The Annals of Statistics
1978, Vol. 6, No. 6, 1286–1301

SOME ALGORITHMIC ASPECTS OF THE THEORY OF OPTIMAL DESIGNS[1]

BY CHIEN-FU WU

University of Wisconsin, Madison

The approximate optimal design problem is treated as a constrained convex programming problem. A general class of optimal design algorithms is proposed from this point of view. Asymptotic convergence to optimal designs is also proved. Related problems like the implementability problem for the infinite support case and the general step-length algorithms are discussed.

1. Introduction. The approximate optimal design problem, as first formulated by Kiefer and Wolfowitz (1959), is an important approach to the optimality problem of statistical design of experiments. Except for some cases with nice response functions and design regions, it is very hard to obtain a closed form of the solution. For different problems quite different techniques have to be used. It is thus desirable to find iterative methods for obtaining optimal designs which do not rely on the special structure of each problem. A few methods have been proposed to cope with this type of problem, but they are either special for one design criterion or utilize one optimization method.

The purpose of the present work is to investigate the problem more systematically from the viewpoint of optimization theory and algorithms. The optimal design problem, in the sense of Kiefer and Wolfowitz, is reformulated as a convex programming problem with some distinct features. To cope with these distinct features, a class of optimal design algorithms is developed along a different line from the standard convex programming methods. In Section 2 both the iterative methods on a finite support and the changes of design supports are considered. These two are combined to form a general class of optimal design algorithms. Several modifications of the algorithms are also discussed. The associated proofs of convergence to optimal designs are given in Section 3. For the infinite support case, the implementability problem for the above algorithms is resolved by discretizing the support set. Related convergence results are given in Section 4. Most of the algorithms considered above involve a line search along the direction of iteration. For numerical or statistical reasons, they may be replaced by a prescribed sequence of step lengths before the execution of the algorithm. In Section 5 some general convergence results are given for these general step-length algorithms.

A linear experiment is given by

$$y = \theta^T x + \varepsilon$$

Received September 1977; revised December 1977.

[1]This research was supported by the U.S. Army under Contract No. DAAG 29-75-C-0024.

AMS 1970 *subject classifications*. Primary 62K05; Secondary 90C50.

Key words and phrases. General equivalence theorem, optimal design algorithms, nonlinear programming, general step-length algorithms.

where θ is a $k \times 1$ vector, x is from a compact set $\mathcal{X}$ of $k \times 1$ vectors and ε is a random error. Errors corresponding to different observations are assumed to be uncorrelated, with equal variances and zero means. A design ξ is defined to be a probability measure on $\mathcal{X}$ and its corresponding information matrix $M(\xi)$ is defined to be $\int_{\mathcal{X}} xx^T\xi(dx)$. Let $\mathfrak{M}$ be the collection of $M(\xi)$ for all probability measures ξ on $\mathcal{X}$. ξ^* is called Φ-optimal if it achieves

$$\inf\{\Phi(M(\xi)) : M(\xi) \in \mathfrak{M}\}. \tag{1.1}$$

Typical examples of Φ are $\Phi(M) = -\log\det M$ (D-optimality), $\Phi(M) = \operatorname{tr}(AM^{-p})$ for A positive definite and $p > 0$ (Kiefer's Φ_p-optimality or L-optimality for $p = 1$) and $\Phi(M) =$ maximum eigenvalue of M^{-1} (E-optimality).

Throughout the paper we assume that Φ *is convex and bounded below on* $\mathfrak{M}$ *and* Φ *is differentiable in a neighborhood of* $\mathfrak{M}^+$ *in the space of nonnegative definite* $k \times k$ *matrices.* $\mathfrak{M}^+ = \{M : M \in \mathfrak{M}$ and $\Phi(M) < \infty\}$. For a $k \times 1$ vector v (or a $k \times k$ matrix M), *the* l_2*-norm* $\|v\| = \sum_{i=1}^k v_i^2$ (or $\|M\| = \sum_{i,j=1}^k m_{ij}^2$) *is assumed.*

When Φ is well defined and differentiable, the $k \times k$ matrix $\nabla\Phi$ is defined as

$$(\nabla\Phi)_{ij} = \frac{\partial\Phi(M)}{\partial m_{ij}}.$$

The directional derivative

$$\begin{aligned}\frac{\partial}{\partial\alpha}\Phi((1-\alpha)M(\xi) + \alpha M(\xi'))|_{\alpha=0^+} &= \langle \nabla\Phi(M(\xi)), M(\xi' - \xi)\rangle \\ &= \int_{\mathcal{X}} x^T\nabla\Phi(M(\xi))x\xi'(dx) - \operatorname{tr}(\nabla\Phi(M(\xi))M(\xi))\end{aligned}$$

where $\langle \nabla\Phi(M), M\rangle = \operatorname{tr}(\nabla\Phi(M)M)$ and $\langle \nabla\Phi(M), xx^T\rangle = \operatorname{tr}(\nabla\Phi(M)xx^T) = x^T\nabla\Phi(M)x$ since l^2-norm is assumed.

For convenience in presenting the algorithms, we shall use the following standard notation:

$$\begin{aligned} d(x, \xi) &= -x^T\nabla\Phi(M(\xi))x, \\ \bar{d}(\xi) &= \max_{x\in x} d(x, \xi), \\ d^{\#}(\xi) &= -\operatorname{tr}(\nabla\Phi(M(\xi))M(\xi)). \end{aligned}$$

In terms of the d-notation, the celebrated *general equivalence theorem* states that: ξ^* is Φ-optimal $\Leftrightarrow \xi^*\{x : d(x, \xi^*) = \bar{d}(\xi^*)\} = 1 \Leftrightarrow \bar{d}(\xi^*) = d^{\#}(\xi^*)$. (See Kiefer, 1974).

Structurally, problem (1.1) is a constrained minimization problem over a convex set generated by $\{xx^T\}_{x\in\mathcal{X}}$. Some well-known optimization methods can certainly be used in generating useful algorithms. But in the design situation, the optimal design ξ^*, instead of its information matrix $M(\xi^*)$, is the main concern. This excludes the use of duality theory, since we want to update the design at every iteration. Another distinct feature of problem (1.1) is that $\{xx^T : x \in \mathcal{X}\}$ gives the set of extreme points of M. When the design support has to be adjusted, iterations along the directions of the vertices of $\mathfrak{M}$ seem indispensable. To cope with these novel features, a class of optimal design algorithms is suggested in the next section.

2. A class of optimal design algorithms. When $\mathcal{X}$ has finitely many points $\{x_i\}_{i=1}^n$, (1.1) can be rephrased as the following:

$$\min\{\Phi(\Sigma_{i=1}^n \lambda_i x_i x_i^T) : \Sigma_{i=1}^n \lambda_i = 1, \lambda_i \geqslant 0\}. \tag{2.1}$$

This is a typical constrained minimization problem with a nice constraint set, a simplex. We can certainly apply some well-known methods in nonlinear programming to this problem. Let ξ_0 be the initial design measure with positive components. In Wu (1978), the following class of iterative directions is considered:

$$h = (\mathbf{1}^T\Lambda\mathbf{1})(\Lambda d) - (\mathbf{1}^T\Lambda d)(\Lambda\mathbf{1}) \tag{2.2}$$

where $\mathbf{1}^T = (1, \cdots, 1)$, $d^T = (d_1, \cdots, d_n)$ with $d_i = d(x_i, \xi_0)$ and Λ is a positive definite $n \times n$ matrix which may depend on ξ_0. Some reasons for considering (2.2) are:

(I) h is a legitimate direction of iteration for probability measures, since $\Sigma_{i=1}^n h_i = (\mathbf{1}^T\Lambda\mathbf{1})(\mathbf{1}^T\Lambda d) - (\mathbf{1}^T\Lambda d)(\mathbf{1}^T\Lambda\mathbf{1}) = 0$.

(II) $(\partial/\partial\alpha)\Phi(M(\xi_0) + \alpha\Sigma_{i=1}^n h_i x_i x_i^T)|_{\alpha=0^+} = \Sigma_{i=1}^n h_i x_i^T \ \nabla\ \Phi(M(\xi_0))x_i = -[(\mathbf{1}^T\Lambda\mathbf{1})(d^T\Lambda d) - (\mathbf{1}^T\Lambda d)^2] \leqslant 0$ with equality $\Leftrightarrow d$ is proportional to $\mathbf{1}$ (since Λ is positive definite) $\Leftrightarrow \xi_0$ is Φ-optimal on $\mathcal{X}$ (from the general equivalence theorem). As long as the initial ξ_0 is not Φ-optimal, iteration along the direction h will improve over ξ_0.

The Λ matrix in (2.2) can be quite arbitrary. For different statistical or numerical reasons, different Λ's can be considered. Examples of (2.2) include the gradient projection method and its normalized version, conjugate gradient projection method and its adaptive version, Newton and quasi-Newton methods. Derivations and numerical comparisons of these methods and other popular methods in this area are made in [11].

For an arbitrary design region $\mathcal{X}$, the algorithm may start on a proper subset of $\mathcal{X}$ and change its design support when necessary. This is especially needed when $\mathcal{X}$ is infinite. The idea of the general optimal design algorithm is to combine the iterative methods (2.2) for finite design region and the methods of changing design support discussed below.

If ξ_n is not an optimal design, from the general equivalence theorem, $\{x \in \mathcal{X} : d(\xi_n, x) = \bar{d}(\xi_n)\}$ is not included in the support of ξ_n. To improve over ξ_n, a new support point y_n with a large $d(y_n, \xi_n)$ value should be introduced. That is

$$\xi_{n+1} = (1 - \alpha_n)\xi_n + \alpha_n\xi_{y_n}, \tag{2.3}$$

where ξ_{y_n} is concentrated at y_n with $d(y_n, \xi_n) - d^\#(\xi_n) \geqslant r(\bar{d}(\xi_n) - d^\#(\xi_n))$, $0 < r \leqslant 1$ and α_n is chosen to minimize $\Phi((1-\alpha)M(\xi_n) + \alpha M(\xi_{y_n}))$ over $0 \leqslant \alpha \leqslant 1$. Equation (2.3) with $r = 1$ was suggested by Fedorov (1969) and Wynn (1970). An important modification of (2.3) is due to Atwood (1973), who suggested choosing y_n with low $d(y_n, \xi_n)$ value and then allowing α_n to be negative. For design problems of moderate size, this method is highly recommended in [11].

Since all the extreme points of the convex set $\mathfrak{M}$ are of the form xx^T, algorithm (2.3) only adjusts the design measure along the direction of the vertices of $\mathfrak{M}$. Therefore, it is called the *vertex direction method*. The general idea of adjusting along the direction of the vertices was proposed by Frank and Wolfe (1956) in a different context. Due to its numerical deficiencies, the F-W method is not used too often in solving optimization problems. But in the optimal design context, in order to improve over the current design, some new points have to be introduced. The vertex direction method emerges as an indispensable tool. In fact, our choices of y_n and Atwood's choices of α_n seem to be new in optimization literature. Other distinct features of the vertex direction method are discussed in the numerically oriented paper [11].

The following optimal design algorithm combines methods (2.2) and (2.3). For design $\xi_j = (\xi_i^{(j)})_i$ and $h(\xi_j) = (h_i(\xi_j))_i$ calculated according to (2.2), a line search for $\Phi(M(\xi_j + uh(\xi_j)))$ is performed on $[0, \bar{u}]$ where

$$\bar{u} = \min\{\xi_i^{(j)}/|h_i(\xi_j)| : h_i(\xi_j) < 0\}. \tag{2.4}$$

Note that $\xi_j + uh(\xi_j)$ is still a probability measure for $0 \leqslant u \leqslant \bar{u}$. If no optimal design exists on the support of ξ_j and $\Phi(M(\xi_j))$ can not be improved very much on its current support, the effort of the above line search on $[0, \bar{u}]$ is almost futile. So when $\bar{u}\Sigma_{i=1}^n h_i(\xi_j)d(x_i, \xi_j)$, the directional derivative of Φ at $M(\xi_j)$ along $h(\xi_j)$, is small, the support of ξ_j is augmented by the Fedorov-Wynn type algorithm and then method (2.2) is repeated; otherwise, a line search on $[0, \bar{u}]$ is carried out. This is the idea behind the following general algorithm. h_j is the abbreviation for $h(\xi_j)$.

ALGORITHM 1. Choose $\varepsilon_0 > 0$, $\quad 0 < r \leqslant 1$.

Step 0. $j = 0$, choose a ξ_0 with $\Phi(M(\xi_0)) < \infty$.

Step 1. Compute h and $\bar{u}$ of ξ_j according to (2.2) and (2.4); if $\bar{u}\Sigma_{i=1}^n h_i(\xi_j)d(x_i, \xi_j) \leqslant \varepsilon_0$, go to 3; else, go to 2.

Step 2. Let $\xi_{j+1} = \xi_j + u_j h_j$ where u_j minimizes $\Phi(M(\xi_j + uh_j))$ over $0 \leqslant u \leqslant \bar{u}$, $j = j + 1$, go to 1.

Step 3. If $\bar{d}(\xi_j) = d^\sharp(\xi_j)$, stop; else, choose a y_j with $d(y_j, \xi_j) - d^\sharp(\xi_j) \geqslant r(\bar{d}(\xi_j) - d^\sharp(\xi_j))$, let $\xi_{j+1} = (1 - \alpha_j)\xi_j + \alpha_j \xi_{y_j}$ with α_j minimizing $\Phi(M((1 - \alpha)\xi_j + \alpha\xi_{y_j}))$ over $0 \leqslant \alpha \leqslant 1, j = j + 1$, go to 1.

REMARKS. 1. The one dimensional minimization in Steps 2 and 3 can be replaced by other existing line search methods. The corresponding convergence proofs can be adapted from Theorem 1 in an analogous way to similar results in nonlinear programming. Interested readers are referred to [7].

2. The r in Step 3 of Algorithm 1 can be chosen to depend on j. A sufficient condition on $\{r_j\}$ for convergence to optimal designs is given in Theorem 2.

3. Algorithm 1 can be interpreted as a sequence of minimizations over a sequence of "polygons" inscribed in the convex set $\mathfrak{M}$. Although it can be considered as a special case of the interior penalty functions method (see [7]), the

idea of using the general equivalence theorem to decide when and how to change design supports (which is equivalent to choices of penalty function) seems to be new.

A few modifications of the basic Algorithm 1 are given below. *All the dotted lines in the following modified algorithms refer to the basic Algorithm* 1.

(i) *A proper balance between Steps* 2 *and* 3. From Remark 1 after Theorem 1 in the next section, Algorithm 1 will skip Step 2 after a finite number of steps. To benefit from the numerical efficiencies of the optimization algorithms in Step 2, a proper balance between Steps 2 and 3 is needed. We can prolong the sojourn of the algorithm at Step 2 by taking ε_0 small, but the gain caused by doing Step 2 may become very insignificant. On the other hand, big ε_0 value will drive the algorithm to "permanent residence" at Step 3 too early. This dilemma is resolved by taking a sequence $\{\varepsilon_i\}_i$ in the following way.

ALGORITHM 1.1. Choose sequences $\{\varepsilon_i\}_i$ and $\{a_i\}_i$ both decreasing to zero.
Step 0. $k = 0, j = 0,$---.
Step 1. ---; if $\bar{u}\Sigma_{i=1}^n h_i(\xi_j)d(x_i, \xi_j) \leqslant \varepsilon_k$, go to 3; else go to 2.
Step 2. Same as in Algorithm 1.
Step 3. ---; else, let $a_{p+1} \leqslant \bar{d}(\xi_j) - d^\#(\xi_j) < a_p$, set $k = \max(k, p)$, choose ξ_{j+1} according to (2.3), $j = j + 1$, go to 1.

(ii) *Choices of new design points*. The main problem with Algorithm 1 is the evaluation of $\bar{d}(\xi_j)$ and the search for a new point y_j. In terms of the asymptotic convergence, it suffices to choose a y_j with its d-value not too far away from the $\bar{d}$ value. The idea is borrowed from nonlinear programming where an iterative direction, not far away from the steepest descent direction, is often used.

ALGORITHM 1.2. Let $0 < r < 1, \quad \varepsilon_1 > 0$.
Steps 0, 1, 2. Same as in Algorithm 1.
Step 3. ---; else, choose a y_j with $d(y_j, \xi_j) - d^\#(\xi_j) \geqslant \varepsilon_1$ or $r(\bar{d}(\xi_j) - d^\#(\xi_j))$, ----.

With this modification, the effort of finding y_j is substantially reduced since the choices of y_j with $d(y_j, \xi_j) - d^\#(\xi_j) \geqslant \varepsilon_1$ may be achieved in a finite number of steps.

The following special case of Step 3 in Algorithm 1.2 is of particular interest (see also Section 5).

Step 3′. ---; else, choose a y_j with

$$\frac{d(y_j, \xi_j) - d^\#(\xi_j)}{|M(\xi_{y_j} - \xi_j)|} = \max_{y \in \mathfrak{X}} \frac{d(y, \xi_j) - d^\#(\xi_j)}{|M(\xi_y - \xi_j)|}.$$

Let $\bar{d}(\xi_j) = d(\bar{y}_j, \xi_j)$ for some $\bar{y}_j \in \mathfrak{X}$. Therefore

$$d(y_j, \xi_j) - d^\#(\xi_j) \geqslant \frac{d(\bar{y}_j, \xi_j) - d^\#(\xi_j)}{|M(\xi_{\bar{y}_j} - \xi_j)|}|M(\xi_{y_j} - \xi_j)| \text{ (from the choice of } y_j)$$
$$\geqslant r_0[\bar{d}(\xi_j) - d^\#(\xi_j)]$$

where $r_0 = \inf\{|M(\xi_x - \xi)|/|M(\xi_z - \xi)|: x, z \in \mathcal{X}$ and $\Phi(M(\xi)) \leqslant \Phi(M(\xi_0)) < \infty\}$ is positive if $\Phi(M(\xi_x))$ is assumed to be infinite for all $x \in \mathcal{X}$.

(iii) *Reduction of support size*. The ultimate goal of the optimal design problem is two-fold: to find an ξ^* solving (1.1) and having the smallest possible support size. (If the experiment is designed also to detect incorrect models, the design support can not be too small. But still, designs with too large support are not desirable.) The first is what we are striving to solve in this paper. The second is a much harder problem. Our Algorithm 1 will sometimes reduce the support size if the u_j in Step 2 of Algorithm 1 is $\bar{u}$. The following modification will increase the chance of reducing the support size.

ALGORITHM 1.3.

Step 0, 2, 3. Same as in Algorithm 1.
Step 1. ---; if $\bar{u}\Sigma_{i=1}^n h_i(\xi_j)d(x_i, \xi_j) \leqslant \varepsilon_0$, go to 4; else, ---- .
Step 4. If $\Phi(M(\xi_j)) \leqslant \Phi(M(\xi_j + \bar{u}h_j))$, go to 3; else, let $\xi_{j+1} = \xi_j + \bar{u}h_j, j = j + 1$, go to 1.

This is particularly useful when $\bar{u}$ is small and $\Sigma_{i=1}^n h_i d_i$ is moderate. In this case, it is very probable that $\Phi(M(\xi_j + \bar{u}h_j))$ is smaller than $\Phi(M(\xi_j))$.

3. Convergence proofs. To simplify the notation, M_j, Φ_j, and h_j will sometimes be used for $M(\xi_j)$, $\Phi(M(\xi_j))$ and $h(\xi_j)$. $d_i^{(j)}$ and $h_i^{(j)}$ will be used for $d(x_i, \xi_j)$ and $h_i(\xi_j)$ when this does not cause any confusion. The following definition is needed in the proof of Theorem 1.

DEFINITION 1 (Armijo search, [7]). Given $0 < \alpha < 1$ and $0 < r < 1$, define β_j to be r^m with m the *first* integer such that

$$\Phi\big(M(\xi_j + r^m\bar{u}h_j)\big) - \Phi\big(M(\xi_j)\big) \leqslant \alpha r^m\bar{u}\langle \nabla \Phi(M_j), \Sigma_{i=1}^n h_i(\xi_j)x_i x_i^T\rangle$$
$$= -\alpha r^m\bar{u}\Sigma_{i=1}^n h_i(\xi_j)d(x_i, \xi_j)$$

in Step 2 of Algorithm 1 or

$$\Phi\big(M\big[(1 - r^m)\xi_j + r^m\xi_{y_j}\big]\big) - \Phi\big(M(\xi_j)\big)$$
$$\leqslant \alpha r^m\langle \nabla \Phi(M_j), M(\xi_{y_j} - \xi_j)\rangle = -\alpha r^m\big(d(y_j, \xi_j) - d^\#(\xi_j)\big),$$

in Step 3 of Algorithm 1.

$\Sigma_{i=1}^n h_i^{(j)}d_i^{(j)} \geqslant \varepsilon_0$ follows from Step 2 of Algorithm 1 and $d(y_j, \xi_j) - d^\#(\xi_j) \geqslant r(\bar{d}(\xi_j) - d^\#(\xi_j)) > 0$ follows from Step 3 of Algorithm 1. This, together with $0 < \alpha < 1$ and the definition of tangent, implies that the β_j in Definition 1 is well defined and positive.

Define F_K to be the compact set $\{M : M \in \mathfrak{M}, \Phi(M) \leqslant K < \infty\}$. Uniform continuity and continuity on F_K are thus equivalent.

THEOREM 1. *Suppose $\nabla \Phi$ exists and is continuous on $F_{\Phi(M(\xi_0))}$. Let $\{\xi_j\}$ be the sequence constructed by Algorithm 1. Then $\Phi(M_j)$ converges monotonically to the optimal value Φ^* (the Φ-value of an optimal design).*

PROOF. If $\{\xi_j\}$ stops at ξ_k, $\bar{d}(\xi_k) = d^{\#}(\xi_k)$ and ξ_k is optimal. Otherwise, there exists an infinite sequence $\{\xi_j\}$ with $\Phi(M(\xi_j))$ monotonically decreasing to Φ'; monotonicity is clear from the nature of line minimization and the choice of iterative directions. If Φ' is not the optimal value, from the compactness of $\mathfrak{M}$, there exists an infinite subsequence $\{\xi_{n_j}\}$ which converges to a nonoptimal ξ'. From the general equivalence theorem, there exists an $a > 0$ and an i_0 such that

$$\bar{d}(\xi_{n_j}) - d^{\#}(\xi_{n_j}) \geqslant a \qquad \text{for all} \quad n_j \geqslant i_0.$$

(i) Let β_j be defined as in Definition 1. In Step 2 of Algorithm 1,

$$\begin{aligned}\Phi(M_{j+1}) - \Phi(M_j) &\leqslant \Phi(M(\xi_j + \beta_j \bar{u} h_j)) - \Phi(M(\xi_j)) \\ &\leqslant -\alpha\beta_j\bar{u}\Sigma_{i=1}^n h_i(\xi_j) d(x_i, \xi_j) \leqslant -\alpha\beta_j\varepsilon_0.\end{aligned} \tag{3.1}$$

In Step 3 of Algorithm 1,

$$\begin{aligned}\Phi(M_{j+1}) - \Phi(M_j) &\leqslant \Phi(M[(1-\beta_j)\xi_j + \beta_j\xi_{y_j}]) - \Phi(M(\xi_j)) \\ &\leqslant -\alpha\beta_j(d(y_j, \xi_j) - d^{\#}(\xi_j)) \leqslant -\alpha\beta_j r(\bar{d}(\xi_j) - d^{\#}(\xi_j)) \\ &\leqslant -\alpha\beta_j r a \qquad \text{for all} \quad j = n_k \geqslant i_0.\end{aligned} \tag{3.2}$$

The first inequality in both steps follows from the fact that u_j is the line minimizer. It remains to prove that β_j (or β_{n_j}) is bounded away from 0 for j large. This will imply that $\{\Phi_j\}$ is not a Cauchy sequence, contradicting $\Phi_j \searrow \Phi'$.

(ii) In Step 2,

$$\begin{aligned}&\Phi(M(\xi_j + \beta\bar{u}h_j)) - \Phi(M(\xi_j)) - (-\alpha\beta\bar{u}\Sigma_{i=1}^n h_i^{(j)} d_i^{(j)}) \\ &\quad = \langle \nabla\Phi(M(\xi_j + \lambda\beta\bar{u}h_j)) - \nabla\Phi(M(\xi_j)), \beta\bar{u}\Sigma_{i=1}^n h_i(\xi_j) x_i x_i^T\rangle \\ &\quad\quad + \langle \nabla\Phi(M_j), \beta\bar{u}\Sigma_{i=1}^n h_i(\xi_j) x_i x_i^T\rangle + \alpha\beta\bar{u}\Sigma_{i=1}^n h_i^{(j)} d_i^{(j)} \\ &\quad\quad\quad 0 \leqslant \lambda \leqslant 1 \text{ from the mean value theorem} \\ &\quad = \beta[\langle \nabla\Phi(M(\xi_j + \lambda\beta\bar{u}h_j)) - \nabla\Phi(M(\xi_j)), \bar{u}M(h_j)\rangle \\ &\quad\quad - (1-\alpha)\bar{u}\Sigma_{i=1}^n h_i^{(j)} d_i^{(j)}].\end{aligned} \tag{3.3}$$

Inside the bracket, the second term $\geqslant (1-\alpha)\varepsilon_0$. Since both $M(\xi_j)$ and $M(\xi_j + \bar{u}h_j)$ are in $\mathfrak{M}$, from the compactness of $\mathfrak{M}$, $|\bar{u}M(h_j)|$ is uniformly bounded by a finite constant M. We can therefore find a $\bar{\beta}$ such that for $0 \leqslant \beta \leqslant \bar{\beta}$

$$M(\xi_j + \beta\bar{u}h_j) \in F_{\Phi(M(\xi_0))}$$

and

$$|\nabla\Phi(M(\xi_j + \beta\bar{u}h_j)) - \nabla\Phi(M(\xi_j))| \leqslant (1-\alpha)\varepsilon_0/M.$$

Here the uniform continuity of $\nabla\Phi$ on $F_{\Phi(M_0)}$ is used. Therefore the first term inside the bracket is $\leqslant (1-\alpha)\varepsilon_0$ and the whole expression in (3.3) is $\leqslant 0$ for $0 \leqslant \beta \leqslant \bar{\beta}$ and all j.

In Step 3,

$$\begin{aligned}
&\Phi\big(M\big[(1-\beta)\xi_j+\beta\xi_{y_j}\big]\big)-\Phi\big(M(\xi_j)\big)-\big(-\alpha\beta\big[d(y_j,\xi_j)-d^{\#}(\xi_j)\big]\big)\\
(3.4)\qquad &=\beta\Big[\big\langle\nabla\Phi\big(M\big[(1-\lambda\beta)\xi_j+\lambda\beta\xi_{y_j}\big]\big)-\nabla\Phi\big(M(\xi_j)\big),M(\xi_{y_j}-\xi_j)\big\rangle\\
&\quad-(1-\alpha)\big(d(y_j,\xi_j)-d^{\#}(\xi_j)\big)\Big],
\end{aligned}$$

$0 \leqslant \lambda \leqslant 1$ from the mean value theorem. Inside the bracket, the second term $\geqslant (1-\alpha)ra$ for all $j = n_k \geqslant i_0$. As argued above, we can find a $\bar{\beta}$ such that the first term inside the bracket is $\leqslant (1-\alpha)ra$ for all $0 \leqslant \beta \leqslant \bar{\beta}$ and all $j = n_k \geqslant i_0$. Therefore, the whole expression in (3.4) is $\leqslant 0$ for $0 \leqslant \beta \leqslant \bar{\beta}$ and all $j = n_k \geqslant i_0$.

Let j_0 be the first m such that $\bar{\beta} \geqslant r^m$; r is from Definition 1. Therefore,

$$\begin{aligned}
(3.5)\qquad \beta_j &\geqslant r^{j_0} > 0 \quad \text{for all} \quad j \quad \text{in Step 2}\\
&> 0 \quad \text{for all} \quad j = n_k \geqslant i_0 \quad \text{in Step 3.}
\end{aligned}$$

(i) and (ii) give the desired result. □

REMARK 1. If Step 2 is carried out at the jth iteration, we have, from (3.1) and (3.5), $\Phi(M_{j+1}) - \Phi(M_j) \leqslant -\alpha\beta^{j_0}\varepsilon_0$. This can not hold for infinitely many j's, since $\{\Phi(M_j)\}$ is Cauchy. Therefore, Step 2 will be skipped eventually. Algorithm 1.1 was designed to cope with this problem.

2. The proof of Theorem 1 also applies to a modified version of the basic Algorithm 1; namely, to replace the line minimization in Steps 2 and 3 of Algorithm 1 by the Armijo search method of Definition 1.

3. The assumptions of Theorems 1 and 2 are met by the usual (smooth) optimality criteria (see [4]). In fact, the second order differentiability condition on F_K with K finite is also satisfied.

Under the same assumption made above, we are going to give the convergence proofs for Algorithms 1.1, 1.2 and 1.3 simply by adjusting the proof of Theorem 1. All these modified algorithms share the same conclusion with the basic Algorithm 1.

PROOF FOR ALGORITHM 1.1. If the Φ' in the proof of Theorem 1 is not the optimal value Φ^*, from the general equivalence theorem, $\lim\inf_j(\bar{d}(\xi_j) - d^{\#}(\xi_j)) > 0$. Then the index k in Algorithm 1.1 will be bounded above by some N_0 and the ε_k will be bounded below by $\varepsilon_{N_0} > 0$. This makes the modification the same as the basic Algorithm 1. We can draw the conclusion from there that $\lim\inf_j(\bar{d}(\xi_j) - d^{\#}(\xi_j)) = 0$ and $\Phi_j \searrow \Phi^*$, etc.

PROOF FOR ALGORITHM 1.2. From Step 3 in Algorithm 1.2, $d(y_j, \xi_j) - d^{\#}(\xi_j) \geqslant \min(\varepsilon_1, r[\bar{d}(\xi_j) - d^{\#}(\xi_j)])$. Simply by making the following changes, the proof of Theorem 1 can be used to give the convergence proof for Algorithm 1.2. (3.2) is replaced by $\Phi(M_{j+1}) - \Phi(M_j) \leqslant -\alpha\beta_j \min(\varepsilon_1, r[\bar{d}(\xi_j) - d^{\#}(\xi_j)]) \leqslant -\alpha\beta_j \min(\varepsilon_1, ra)$. (3.4) is replaced by $\Phi(M[(1-\beta)\xi_j + \beta\xi_{y_j}]) - \Phi(M(\xi_j)) - (-\alpha\beta[d(y_j, \xi_j) - d^{\#}(\xi_j)]) \leqslant \beta\{$1st term (same) $- (1-\alpha)\min[\varepsilon_1, d(y_j, \xi_j) - d^{\#}(\xi_j)]\}$.

PROOF FOR ALGORITHM 1.3. At every stage, there is only a finite number of support points. The algorithm will visit Step 4 of Algorithm 1.3 at most this number of times and then go to the other steps. By omitting the ξ_j constructed in Step 4, the resulting infinite subsequences are constructed in the same way as in Algorithm 1. Its convergence behavior is thus guaranteed by Theorem 1.

If the y_j in Step 3 of Algorithm 1 is chosen such that $d(y_j, \xi_j) - d^{\#}(\xi_j) = r_j(\bar{d}(\xi_j) - d^{\#}(\xi_j))$ with $r_j \to 0$, what conditions on $\{r_j\}$ are required for the convergence to optimal designs? For the sake of simplicity, the following theorem is proved only for the iterative scheme (2.3). That is justified by the fact that Step 2 of Algorithm 1 will be skipped eventually (from Remark 1 following Theorem 1). Even for the modified Algorithm 1.1, Step 3 will be visited infinitely many times.

DEFINITION 2. Let S be a subset of R^q and f be a function on S. f is called Lipschitz on S with Lipschitz constant L iff $|f(x) - f(y)| \leqslant L|x - y|$ for all $x, y \in S$.

An analogue of the following theorem in the unconstrained case was obtained by Zoutendijk (page 48, [7]).

THEOREM 2. *Suppose* $\nabla \Phi$ *is Lipschitz on* $F_{\Phi(M(\xi_0))}$ *with Lipschitz constant* L. *Let* $\{\xi_j\}$ *be constructed according to* (2.3) *with* $d(y_j, \xi_j) - d^{\#}(\xi_j) = r_j(\bar{d}(\xi_j) - d^{\#}(\xi_j))$, $r_j > 0$ *and* $\sum_{j=0}^{\infty} r_j^2 = \infty$. *Then* $\Phi(M_j)$ *converges monotonically to the optimal value* Φ^*.

PROOF. Define the convex function $g(u) = \Phi((1 - u)M_j + uM(\xi_{y_j}))$ and $[0, \bar{u}] = \{u : g(u) \leqslant g(0), u \in [0, 1]\}$. $\bar{u} > 0$ since $r_j > 0$. In particular, the minimizing α_j in (2.3) is in $[0, \bar{u}]$. It is also true that

$$(1 - u)M(\xi_j) + uM(\xi_{y_j}) \in F_{\Phi(M_0)} \quad \text{for} \quad 0 \leqslant u \leqslant \bar{u}.$$

$$\begin{aligned} g(u) - g(0) &= u\textstyle\int_0^1 \langle \nabla \Phi(M_j + tuM(\xi_{y_j} - \xi_j)), M(\xi_{y_j} - \xi_j)\rangle\, dt \\ &= u\textstyle\int_0^1 \langle \nabla \Phi(M_j + tuM(\xi_{y_j} - \xi_j)) - \nabla \Phi(M_j), M(\xi_{y_j} - \xi_j)\rangle\, dt \\ &\quad - u(d(y_j, \xi_j) - d^{\#}(\xi_j)) \\ &\leqslant u^2 L\textstyle\int_0^1 t|M(\xi_{y_j} - \xi_j)|^2\, dt - ur_j(\bar{d}(\xi_j) - d^{\#}(\xi_j)) \\ &\leqslant u^2 LD/2 - ur_j(\bar{d}(\xi_j) - d^{\#}(\xi_j)) \end{aligned}$$

where $D = 4 \max_{M \in \mathfrak{M}} |M|^2$, $0 \leqslant u \leqslant \bar{u}$.

Define $H(u) = u^2 LD/2 - ur_j(\bar{d}(\xi_j) - d^{\#}(\xi_j))$. $H(0) = 0$; $H(\bar{u}) \geqslant g(\bar{u}) - g(0) = 0$ and $H'(0) < 0$. Therefore, $\hat{u}$, the minimizing u-value of H over the real line, is in $(0, \bar{u})$ and $H(\hat{u}) = -r_j^2(\bar{d}_j - d_j^{\#})^2/2LD$.

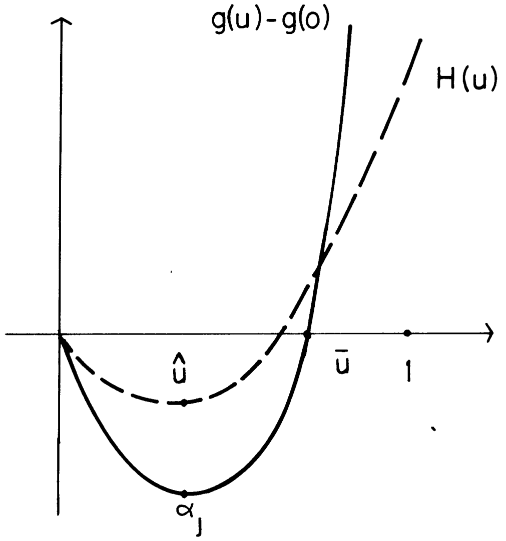

FIG. 1

Figure 1 will help to understand the above arguments.

$$\begin{aligned}\Phi(M_{j+1}) - \Phi(M_j) &= g(\alpha_j) - g(0)\\ &= \min\{ g(u) - g(0) \colon 0 \leqslant u \leqslant \bar{u}\}\\ &\leqslant \min\{H(u)\colon 0 \leqslant u \leqslant \bar{u}\} = H(\hat{u})\\ &= -r_j^2\left(\bar{d}_j - d_j^{\#}\right)^2/2LD.\end{aligned} \tag{3.6}$$

If $\lim\inf_j(\bar{d}_j - d_j^{\#}) > 0$, (3.6) and $\Sigma_{j=0}^{\infty} r_j^2 = \infty$ will imply that $\Sigma_j(\Phi_{j+1} - \Phi_j) = -\infty$, a contradiction. Therefore, there exist a subsequence $\{n_j\}$ with $\bar{d}_{n_j} - d_{n_j}^{\#} \to 0$. From the compactness of $\mathfrak{M}$, there exists a subsequence $n_{j(i)}$ with $\xi_{n_{j(i)}}$ converging to some ξ'. Therefore $\bar{d}(\xi') - d^{\#}(\xi') = 0$ and $\Phi(M(\xi_{n_{j(i)}})) \searrow \Phi^*$. From the monotonicity of $\{\Phi_i\}$, this implies that $\Phi_i \searrow \Phi^*$. □

This theorem is particularly useful when the line minimization can be executed easily. Attention should then be focused on the choices of $\{r_j\}$ (or equivalently the choices of $\{x_j\}$).

4. Discretization for the infinite support case. When $\mathfrak{X}$ is not a finite set, Step 3 of Algorithm 1 involves a search for the maximum d-value over the whole region. In general this is not implementable on a digital computer. The usual practice is to choose a finite subset $\mathfrak{X}_i$ of $\mathfrak{X}$, work there for a while and if not satisfied, augment the $\mathfrak{X}_i$, and so forth. In this section, theoretical guidelines for choosing $\mathfrak{X}_i$ and the convergence to optimal designs are given. Since the actual choice of $\mathfrak{X}_i$ will depend on each specific problem, we do not intend to provide any recipe here. Although the algorithms described below are not directly applicable, they do give some clues to the computational aspects of the problem and clarify certain theoretical points.

Choose an increasing sequence of finite sets $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \ldots$ such that

$$\sup_{x \in \mathcal{X}_n} d(x, \xi) = \bar{d}_n(\xi) \geqslant \bar{d}(\xi) - \frac{1}{n}, \tag{4.1}$$

for all $M(\xi) \in F_C$, $C < \infty$. (F_C is defined in Section 3). Therefore

$$\sup_n \bar{d}_n(\xi) = \bar{d}(\xi).$$

Existence of such $\mathcal{X}_n$'s is guaranteed by the following lemma.

LEMMA 1. *If Φ is continuously differentiable in a neighborhood of F_C, there exists a sequence $\{\mathcal{X}_n\}$ satisfying* (4.1).

The result follows from the boundedness of $\nabla \Phi$ on F_C and the compactness (hence totally boundedness) of the set $\{xx^T : x \in \mathcal{X}\}$.

Having chosen the $\mathcal{X}_n$ in (4.1), Step 3 in Algorithm 1 can be modified in various ways involving finding the maximum over $\mathcal{X}_n$ each time. One such algorithm is presented below.

ALGORITHM 2. Choose $n_0 < \infty$.

Step 0. $j = 0$, $n = n_0$, choose a ξ_0 with $\Phi(M(\xi_0)) < \infty$.

Steps 1, 2. Same as in Algorithm 1.

Step 3. If $\bar{d}_n(\xi_j) - d^{\#}(\xi_j) \geqslant 1/n$, choose a y_j from $\mathcal{X}_n$ with $d(y_j, \xi_j) = \bar{d}_n(\xi_j)$, let $\xi_{j+1} = (1 - \alpha_j)\xi_j + \alpha_j \xi_{y_j}$ with α_j minimizing $\Phi(M[(1 - \alpha)\xi_j + \alpha\xi_{y_j}])$ over $0 \leqslant \alpha \leqslant 1$, $j = j + 1$, go to 1; else, $n = n + 1$, go to 3.

COMMENT. When the algorithm loops at Step 3, we simply work on $\mathcal{X}_n - \mathcal{X}_{n-1}$.

THEOREM 3. *Let $\{\xi_j\}$ be constructed according to Algorithm 2. Under the assumptions of Lemma 1 with $C = \Phi(M_0)$, $\Phi(M_j)$ converges monotonically to the optimal value Φ^*.*

PROOF. All the steps referred to in the proof are from Algorithm 2.

(i) ∞-loop at Step 3: if this happens at ξ_j, $\bar{d}_n(\xi_j) - d^{\#}(\xi_j) \leqslant 1/n$ for all $n \geqslant n_0$ implies $\bar{d}(\xi_j) = d^{\#}(\xi_j)$ by (4.1). From the general equivalence theorem, ξ_j is optimal.

(ii) It is clear that $\{\Phi_j\}$ is a monotone decreasing sequence. The proof of Theorem 1 also shows that Step 1 in Algorithm 2 will be skipped eventually (otherwise the sequence $\{\Phi_j\}$ is not Cauchy). In terms of the asymptotic considerations, we may assume that all iterations take place at Step 3.

If an ∞-loop at Step 3 does not happen, Algorithm 2 will construct a sequence $\{\xi_j\}_N^\infty$ with $\{n_j\}$, the corresponding indices of the set $\mathcal{X}_n$. From Step 3 and (4.1),

$$d(y_j, \xi_j) - d^{\#}(\xi_j) \geqslant \frac{1}{n_j} \geqslant \bar{d}(\xi_j) - \bar{d}_{n_j}(\xi_j).$$

Therefore,

$$d(y_j, \xi_j) - d^\#(\xi_j) \geqslant \big(d(y_j, \xi_j) - d^\#(\xi_j) + \bar{d}(\xi_j) - \bar{d}_{n_j}(\xi_j)\big)/2$$
$$= \big(\bar{d}(\xi_j) - d^\#(\xi_j)\big)/2 \qquad \text{for all} \quad j \geqslant N.$$

This choice of y_j is a special case of Algorithm 1 with $r = \frac{1}{2}$. Convergence of the present algorithm thus follows from there. □

5. General step-length algorithms. All the algorithms discussed before are equivalent to the following iterative scheme:

$$\xi_{i+1} = (1 - \alpha_i)\xi_i + \alpha_i \eta_i, \tag{5.1}$$

where η_i is chosen in Steps 1 or 3 of Algorithm 1 according to various methods considered there and α_i is determined by a line search method. For economic reasons, we may omit the line searches for α_i and instead choose a prescribed sequence $\{\alpha_i\}$ before the execution of the algorithm. An important special case is the choice $\alpha_i = (i + 1)^{-1}$. This occurs in sequential design of experiments. Suppose we have already performed the experiments at $\{x_i\}_{i=1}^n$ up to the nth iteration (therefore, $M(\xi_n) = n^{-1}\Sigma_{i=1}^n x_i x_i^T$) and the next experiment is chosen to be performed at x_{n+1}; then $M(\xi_{n+1}) = (n + 1)^{-1}\Sigma_{i=1}^{n+1} x_i x_i^T = (1 - (n + 1)^{-1}M(\xi_n) + (n+1)^{-1}x_{n+1}x_{n+1}^T$. Here the step length is set at $(i+1)^{-1}$ and no line search methods are necessary. Such algorithms are called general step-length algorithms. If ξ_i is not optimal for any finite i (therefore $\alpha_i > 0$, for all i), we want to know:

(5.2) What conditions on $\{\alpha_i\}$ will give $\xi_i \to \xi^*$?

In optimization theory, this type of problem was treated by Levitin and Polyak (1966) for the unconstrained case with Φ bounded and by Polyak (1967) for the constrained case. In the context of optimal design theory, it was treated by Fedorov (1969), Wynn (1970), Tsay (1976) and Wu and Wynn (1978). The results of this section are closely related to those in [12] but are different in form.

Since the Φ-sequence is no longer monotone, the method of proof in Section 3 cannot be used here. A further complication is the unboundedness of Φ on a portion of $\mathfrak{M}$. This is typically true for D-optimality or Φ_p-optimality. To cope with this novel feature, new techniques are developed in the proof of Theorem 5 and in [12]. The author believes that the technique will be useful in other optimization problems which share the same feature.

The following notation is needed for the proof of Theorem 4.

$$F_K = \{M : M \in \mathfrak{M}, \Phi(M) \leqslant K\}.$$

$$\partial F_K = \text{boundary of} \quad F_K \quad \text{in } \mathfrak{M}.$$

$$\text{dist}(A, B) = \min\{|M_1 - M_2| : M_1 \in A, M_2 \in B \qquad \text{for} \quad A, B \subseteq \mathfrak{M}.\}$$

$$D = 4\max\{|M|^2 : M \in \mathfrak{M}\}.$$

$$P(K) = \max_{M \in F_K} |\nabla \Phi(M)| \big(= \max_{M \in \partial F_K} |\nabla \Phi(M)| \text{ if} \quad \Phi \quad \text{is convex}\big).$$

Let the η_i and α_i in (5.1) be chosen according to the following:

$$d(\eta_i, \xi_i) - d^{\#}(\xi_i) = r_i(\bar{d}(\xi_i) - d^{\#}(\xi_i)),\ \alpha_i, r_i > 0, \qquad (5.3)$$

$$\alpha_i / r_i \to 0 \quad \text{and} \quad \Sigma_{i=0}^{\infty} \alpha_i r_i = \infty,$$

where $d(\eta_i, \xi_i) = -\langle \nabla \Phi(M(\xi_i), M(\eta_i)\rangle$.

THEOREM 4. *Suppose* $\nabla \Phi$ *is Lipschitz on* $F_{K'}$ *with Lipschitz constant* $L(K')$, $\Phi(M_0) < K < K' < \infty$. *If*

$$\sup_i \alpha_i / r_i \leqslant 2(K - \Phi^*)/L(K')D \qquad (5.4)$$

and

$$\sup_i \alpha_i \leqslant (K' - K)/P(K')D^{1/2},$$

then the sequence constructed according to (5.1) *and* (5.3) *lies in* F_K *and* $\Phi(M(\xi_j))$ *converges to the optimal value* Φ^*.

PROOF. The idea of the proof is to bound the sequence from going too far and proceed as if Φ is bounded.

(i) $M(\xi_i) \in F_K \Rightarrow M(\xi_{i+1}) \in F_{K'}$:

$$|M_{i+1} - M_i| = \alpha_i |M(\eta_i - \xi_i)| \leqslant \alpha_i D^{1/2} \leqslant (K' - K)/P(K').$$

It is then sufficient to show that $(K' - K)/P(K') \leqslant \text{dist}(\partial F_K, \partial F_{K'})$. This is geometrically quite obvious. An analytic proof is the following.

Choose $M \in \partial F_K$, $M' \in \partial F_{K'}$ and let $g(u) = \Phi((1-u)M + uM')$.

$$0 \leqslant g(1) - g(0) = K' - K = \int_0^1 \langle \nabla \Phi(M + u(M' - M)), M' - M\rangle\, du$$
$$\leqslant \max_{0 \leqslant u \leqslant 1} |\nabla \Phi(M + u(M' - M))| \cdot |M' - M| \leqslant P(K') \cdot |M' - M|.$$

This implies that

$$\text{dist}(\partial F_K, \partial F_{K'}) \geqslant (K' - K)/P(K'). \qquad (5.5)$$

(ii) $M_i \in F_K$ for all i: since $M(\xi_0) \in F_K$ by assumption, we need only to show that $M(\xi_i) \in F_K$ implies $M(\xi_{i+1}) \in F_K$ for any i. From (i), $M(\xi_{i+1}) \in F_{K'}$ and the Lipschitz condition on $F_{K'}$ can be invoked.

$$\begin{aligned}\Phi(M_{i+1}) - \Phi(M_i) &= \alpha_i \int_0^1 \langle \nabla \Phi(M(\xi_i) + t\alpha_i M(\eta_i - \xi_i)) \\ &\quad - \nabla \Phi(M(\xi_i)), M(\eta_i - \xi_i)\rangle\, dt + \alpha_i \langle \nabla \Phi(M(\xi_i)), M(\eta_i - \xi_i)\rangle \\ &\leqslant \alpha_i L(K') \int_0^1 t\alpha_i |M(\eta_i) - M(\xi_i)|^2\, dt - \alpha_i(d(\eta_i, \xi_i) - d^{\#}(\xi_i)) \qquad (5.6) \\ &\leqslant \alpha_i^2 L(K')D/2 - \alpha_i r_i(\bar{d}(\xi_i) - d^{\#}(\xi_i)) \\ &\leqslant -\alpha_i r_i(\Phi(M(\xi_i)) - \Phi^* - \alpha_i L(K')D/2r_i).\end{aligned}$$

The last inequality follows from the convexity of Φ.

CASE 1. If $\Phi(M_i) \geqslant \Phi^* + \alpha_i L(K')D/2r_i$, then $\Phi(M_{i+1}) \leqslant \Phi(M_i) \leqslant K$ from (5.6).

CASE 2. If $\Phi(M_i) \leqslant \Phi^* + \alpha_i L(K')D/2r_i$, then

$$\begin{aligned}\Phi(M_{i+1}) &\leqslant (1-\alpha_i r_i)\Phi(M_i) + \alpha_i r_i \Phi^* + \alpha_i^2 LD/2 \text{ (from (5.6))}\\ &\leqslant (1-\alpha_i r_i)(\Phi^* + \alpha_i LD/2r_i) + \alpha_i r_i \Phi^* + \alpha_i^2 LD/2\\ &= \Phi^* + \alpha_i LD/2r_i \leqslant \Phi^* + \frac{2(K-\Phi^*)}{LD}\frac{LD}{2} = K.\end{aligned}$$

(iii) $\lim\inf_i(\Phi(M_i) - \Phi^*) = 0$: from (ii) we can apply the Lipschitz condition to all the M_i's.

If $\lim\inf_i(\Phi(M_i) - \Phi^*) = a > 0$, from (5.6) and $\alpha_i / r_i > 0$, there exists an N_0 such that

$$\Phi(M_{i+1}) - \Phi(M_i) \leqslant -\alpha_i r_i \frac{a}{2} \qquad \text{for all } i \geqslant N_0.$$

This implies that $\Sigma_{i=0}^{\infty}(\Phi(M_{i+1}) - \Phi(M_i)) = -(a/2)\Sigma_{i=0}^{\infty}\alpha_i r_i = -\infty$, contradicting the fact that Φ is bounded below.

(iv) $\lim_i \Phi(M_i) = \Phi^*$: from (iii), there exists a subsequence $\{n_i\}$ such that $\Phi(M_{n_i}) \to \Phi^*$. For each i, either $\Phi_{i+1} \leqslant \Phi_i$ or $\Phi_i \leqslant \Phi^* + \alpha_i L(K)D/2r_i$ holds (from (5.6)). By induction, this implies that

$$\Phi_k \leqslant \max_{n_i \leqslant j \leqslant k-1}\{\Phi_{n_i}, A_j\}$$

where $A_j = \Phi^* + \alpha_j L(K)D/2r_j$ and $n_i \leqslant k$. Since both $\Phi(M_{n_i})$ and $\sup_{j \geqslant n_i} A_j$ converge to Φ^* as $i \to \infty$, $\lim_i \Phi(M_i) = \Phi^*$ is established. □

When $\nabla \Phi$ is Lipschitz on $\mathfrak{M}$, (i) and (ii) in the above proof are not necessary, so we can conclude the same result without assuming (5.4).

COROLLARY 1. *Suppose $\nabla \Phi$ is Lipschitz on $\mathfrak{M}$. Let $\{\xi_i\}$ be constructed according to* (5.1) *and* (5.3). *Then $\Phi(M(\xi_i))$ converges to the optimal value Φ^*.*

For $r_i \geqslant r > 0$, (5.3) becomes

(5.7) $$\alpha_i > 0, \qquad \alpha_i \to 0 \qquad \text{and} \quad \Sigma_{i=0}^{\infty}\alpha_i = \infty.$$

This is the condition under which the same type of problem was treated by the aforementioned authors. In [12], (5.7), without the extra assumption (5.4), is shown to be a sufficient condition for optimality convergence for many well-known optimality criteria.

The undesirable upper bounds in Theorem 4 can be removed by imposing a further condition either on Φ or on the choices of η_i in (5.1). The first approach was adopted in Wu and Wynn (1978). The second approach originated with Shor and was further expounded by Polyak (1967). The following is a restatement and extension of Polyak's results in the design context.

DEFINITION 3. A linear functional u of R^q is a support functional to a subset Q of R^q at the point z_0 iff $\langle u, z - z_0\rangle \geqslant 0$ for all $z \in Q$.

If Φ is convex and differentiable in R^q and $Q = \{z : z \in R^q, \Phi(z) \leqslant \Phi(z_0)\}$, then $-\nabla \Phi(z_0)$ is a support functional to Q at z_0.

DEFINITION 4. Let $\mathcal{D}$ be a subset of R^q. $z_0 \in \mathcal{D}$ is a projection of $y \in R^q$ onto $\mathcal{D}$ iff $|y - z_0| = \min\{|y - z| : z \in \mathcal{D}\}$. z_0 is denoted by $P_{\mathcal{D}}(y)$.

Two design sequences are constructed by utilizing the concept of a support functional. Throughout this subsection, we assume that Φ is convex and continuous in $\mathfrak{N}_0$, a neighborhood of $\mathfrak{M}$ in the space of $k \times k$ matrices

A. $M(\xi_{i+1}) = P_{\mathfrak{M}}(M(\xi_i) + \alpha_i N_i/|N_i|)$ where $\alpha_i > 0$, N_i is a support functional to the set $\{M : M \in \mathfrak{N}_0, \Phi(M) \leqslant \Phi(M(\xi_i))\}$ at $M(\xi_i)$.

B. $M(\xi_{i+1}) = M(\xi_i) + \alpha_i(M(\eta_i) - M(\xi_i))/|M(\eta_i - \xi_i)|$, $M(\eta_i) \in \mathfrak{M}$, $0 < \alpha_i \leqslant |M(\eta_i) - M(\xi_i)|$ and $M(\eta_i) - M(\xi_i)$ is a support functional to the set

$$\{M : M \in \mathfrak{N}_0, \Phi(M) \leqslant \Phi(M(\xi_i))\}$$

at $M(\xi_i)$.

THEOREM 5. *For $M(\xi_i)$ constructed in* A *or* B *and $\{\alpha_i\}$ satisfying $\alpha_i \to 0$ and $\sum_{i=0}^{\infty}\alpha_i = \infty$, $\Phi(M(\xi_i))$ converges to the optimal value Φ^*.*

This result is essentially the same as Theorems 1 and 2 of Polyak (1967). In the design context, a proof was given in Theorem 8 in Wu (1976). All the technical details are omitted here.

Two choices of the support functional in A or B are given below.

EXAMPLE 1. When Φ is differentiable, the N_i in case A can be taken to be $-\nabla\Phi(M(\xi_i))$. But the projection onto $\mathfrak{M}$ which defines $M(\xi_{i+1})$ involves a search over $\mathfrak{M}$.

EXAMPLE 2. If the boundary of $\mathfrak{M}$ in $\mathcal{L}(\mathfrak{M})$, the linear manifold spanned by $\mathfrak{M}$, is equal to $\{xx^T : x \in \mathcal{X}\}$, the convergence condition for the iterative scheme (5.8) is quite interesting.

$$\xi_{i+1} = (1 - \alpha_i)\xi_i + \alpha_i\xi_{x_i}, \tag{5.8}$$

where x_i is chosen to maximize $[d(x, \xi_i) - d^{\#}(\xi_i)]/|xx^T - M(\xi_i)|$ over $\mathcal{X}$. In order that Theorem 5 can be applied, we have to show that $M(\xi_{x_i} - \xi_i)$ is a support functional. From the above assumption on the boundary of $\mathfrak{M}$,

$$\begin{aligned}\max_{x\in\mathcal{X}} & \frac{-\langle \nabla\Phi(M_i), xx^T - M_i\rangle}{|xx^T - M_i|} \\ &= \max_{M\in\partial\mathfrak{M}} \frac{-\langle \nabla\Phi(M_i), M - M_i\rangle}{|M - M_i|} \\ &= \max_{M\in\mathfrak{M}} \frac{-\langle \nabla\Phi(M_i), M - M_i\rangle}{|M - M_i|}.\end{aligned} \tag{5.9}$$

The last equality follows from the scale invariance of the normalized directional derivative in (5.9). Therefore $x_i x_i^T - M(\xi_i)$ lies in the projection direction of $-\nabla\Phi(M_i)$ onto $\mathcal{L}(\mathfrak{M})$, i.e., there exists a $\lambda_i > 0$ such that $x_i x_i^T - \nabla\Phi(M_i)$ is the projection of $-\lambda_i\nabla\Phi(M_i)$ onto $\mathcal{L}(\mathfrak{M})$. If the initial M_0 is nonsingular, all the succeeding M_i's are nonsingular and hence lie in the interior of $\mathfrak{M}$ in $\mathcal{L}(\mathfrak{M})$. $-\lambda_i\nabla\Phi(M_i) - (x_i x_i^T - M_i)$ is therefore orthogonal to $\mathcal{L}(\mathfrak{M})$.

$$\langle -\lambda_i\nabla\Phi(M_i) - (x_i x_i^T - M_i), M - M_i\rangle = 0 \qquad \text{for all} \quad M \in \mathfrak{M}.$$

From the convexity of Φ, we also have

$$\langle \nabla \Phi(M_i), M - M_i \rangle \leqslant 0 \quad \text{for all} \quad M \in \mathfrak{M}_0 \quad \text{satisfying} \quad \Phi(M) \leqslant \Phi(M_i).$$

Together they give $\langle x_i x_i^T - M_i, M - M_i \rangle \geqslant 0$ for all $M \in \mathfrak{M}_0$ satisfying $\Phi(M) \leqslant \Phi(M_i)$. Therefore $x_i x_i^T - M_i$ is a support functional to the set $\{M : M \in \mathfrak{M}_0, \Phi(M) \leqslant \Phi(M_i)\}$ at M_i. According to Theorem 5 with design sequence constructed in B, the convergence conditions are:

$$\alpha_i |x_i x_i^T - M_i| \to 0 \quad \text{and} \quad \Sigma_{i=0}^{\infty} \alpha_i |x_i x_i^T - M_i| = \infty.$$

If x_i is chosen to maximize $d(x, \xi_i) - d^{\#}(\xi_i)$, it does not necessarily satisfy the assumptions of Theorem 5 and hence the result there can not be applied. Here we do see the difference of the asymptotic behaviors of procedures involving two different choices of x_i: to maximize $d(x, \xi_i) - d^{\#}(\xi_i)$ or $[d(x, \xi_i) - d^{\#}(\xi_i)]/|xx^T - M(\xi_i)|$.

Acknowledgments. This paper is based on part of the author's Ph.D. dissertation written at the University of California, Berkeley, under the supervision of Professor Peter J. Bickel. The author would like to thank Professors P. J. Bickel, J. Kiefer and H. P. Wynn for encouragement and help.

REFERENCES

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

[2] Fedorov, V. V. (1969, Transl. 1972). *Theory of Optimal Experiments*. Academic Press, New York.

[3] Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Res. Logist. Quart.* **3** 95–110.

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

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

[6] Levitin, E. S. and Polyak, B. T. (1966). Constrained minimization methods. *USSR Comput. Math. and Math. Phys.* **6 5** 1–50.

[7] Polak, E. (1971). *Computational Methods in Optimization—A Unified Approach*. Academic Press, New York.

[8] Polyak, B. T. (1967). A general method of solving extremum problems. *Soviet Math. Dokl.* **8** 593–597.

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

[10] Wu, C. F. (1976). Contributions to optimization theory with applications to optimal design of experiments. Ph.D. thesis, University of California, Berkeley.

[11] Wu, C. F. (1978). Some iterative procedures for generating non-singular optimal designs. To appear in *Commun. Statist. A*.

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

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

Department of Statistics
University of Wisconsin
1210 W. Dayton St.
Madison, Wisconsin 53706