i(\theta^*, \theta_1) - \sum_{i: X_i \in \Omega_0} |F_{\theta_1}(X_i) - F_{\theta^*}(X_i)| \\ &\geq \sum_{i=1}^n Z_i(\theta^*, \theta_1) - MK\eta. \end{aligned}$$

Thus, if $\eta = \varepsilon_*/(2MK)$, another application of Azuma's inequality gives

$$P_{\theta^*}(T_n(\theta_1) \leq n\varepsilon_*) \leq P_{\theta^*} \left(\sum_{i=1}^n Z_i(\theta^*, \theta_1) \leq -n\varepsilon_*/2 \right) \leq e^{-n\varepsilon_*^2/8}.$$

As Θ is compact, we can cover $\{\theta : \|\theta - \theta_0\| > \varepsilon\}$ by finitely many balls of radius $\varepsilon_*/(2MK)$. Construct tests as in (A.1) for θ_1 equal the center of each ball. Each such test has type I error probability bounded by $e^{-n\varepsilon_*^2/2}$ and type II error probability on the respective ball bounded by $e^{-n\varepsilon_*^2/8}$. Then the maximum of these finitely many tests satisfies Condition 2 of Lemma 1.

To verify Condition 1 of Lemma 1, note that the log-likelihood ratio is given by $A^{(n)}(\theta_0, \theta) = \sum_{i=1}^n A_i(\theta_0, \theta)$, where

$$A_i(\theta_0, \theta) = Y_i \log \frac{F_{\theta_0}(X_i)}{F_{\theta}(X_i)} + (1 - Y_i) \log \frac{1 - F_{\theta_0}(X_i)}{1 - F_{\theta}(X_i)}.$$

Then

$$E(A_i(\theta_0, \theta) | \mathcal{D}_{i-1}) = F_{\theta_0}(X_i) \log \frac{F_{\theta_0}(X_i)}{F_{\theta}(X_i)} + (1 - F_{\theta_0}(X_i)) \log \frac{1 - F_{\theta_0}(X_i)}{1 - F_{\theta}(X_i)},$$

which is bounded above by a multiple of $\|\theta - \theta_0\|^2$ by Lemma 2 and the assumption that $\sup\{\|\partial F_{\theta}(x)/\partial \theta\| : \theta \in \Theta, x \in \Omega\} = M_1 < \infty$. Because

$$A^{(n)}(\theta_0, \theta) = \sum_{i=1}^n \{A_i(\theta_0, \theta) - E(A_i(\theta_0, \theta) | \mathcal{D}_{i-1})\} + \sum_{i=1}^n E(A_i(\theta_0, \theta) | \mathcal{D}_{i-1})$$

and

$$n^{-1} \sum_{i=1}^n \{A_i(\theta_0, \theta) - E(A_i(\theta_0, \theta) | \mathcal{D}_{i-1})\} \rightarrow 0 \quad \text{a.s. } [P_{\theta_0}]$$

by the strong law of large numbers for martingales, it follows that for some constant C_0 , we have $n^{-1} A^{(n)}(\theta_0, \theta) \leq C_0 \delta^2$ a.s. for all sufficiently large n and $\|\theta - \theta_0\| < \delta$. Thus Condition 1 holds.

Condition 3 of Lemma 1 is clearly satisfied. Theorem 1 then follows. \square

Proof of Proposition 1. We will first show that the design vector is bounded away from the vertices in the unconstrained case. It is enough to prove (10) for $k = 1$. Fix $\theta \in \Theta$. Consider $Q_\theta(\xi_{n+1}^1) - Q_\theta(\xi_{n+1}^j)$ for $j \neq 1$ where

$$\xi_{n+1}^j = \frac{n}{n+1} \xi_n + \frac{1}{n+1} \mathbf{e}_j, \tag{A.2}$$

and $Q_\theta(\cdot)$ is defined just prior to (5). By simple algebra,

$$\begin{aligned} Q_\theta(\xi_{n+1}^1) - Q_\theta(\xi_{n+1}^j) &= \frac{2n\beta^{-2}}{(n+1)^2} a_1 a_j (d_1 - d_j)^2 (\xi_{1,n} - \xi_{j,n}) \\ &\quad + \frac{2n\beta^{-2}}{(n+1)^2} \sum_{i=j+2; i \neq j}^K a_i \xi_{i,n} [a_1 (d_1 - d_i)^2 - a_j (d_j - d_i)^2] \\ &= \frac{2n\beta^{-2}}{(n+1)^2} \sum_{i=1}^K a_i \xi_{i,n} [a_1 (d_1 - d_i)^2 - a_j (d_j - d_i)^2] \\ &:= \frac{2n\beta^{-2}}{(n+1)^2} l_j(\xi_n). \end{aligned} \tag{A.3}$$

The linear function $l_j(\xi_n)$ converges to a negative limit as ξ_n tends to \mathbf{e}_1 , i.e.,

$$\lim_{\xi_n \rightarrow \mathbf{e}_1} l_j(\xi_n) = -a_1 a_j (d_1 - d_j)^2 < 0 \quad \text{for } j = 2, 3, \dots, K.$$

Thus there exists $0 \leq \eta_j(\theta) < 1$ such that for $\xi_{1,n} > 1 - \eta_j(\theta)$ we have

$$Q_\theta(\xi_{n+1}^1) < Q_\theta(\xi_{n+1}^j).$$

Let $\eta = \sup\{\eta_j(\theta) : \theta \in \Theta, 2 \leq j \leq K\}$. Because Θ is compact we have $0 < \eta < 1$. Now if $\xi_{1,n} > 1 - \eta$ then $\xi_{1,n+1} = n\xi_{1,n}/(n+1)$. Thus $\xi_{1,n+1} - \xi_{1,n} = \xi_{1,n}/(n+1) > (1 - \eta)/(n+1)$. Because the harmonic series $\sum n^{-1}$ is divergent, we have $\xi_{1,n} < 1 - \eta$ infinitely often. Choose $N_{1,\eta}$ such that $N_{1,\eta}^{-1} < \eta/2$. Let $N_\eta = \inf\{n : n > N_{1,\eta} \text{ and } \xi_{1,n} < 1 - \eta\}$. Then for $n > N_\eta$ we have

$$\xi_{1,n} = \begin{cases} (1 - n^{-1})\xi_{1,n-1} + n^{-1} < 1 - \eta/2 & \text{if } 1 - \eta < \xi_{1,n-1} < 1 - \eta/2, \\ (1 - n^{-1})\xi_{1,n-1} < 1 - \eta/2 & \text{if } \xi_{1,n-1} < 1 - \eta. \end{cases}$$

Clearly, the constraint bounds the design vector away from any of the vertices $\mathbf{e}_{K^*+1}, \dots, \mathbf{e}_K$, where K^* is defined in (9). To see that the design vector is bounded away from the vertices $\mathbf{e}_1, \dots, \mathbf{e}_{K^*}$ one needs to only consider $j \in \{1, 2, \dots, K^*\}$ in the above proof. Thus, (10) continues to hold even when the iterations in the design space are performed in conjunction with the constraint (8). \square

Proof of Corollary 1. By continuity, there exist a neighborhood U^* of θ_0 such that $\mu_R(\theta) < d_i, i = 1, \dots, K^*$ and $\mu_R(\theta) \geq d_i, i = K^* + 1, \dots, K$, for every $\theta \in U^*$. By Theorem 1, there exists N^* , possibly depending on ε , where ε is defined in (8), such that for $n \geq N^*$ the posterior probability $\pi(U^* | \mathcal{D}_n)$ is greater than $1 - \varepsilon/2$. Hence, by the constraint (8), for $n \geq N^*$, the allocation will be restricted to d_1, \dots, d_{K^*} . \square

To prove Theorem 2, we require the following lemma.

Lemma 3. Let $\{s_n\}$ be a real sequence and let $s_n \leq M_2 < \infty$ for all n . Let $A_n = s_{n+1} - s_n$. If $\sum_n A_n I(A_n < 0) > -\infty$ then $\lim_{n \rightarrow \infty} s_n = s$ for some $s \in (-\infty, \infty)$.

Proof. Let $M_3 = \sum_n \Delta_n I (\Delta_n < 0)$. Then $s_n = \sum_{k=2}^n \Delta_k + s_1 > M_3 + s_1$. Therefore, $\{s_n\}$ is a bounded sequence, and, hence have accumulation points. Suppose $\lim_{n \rightarrow \infty} s_n$ does not exist. Then s_n has at least two distinct accumulation points. Let $t_1 < t_2$ be two accumulation points of $\{s_n\}$. Let for each n, m_n denote the number of crossings from $[t_2 - (t_2 - t_1)/4, \infty)$ to $(-\infty, t_1 + (t_2 - t_1)/4]$. Then, $\sum_{k=1}^n \Delta_k I (\Delta_k < 0) < -m_n(t_2 - t_1)/2$. Because t_1 and t_2 are accumulation points, $\lim_{n \rightarrow \infty} m_n = \infty$. Hence, $\lim_{n \rightarrow \infty} \sum_{k=1}^n \Delta_k I (\Delta_k < 0) = -\infty$, and we have a contradiction. \square

Proof of Theorem 2. Define the partition of \mathbb{N} , the set of natural numbers, $\mathbb{N} = I \cup J$, where $I = \{i : Q(\xi_i) - Q(\xi_{i-1}) \geq 0\}$ and $J = I^c$ in \mathbb{N} . Because $Q(\cdot)$ is bounded above, by Lemma 3 it is then enough to show that the sequence $\{Q(\xi_{j+1}) - Q(\xi_j)\}_{j \in J}$ is summable. Let $n \in J$. Let $\xi_{n+1}^l = (1 - \lambda_n)\xi_n + \lambda_n \mathbf{e}_l, l = 1, 2, \dots, K$, be the K possible candidates at the n th iteration. Let $\Delta_n^l = Q(\xi_{n+1}^l) - Q(\xi_n)$ and $\Delta_n = (\Delta_n^1, \dots, \Delta_n^K)'$. Because Q admits a bounded second mixed derivatives we have $\Delta_n = \lambda_n(I - 1' \xi) \partial Q(\xi) / \partial \xi|_{\xi=\xi_n} + O(\lambda_n^2)$, where 1 is the vector of ones and I is the identity matrix. Therefore $\sum_{l=1}^K \xi_{n,l} \Delta_n^l = O(\lambda_n^2)$. By assumption there exists a positive constant M_4 such that $Q(\xi) < M_4$ for $\xi \in \Xi$. Let $-M_4 < \Delta_n^{(1)} \leq \Delta_n^{(2)} \leq \dots \leq \Delta_n^{(K)} < 0$ be the ordered values of $\Delta_n^l, l = 1, 2, \dots, K$. Also $\max\{\xi_{n,i} : 1 \leq i \leq K\} > (K + 1)^{-1}$ for all n . Let i' be the index of the ordered Δ_n^l corresponding to $\max\{\xi_{n,i} : 1 \leq i \leq K\}$. Then $|Q(\xi_{n+1}) - Q(\xi_n)| = |\Delta_n^{(i')}| \leq |\Delta_n^{(i')}|$ so that $|Q(\xi_{n+1}) - Q(\xi_n)| < (K + 1) |\sum_{l=1}^K \xi_{n,l} \Delta_n^l| = O(\lambda_n^2)$. Because $\{\lambda_n^2\}_{n \in J}$ is a subsequence of the summable sequence $\{\lambda_n^2\}_{j \in \mathbb{N}}$ we have

$$\lim_{n \rightarrow \infty} Q(\xi_n) = Q(\xi^*).$$

Clearly, by continuity of $Q(\cdot), d(B_\varepsilon(\xi^*), A(\xi^*)) \rightarrow 0$ as $\varepsilon \rightarrow 0$, where $d(\cdot, \cdot)$ is a set distance and $B_\varepsilon(\xi^*) = \{\xi \in \Xi : |Q(\xi) - Q(\xi^*)| < \varepsilon\}$. Hence the result follows. Suppose we define connectivity on the set $A(\xi^*)$ as follows: two points ξ_1 and ξ_2 are connected if there exists a path $P_\xi : \xi_1 \rightsquigarrow \xi_2$ from ξ_1 to ξ_2 such that $Q(\xi) = Q(\xi_1)$ for all $\xi \in P_\xi$. Then the sequence of iterations must converge to a connected subset of $A(\xi^*)$. \square

Proof of Proposition 2. The observed information can be written as

$$\begin{aligned} M_n^*(\theta) &= \sum_{i=1}^K \xi_{ni} \left[\frac{(\hat{F}(z_i) - F(z_i))\beta^{-2}}{F(z_i)(1 - F(z_i))} \right] \begin{pmatrix} f'(z_i) & f(z_i) + z_i f'(z_i) \\ f(z_i) + z_i f'(z_i) & 2z_i f(z_i) + z_i^2 f'(z_i) \end{pmatrix} \\ &\quad + \sum_{i=1}^K \xi_{ni} \left[\frac{\hat{F}(z_i)}{F^2(z_i)} + \frac{(1 - \hat{F}(z_i))}{(1 - F(z_i))^2} \right] (\beta^{-2} f^2(z_i)) \begin{pmatrix} 1 & z_i \\ z_i & z_i^2 \end{pmatrix} \\ &:= B_n(\theta) + D_n(\theta), \end{aligned} \tag{A.4}$$

where $z_i = (d_i - \alpha)/\beta, \hat{F}(z_i) = T_i(n)/N_i(n)$ is the observed proportion of 1's at dose level d_i and $T_i(n)$ and $N_i(n)$ are the number of 1's and total number of allocations to design point d_i , respectively. We show that $B_n(\theta_0) \rightarrow 0$ a.s. $[P_{\theta_0}]$. It suffices to show that for each $i, \xi_{ni}(\hat{F}(z_i) - F(z_i)) \rightarrow 0$, a.s. $[P_{\theta_0}]$. Suppose ξ_{ni} does not go to zero almost surely for some design point d_i . Then, $N_i(n) \rightarrow \infty$ almost surely. Because, given the design point d_i , the observations are i.i.d. Bernoulli($F(z_i)$), by the strong law, we have $(\hat{F}(z_i) - F(z_i))$ tending to zero almost surely. Hence, $B_n(\theta_0) \rightarrow 0$ almost surely. The eigenvalues of $B_n(\theta)$ and $D_n(\theta)$ are continuous functions of θ . Thus, for large n we can find a neighborhood of θ_0 where $B_n(\theta)$ is arbitrarily small and the eigenvalues of $B_n(\theta) + D_n(\theta)$ are uniformly close to those of $D_n(\theta)$. Specifically, given any $\varepsilon > 0$, we can find a $\delta > 0$, such that $|\lambda_{\max}(M_n^*(\theta)) - \lambda_{\max}(D_n(\theta))| < \varepsilon$ and $|\lambda_{\min}(M_n^*(\theta)) - \lambda_{\min}(D_n(\theta))| < \varepsilon$ for all $\theta \in N_\delta(\theta_0)$. Thus, it is enough to show that the eigenvalues of $D_n(\theta)$ are bounded and bounded away from zero. Now, $D_n(\theta)$ is a convex combination of finitely many rank one matrices, i.e., $D_n(\theta) = \sum_{i=1}^K \xi_{ni} c_{ni}(\theta) A_i$, where A_i are the rank one matrices $(1, z_i)'(1, z_i)$ and

$$c_{ni}(\theta) = \left[\frac{\hat{F}(z_i)}{F^2(z_i)} + \frac{(1 - \hat{F}(z_i))}{(1 - F(z_i))^2} \right] (\beta^{-2} f^2(z_i)).$$

The nonzero eigenvalue of $c_{ni}(\theta) A_i$ is $c_{ni}(\theta)(1 + z_i^2)$. Thus, the largest eigenvalue of $D_n(\theta)$ satisfies

$$\lambda_{\max}(D_n(\theta)) \leq \max_{1 \leq i \leq K} c_{ni}(\theta)(1 + z_i^2).$$

Since Θ is compact it follows that $\lambda_{\max,n} \leq L_{\max} < \infty$ for some constant L_{\max} . The matrix $D_n(\theta)$ is singular if and only if one of the allocation proportions is one and the rest are zero. Thus, the minimum eigenvalue, $\lambda_{\min}(D_n(\theta))$ of $D_n(\theta)$ is a continuous nonnegative function of the design vector ξ_n over the compact set Ξ taking zero value only at the vertices \mathbf{e}_j of the simplex. By Proposition 1, the vectors ξ_n are bounded away from the vertices. Thus, there exists a constant $L_{\min,1}$ such that $0 < L_{\min,1} < \lambda_{\min}(D_n(\theta))$ for all n . \square

Proof of Theorem 3. Observe that, with $\pi_n(u) = e^{w_n(u)} / \int_{\mathbb{R}^2} e^{w_n(t)} dt$ we have

$$\begin{aligned} \int |\pi_n^*(u) - \pi_n(u)| du &\leq \int \left| \frac{e^{w_n(u)} \pi(\hat{\theta}_n + (nM_n^*)^{-1/2}u)}{\int e^{w_n(t)} \pi(\hat{\theta}_n + (nM_n^*)^{-1/2}u) dt} - \frac{e^{w_n(u)} \pi(\hat{\theta}_n + (nM_n^*)^{-1/2}u)}{\int e^{w_n(t)} \pi(\hat{\theta}_n) dt} \right| du \\ &\quad + \int \left| \frac{e^{w_n(u)} \pi(\hat{\theta}_n + (nM_n^*)^{-1/2}u)}{\int e^{w_n(t)} \pi(\hat{\theta}_n) dt} - \frac{e^{w_n(u)} \pi(\hat{\theta}_n)}{\int e^{w_n(t)} \pi(\hat{\theta}_n) dt} \right| du \\ &\leq 2 \frac{\int e^{w_n(u)} |\pi(\hat{\theta}_n) - \pi(\hat{\theta}_n + (nM_n^*)^{-1/2}u)| du}{\int e^{w_n(u)} \pi(\hat{\theta}_n) du}. \end{aligned}$$

Since $\hat{\theta}_n \rightarrow \theta_0$ a.s. $[P(\theta_0)]$ and $\pi(\theta)$ is continuous and positive at θ_0 , the left-hand side of the above display goes to zero. Thus, we may replace $\pi_n^*(u)$ by $\pi_n(u)$ in (17). Let $f_n(u) = \exp(w_n(u))$. Consider the ratio $[\int_{\|u\| > \delta\sqrt{n}} f_n(u) du / \int_{\mathbb{R}^2} f_n(u) du]$. The ratio is the posterior probability $P(\theta \notin N_{\delta,n}(\hat{\theta}_n) | \mathcal{D}_n)$, where $N_{\delta,n}(\theta) = \{\theta' : (\theta' - \theta)' M_n^* (\theta' - \theta) \leq \delta^2\}$. Since $\hat{\theta}_n \rightarrow \theta_0$ a.s. $[P(\theta_0)]$, there exists $\delta' > 0$ such that for all sufficiently large n , $N_{\delta',n}(\theta_0) \subseteq N_{\delta,n}(\hat{\theta}_n)$ a.s. $[P(\theta_0)]$. By Proposition 2, there exists $\delta'' > 0$, such that $N_{\delta'',n}(\theta_0) \subseteq N_{\delta',n}(\theta_0)$, where, as before, $N_{\delta'',n}(\theta_0) = \{\theta : (\theta - \theta_0)'(\theta - \theta_0) \leq (\delta'')^2\}$. Therefore,

$$P(\theta \notin N_{\delta,n}(\hat{\theta}_n) | \mathcal{D}_n) \leq P(\theta \notin N_{\delta',n}(\theta_0) | \mathcal{D}_n) \leq P(\theta \notin N_{\delta'',n}(\theta_0) | \mathcal{D}_n).$$

The uniform prior is proper because of the compactness of the parameter space. Thus, by Theorem 1, $P(\theta \notin N_{\delta'',n}(\theta_0) | \mathcal{D}_n) \rightarrow 0$. Therefore we have

$$\frac{\int_{\|u\| \leq \delta\sqrt{n}} f_n(u) du}{\int_{\mathbb{R}^2} f_n(u) du} \rightarrow 1. \tag{A.5}$$

We can choose δ such that for large enough n , the region $\|u\| \leq \delta\sqrt{n}$ is contained in a neighborhood of θ_0 where the derivatives of the log-likelihood with respect to elements of θ are uniformly bounded, i.e., there exists M_5 such that if $\theta \in N_{\delta,n}(\hat{\theta})$ then

$$n^{-1} \left| \frac{\partial^3 l_n(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| < M_5, \quad i, j, k = 1, 2,$$

and $\theta_1 = \alpha$ and $\theta_2 = \beta$. Also, we can choose δ such that, by Proposition 2, $\|(\theta - \hat{\theta})\| \leq n^{-1/2} L_{\min}^{-1/2} \|u\|$ for all $\theta \in N_{\delta,n}(\hat{\theta})$. Then, for $\theta \in N_{\delta,n}(\hat{\theta})$ we have

$$|R_n(u)| \leq 2M_5 L_{\min}^{-3/2} n^{-1/2} \|u\|^3 \leq 2M_5 L_{\min}^{-3/2} \delta (u'u).$$

We can choose δ such that $2M_5 L_{\min}^{-3/2} \delta < 1/2$. Then

$$\begin{aligned} \int_{\|u\| \leq \delta\sqrt{n}} \exp\left\{-\frac{1}{2}u'u + R_n(u)\right\} du &\leq \int \exp\left\{-\left(\frac{1}{2} - 2M_5 L_{\min}^{-3/2} \delta\right)u'u\right\} du \\ &= 2\pi(1 - 4M_5 L_{\min}^{-3/2} \delta)^{-1} < \infty. \end{aligned}$$

Thus the sequence of functions $I(\|u\| \leq \delta\sqrt{n}) \exp\{-\frac{1}{2}u'u + R_n(u)\}$ are uniformly integrable and hence

$$\int_{\|u\| \leq \delta\sqrt{n}} \exp\left\{-\frac{1}{2}u'u + R_n(u)\right\} du \rightarrow \int \exp\left\{-\frac{1}{2}u'u\right\} du = 2\pi. \tag{A.6}$$

Combining (5) and (6) we have that $\int f_n(u) du \rightarrow 2\pi$. Therefore $\pi_n(u) \rightarrow \phi(u)$ a.s. $[P(\theta_0)]$. Thus, (17) follows from Scheffe's theorem. \square

Next we prove the asymptotic normality of the MLE $\hat{\theta}_n$ which is required in the proof of asymptotic normality of the Bayes estimator (Corollary 2). The proof of the theorem for asymptotic normality of the MLE (Theorem 4) relies on the following proposition, which is a straightforward generalization of Theorem 5.7 in van der Vaart (1998, p. 45). The proof of the proposition follows immediately from the proof of Theorem 5.7 of van der Vaart (1998, p. 46) and is omitted.

Let Θ be a metric space with a metric d and let θ_0 be an interior point of Θ .

Proposition 6. Let G_n and H_n be random functions of θ such that for every $\varepsilon > 0$ the following conditions hold:

1. $\sup\{|G_n(\theta) - H_n(\theta)| : \theta \in \Theta\} \xrightarrow{P} 0$,
 2. there exists a $\delta > 0$, such that $\sup\{H_n(\theta) : d(\theta, \theta_0) > \delta\} < H_n(\theta_0)$ for each $n \geq 1$.
- Then any sequence of estimators $\hat{\theta}_n$ with $G_n(\hat{\theta}_n) \geq G_n(\theta_0) - o_p(1)$ converges in probability to θ_0 .

The generalization allows one to consider sequence of random functions H_n instead of a fixed deterministic function of θ . This is possible only for sequences of function satisfying Condition 2 which is quite restrictive in the sense that it requires all member of the sequence to have a fixed maximum, θ_0 . However, in our application Condition 2 is satisfied.

Theorem 4. *Suppose the data are obtained following the sequential scheme where conditional on the first r observations, the $(r + 1)$ th design point is chosen as (4) and the corresponding response arises as (1). Let $\hat{\theta}_n$ be the MLE of θ . Then $\hat{\theta}_n$ is consistent at θ_0 and $(nM_n^*(\theta_0))^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow N(0, I)$ in law.*

Proof. We will first show that the MLE is consistent. The asymptotic normality will follow almost immediately from the consistency result. To prove consistency of the MLE, we will verify the conditions of Proposition 6. Note that $\hat{\theta}_n = \arg \max\{G_n(\theta) : \theta \in \Theta\}$, where

$$G_n(\theta) = n^{-1} \sum_{i=1}^n \left[Y_i \log \frac{F_\theta(X_i)}{F_{\theta_0}(X_i)} + (1 - Y_i) \log \frac{(1 - F_\theta(X_i))}{(1 - F_{\theta_0}(X_i))} \right].$$

Let

$$\begin{aligned} H_n(\theta) &= n^{-1} \sum_{i=1}^n \left[E_{\theta_0}(Y_i | \mathcal{D}_{i-1}) \log \frac{F_\theta(X_i)}{F_{\theta_0}(X_i)} + (1 - E_{\theta_0}(Y_i | \mathcal{D}_{i-1})) \log \frac{(1 - F_\theta(X_i))}{(1 - F_{\theta_0}(X_i))} \right] \\ &= n^{-1} \sum_{i=1}^n \left[F_{\theta_0}(X_i) \log \frac{F_\theta(X_i)}{F_{\theta_0}(X_i)} + (1 - F_{\theta_0}(X_i)) \log \frac{(1 - F_\theta(X_i))}{(1 - F_{\theta_0}(X_i))} \right] \\ &= - \sum_{i=1}^K \xi_{n,i} \text{KL}(\theta_0, \theta; d_i), \end{aligned}$$

where $\text{KL}(\theta_0, \theta; d_i)$ are the Kullback–Leibler divergence measures between the Bernoulli distributions with means $F_{\theta_0}(d_i)$ and $F_\theta(d_i)$, respectively, and $\xi_{n,i}$ is the observed proportion of allocation at dose level d_i . By the property of the Kullback–Leibler distance, Condition 2 of Proposition 6 is satisfied for each function $-\text{KL}(\theta_0, \theta; d_i)$ and hence for any finite convex combination $-\sum_{i=1}^K \xi_{n,i} \text{KL}(\theta_0, \theta; d_i)$. Let $Z_n(\theta)$ be the stochastic process (indexed by θ) defined as

$$Z_n(\theta) := \sqrt{n}(G_n(\theta) - H_n(\theta)) = n^{-1/2} \sum_{i=1}^n (Y_i - F_{\theta_0}(X_i))R(\theta, \theta_0; X_i), \tag{A.7}$$

where

$$R(\theta, \theta_0; X_i) = \log \frac{F_\theta(X_i)/(1 - F_\theta(X_i))}{F_{\theta_0}(X_i)/(1 - F_{\theta_0}(X_i))}$$

is the log odd-ratio between $F_\theta(X_i)$ and $F_{\theta_0}(X_i)$. Condition 1 requires that $\sup\{n^{-1/2}|Z_n(\theta)| : \theta \in \Theta\} \xrightarrow{P} 0$. We will apply Theorem 2.2.4 of [van der Vaart and Wellner \(1996\)](#) to verify this assertion. Since $0 < F_\theta(x) < 1$ for all θ and x and Θ is compact, Taylor’s expansion shows that there exists a constant L_1 such that

$$R(\theta_1, \theta_2; d) \leq L_1(F_{\theta_1}(d) - F_{\theta_2}(d))$$

for all $\theta_1, \theta_2 \in \Theta$ and for all $d \in \Omega$. Another Taylor’s expansion then shows

$$R(\theta_1, \theta_2; d) \leq L_1(F_{\theta_1}(d) - F_{\theta_2}(d)) \leq \sqrt{2}L_1M_1\|\theta_1 - \theta_2\| := C_1\|\theta_1 - \theta_2\|$$

for any $d \in \Omega$ where M_1 is defined in Theorem 1. Fix $\theta_1, \theta_2 \in \Theta$. Define

$$V_i := (Y_i - F_{\theta_0}(X_i))R(\theta_1, \theta_2; X_i), \quad i = 1, 2, \dots, n.$$

Then,

$$E(V_i^4 | \mathcal{D}_{i-1}) \leq C_2 C_1^4 \|\theta_1 - \theta_2\|^4 := C_3 \|\theta_1 - \theta_2\|^4, \tag{A.8}$$

where $C_2 = \max\{E(W_j - F_{\theta_0}(d_j))^4 : 1 \leq j \leq K\}$ for $W_j \sim \text{Bernoulli}(F_{\theta_0}(d_j))$, $j = 1, 2, \dots, K$. Also, $\{V_i\}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{D}_{i-1}\}$. Hence, by Burkholder’s inequality (cf. [Hall and Heyde, 1980](#), p. 23) and by the fact that $(\sum_{i=1}^n a_i)^2 \leq n \sum_{i=1}^n a_i^2$ for any a_1, \dots, a_n , we have for some absolute constant A ,

$$AE \left(\sum_{i=1}^n V_i \right)^4 \leq E \left(\sum_{i=1}^n V_i^2 \right)^2 \leq n \sum_{i=1}^n E(V_i^4).$$

By (A.8), we have $E(\sum_{i=1}^n V_i)^4 \leq n^2 C_4 \|\theta_1 - \theta_2\|^4$ for some constant C_4 . Let $\psi(x) = x^4$. Therefore, the Orlicz norm $\|\cdot\|_\psi$ coincides with the L_4 -norm, and we have

$$\|Z_n(\theta_1) - Z_n(\theta_2)\|_\psi = \left(E \left(n^{-1/2} \sum_{i=1}^n V_i \right)^4 \right)^{1/4} \leq C_4^{1/4} \|\theta_1 - \theta_2\|.$$

By Theorem 2.2.4 of van der Vaart and Wellner (1996) and the comment following the proof and the fact that $Z_n(\theta_0) \equiv 0$ for all n , we have

$$\left\| \sup_{\theta \in \Theta} |Z_n(\theta)| \right\|_\psi \leq K_1 \int_0^{\text{diam } \Theta} \psi^{-1}(D(\varepsilon, \|\cdot\|, \Theta)) d\varepsilon,$$

where K_1 is a constant depending only on ψ and C_3 , $\psi^{-1}(x) = x^{1/4}$ and $D(\varepsilon, \|\cdot\|, \Theta)$ is the packing number of Θ . Since Θ is a compact subset of \mathbb{R}^2 , there exist a constant L_3 such that $D(\varepsilon, \|\cdot\|, \Theta) \leq L_3 \varepsilon^{-2}$. Thus,

$$\left\| \sup_{\theta \in \Theta} |Z_n(\theta)| \right\|_\psi \leq K_1 L_3 \int_0^{\text{diam } \Theta} \varepsilon^{-1/2} d\varepsilon < \infty.$$

Then, by Markov's inequality we have

$$\sup_{\theta \in \Theta} |G_n(\theta) - H_n(\theta)| = \sup_{\theta \in \Theta} n^{-1/2} |Z_n(\theta)| = O_p(n^{-1/2}).$$

By definition, $G_n(\hat{\theta}_n) \geq G_n(\theta_0)$. Hence, by Proposition 6 we have consistency of the MLE. By Taylor's expansion of the derivative of the log-likelihood function at the MLE, we have

$$0 = \sum_{i=1}^n u_i(\theta_0) - nM_n^*(\theta_0)(\hat{\theta}_n - \theta_0) + O_p(\|\hat{\theta}_n - \theta_0\|), \quad (\text{A.9})$$

where $u_i(\theta_0) = \partial/\partial\theta [L_i(\theta) - L_{i-1}(\theta)]$ and $L_i(\theta)$ is the log-likelihood of the first i observations. Then, by the fact that $\|\hat{\theta}_n - \theta_0\| = o_p(1)$ and that $(nM_n^*(\theta_0))^{-1/2} \sum_{i=1}^n u_i(\theta_0) \rightarrow N(0, I)$ (cf. Hall and Heyde, 1980, Section 6.2) we have the result. \square

References

- Abdelbasit, K.M., Plackett, R.L., 1983. Experimental designs for binary data. *J. Amer. Statist. Assoc.* 78, 90–98.
- Babb, J., Rogatko, A., Zacks, S., 1998. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statist. Med.* 17, 1103–1120.
- Bickel, P., Yahav, J., 1969. Some contributions to the asymptotic theory of Bayes solutions. *Z. Wahrsch. Verw. Gebiete* 11, 257–275.
- Box, G.E.P., Hunter, W.G., 1965. Sequential design for experiments for nonlinear models. In: *Proceedings of the IBM Scientific Computing Symposium on Statistics*, October 21–23, 1963, pp. 113–137.
- Chaloner, K., 1989. Bayesian design for estimating the turning point of a quadratic regression. *Comm. Statist. Theory Methods* 18, 1385–1400.
- Chaudhuri, P., Mykland, P.A., 1993. Nonlinear experiments: optimal design and inference based on likelihood. *J. Amer. Statist. Assoc.* 88, 538–546.
- Chaudhuri, P., Mykland, P.A., 1995. On efficient designing of nonlinear experiments. *Statist. Sinica* 5, 421–440.
- Fedorov, V.V., 1972. *Theory of Optimal Experiments*. Academic Press, New York.
- Ford, I., Silvey, S.D., 1980. A sequentially constructed design for estimating a non-linear parametric function. *Biometrika* 67, 381–388.
- Ford, I., Titterton, D.M., Wu, C.F.J., 1985. Inference and sequential design. *Biometrika* 72, 545–551.
- Ford, I., Titterton, D.M., Kistos, C.P., 1989. Recent advances in nonlinear experimental design. *Technometrics* 31, 49–60.
- Ghosh, J.K., Ramamoorthi, R.V., 2003. *Bayesian Nonparametrics*. Springer, New York.
- Haines, L.M., 1998. Optimal design for neural networks. In: Flournoy, N., Rosenberger, W.F., Wong, W.K. (Eds.), *New Developments and Applications in Experimental Design*. Institute of Mathematical Statistics, Hayward, pp. 152–162.
- Haines, L.M., Perovozskaya, I., Rosenberger, W.F., 2003. Bayesian optimal designs for phase I clinical trials. *Biometrics* 59, 591–600.
- Hall, P., Heyde, C.C., 1980. *Martingale Limit Theory and its Application*. Academic Press, New York.
- Hu, F., Rosenberger, W.F., Zhang, L.-X., 2005. Asymptotically best response-adaptive randomization procedures. *J. Statist. Plann. Inference* 136, 1911–1922.
- Hu, I., 1997. Strong consistency in stochastic regression models via posterior covariance matrices. *Biometrika* 84, 744–749.
- Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problem. *Canad. J. Math.* 12, 363–366.
- Lai, T.Z., 1994. Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann. Statist.* 22, 1917–1930.
- Leonard, T., 1982. An inferential approach to the bioassay design problem. Technical Summary Report 2416, Mathematics Research Center, University of Wisconsin-Madison.
- Minkin, S., 1987. On optimal design for binary data. *J. Amer. Statist. Assoc.* 82, 1098–1103.
- O'Quigley, J., Pepe, M., Fisher, L., 1990. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 46, 33–48.
- Rosenberger, W.F., Haines, L.M., 2002. Competing designs for phase I clinical trials: a review. *Statist. Med.* 21, 2757–2770.
- Rosenberger, W.F., Canfield, G.C., Perovozskaya, I., Haines, L.M., Hausner, P., 2005. Development of interactive software for Bayesian optimal phase I clinical trial design. *Drug Inform. J.* 39, 89–98.
- Ross, S.M., 1996. *Stochastic Processes*. Wiley, New York.
- Schwartz, L., 1965. On Bayesian procedures. *Z. Wahrsch. Verw. Gebiete* 4, 10–26.
- Serfling, R., 1981. *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
- Shen, L.Z., O'Quigley, J., 1996. Consistency of continual reassessment method under model misspecification. *Biometrika* 83, 395–405.

- Silvey, S.D., 1980. Optimal Design. Chapman & Hall, London.
- Tsutakawa, R.K., 1972. Design of experiment for bioassay. *J. Amer. Statist. Assoc.* 67, 584–590.
- Tsutakawa, R.K., 1980. Selection of dose levels for estimating a percentage point of a logistic quantal response curve. *Appl. Statist.* 29, 25–33.
- van der Vaart, A., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge.
- van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes. Springer, New York.
- White, L.V., 1975. Optimal design of experiments for non-linear models. Unpublished Ph.D. Thesis, Imperial College, London.
- Whitehead, J., Brunier, H., 1995. Bayesian decision procedures for dose determining experiments. *Statist. Med.* 14, 885–893.
- Wu, C.F.J., 1985. Asymptotic inference from sequential design in a nonlinear situation. *Biometrika* 72, 553–558.
- Wynn, H., 1970. The sequential generation of D -optimal experimental designs. *Ann. Math. Statist.* 41, 1655–1664.
- Zacks, S., 1977. Problems and approaches in design of experiments for estimation and testing in non-linear problems. In: Krishnaiah, P.R. (Ed.), *Multivariate Analysis IV*. North-Holland, Amsterdam, pp. 209–223.
- Zacks, S., Rogatko, A., Babb, J., 1998. Optimal Bayesian-feasible dose escalation for cancer phase I trials. *Statist. Probab. Lett.* 38, 215–220.