The Annals of Statistics
1980, Vol. 8, No. 4,894–912

CONVERGENT DESIGN SEQUENCES, FOR SUFFICIENTLY REGULAR OPTIMALITY CRITERIA, II: SINGULAR CASE[1]

BY CORWIN L. ATWOOD

EG & G Idaho, Inc.

Let an optimality criterion Φ satisfy regularity conditions, including convexity and possession of two continuous derivatives at nonsingular $\mathbf{M}$. $\Phi\{\mathbf{M}\}$ may be minimized at a singular $\mathbf{M}$. A class of design sequences $\{\xi_n\}$ is shown to make $\Phi[\mathbf{M}(\xi_n)]$ converge monotonically to the minimum value. An equivalence theorem for Φ-optimality follows. Techniques which are applicable include vertex direction, conjugate gradient projection, quadratic and "diagonalized quadratic" methods for changing the design weights, and gradient-based methods for making small changes in the support points. Methods are also considered which approximate Φ by a criterion which is infinite for singular $\mathbf{M}$. The results are applied to an example with D_s-optimality.

1. Introduction. This paper generalizes to Φ-optimality the iterative algorithms of Fedorov [7], Wynn [22], and Atwood [3] for D_s-optimality and of Fedorov [7], and Tsay [18] for L-optimality. It extends the results of Atwood [5] on quadratic design sequences and of Wu [20], [21] on gradient projection methods to the case where the Φ-optimal design may be singular, and develops more fully the technique of sliding support points, introduced in [5]. It generalizes to singular Φ-optimality the equivalence theorem of Kiefer [10] for D_s-optimality.

The results are valid for nonsingular as well as singular optimality criteria. In particular, the use of diagonalized quadratic increments and/or sliding support points (Sections 3D and 3E) will probably be helpful with nonsingular criteria such as D-optimality.

Before treating the general results, we summarize the relevant facts about D_s- and L-optimality. These criteria will serve as examples of and motivation for the results of Section 2.

Consider a regression model with uncorrelated observable random variables Y_s, each having mean $\mathbf{f}^T(x)\boldsymbol{\theta}$ and common variance. Here $\boldsymbol{\theta}$ is an unknown k-dimensional parameter and $\mathbf{f}$ is a vector of regression functions on the Euclidean space $\mathscr{X}$ where observations may be taken. A design ξ is a probability distribution on $\mathscr{X}$. If ξ can be realized as a distribution of n x's in $\mathscr{X}$, then the least squares estimator $\hat{\boldsymbol{\theta}}$ has covariance matrix proportional to $\mathbf{M}^{-1}(\xi)$, where the information matrix $\mathbf{M}(\xi)$ is defined as $\mathbf{M}(\xi) = \int \mathbf{f}(x)\mathbf{f}^T(x)\xi(dx)$.

For some designs, e.g., those intended for estimating the first s components of $\boldsymbol{\theta}$, $\mathbf{M}$ may be singular. Let $\mathbf{M}^-$ be a g-inverse of $\mathbf{M}$ as in Rao [13] Section 1b.5, and

Received July 1976; revised July 1978.

[1]Most of this work was done while the author was at the University of California at Davis.

AMS 1970 *subject classifications.* Primary 62K05; secondary 65B99.

Key words and phrases. Singular optimal experimental designs, D_s-optimality, L-optimality, Φ-optimality, iterative design optimization, equivalence theorem for optimal designs.

make the partitions

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix} \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \quad \mathbf{M}^- = \begin{bmatrix} \mathbf{M}^{11} & \mathbf{M}^{12} \\ \mathbf{M}^{21} & \mathbf{M}^{22} \end{bmatrix}$$

where $\boldsymbol{\theta}_1$ is s-dimensional and $\mathbf{M}_{11}$ and $\mathbf{M}^{11}$ are $s \times s$. Define also

$$\mathbf{M}^* = \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^-\mathbf{M}_{21}. \tag{1.1}$$

It is not hard to show that $\mathbf{M}^*$ is unique, regardless of the choice of the g-inverse $\mathbf{M}_{22}^-$. (Set $\mathbf{X} = \mathbf{M}_{22}^-\mathbf{M}_{21}$ in Lemma 6.2 of Karlin and Studden [9]).

If $\mathbf{M}$ is nonsingular, $\mathbf{M}^- = \mathbf{M}^{-1}$. If $\mathbf{M}$ is singular, $\mathbf{M}^-$ is not unique. However, it is not hard to show that $\boldsymbol{\theta}_1$ is estimable iff $\mathbf{M}^*$ is nonsingular, in which case $\mathbf{M}^{11}$ is unique and equal to $(\mathbf{M}^*)^{-1}$ and proportional to the covariance matrix of $\hat{\boldsymbol{\theta}}_1$. (See Section 4a.4 of Rao [13]). Call $\mathbf{M}^*$ the information matrix of $\hat{\boldsymbol{\theta}}_1$. One way to find $\mathbf{M}^*$ or $\mathbf{M}^{11}$ is by taking limits as in Chernoff [6]. Let $\mathbf{A}$ be symmetric such that $\mathbf{M} + \alpha\mathbf{A}$ is nonsingular for α near 0. Then

$$\mathbf{M}^* = \lim_{\alpha\to 0}[\mathbf{M} + \alpha\mathbf{A}]^* \tag{1.2}$$

$$\mathbf{M}^{11} = \lim_{\alpha\to 0}[\mathbf{M} + \alpha\mathbf{A}]^{11}.$$

The limit $\mathbf{M}^*$ is independent of $\mathbf{A}$ and the limit $\mathbf{M}^{11}$ is finite and independent of $\mathbf{A}$ if $\mathbf{M}^*$ is nonsingular.

To consider the continuity of $\mathbf{M}^*$, define the ordering

$$\mathbf{M}_1 \geqslant \mathbf{M}_2 \Leftrightarrow \mathbf{M}_1 - \mathbf{M}_2 \quad \text{is nonnegative definite} \tag{1.3}$$

with strict inequality if $\mathbf{M}_1 \neq \mathbf{M}_2$. For symmetric $\mathbf{M} + \mathbf{B} \geqslant \mathbf{0}$ it is true that $(\mathbf{M} + \mathbf{B})^* \leqslant \mathbf{M}^* + \mathbf{C}$, where $\mathbf{C} \to \mathbf{0}$ as $\mathbf{B} \to \mathbf{0}$, so that if the limit exists,

$$\lim_{B\to 0}(\mathbf{M} + \mathbf{B})^* \leqslant \mathbf{M}^*.$$

Strict inequality can hold if rank $(\mathbf{M} + \mathbf{B}) >$ rank $\mathbf{M}$. So $\mathbf{M}^*$ is only upper semicontinuous in $\mathbf{M}$ at singular $\mathbf{M}$, even though it is continuous along straight paths, i.e., (1.2) holds. As for derivatives, the directional derivative

$$\partial[\mathbf{M} + \alpha\mathbf{A}]^*/\partial\alpha|_{\alpha=0} \tag{1.4}$$

is defined. But $\mathbf{M}^*$ cannot be considered differentiable at singular $\mathbf{M}$, even where $\mathbf{M}^*$ may be continuous, because (1.4) is not linear in $\mathbf{A}$ at singular $\mathbf{M}$. The corresponding assertions about $\mathbf{M}^{11}$ hold where $\mathbf{M}^*$ is nonsingular. These facts are all stated or implicit in [6] and [10].

Simple examples aid the intuition. As an illustration of discontinuity, consider estimation of θ_0 in the model $\mathbf{f}^T(x)\boldsymbol{\theta} = \theta_0 + x\theta_1$ for $0 \leqslant x \leqslant 1$. Let ξ_0 and ξ_x be concentrated at 0 and $x > 0$, respectively. Then $\mathbf{M}[(1 - \alpha)\xi_0 + \alpha\xi_x] \to \mathbf{M}(\xi_0)$ as either $\alpha \to 0$ or $x \to 0$. If $x > 0$ is fixed and $\alpha \to 0$, then $\mathbf{M}^*[(1 - \alpha)\xi_0 + \alpha\xi_x] = 1 - \alpha \to \mathbf{M}^*(\xi_0)$. But if α is fixed and $x \to 0$, then $\mathbf{M}^*[(1 - \alpha)\xi_0 + \alpha\xi_x]$ remains constant. As an illustration of continuity but nondifferentiability, consider $\mathbf{M}^*$ at $\mathbf{M}(\xi^*)$ in Kiefer's example on page 309 of [10].

With this background, we are ready to consider two optimality criteria which satisfy the assumptions of Section 2. They are D_s-optimality:

$$\Phi(\mathbf{M}) = -\log|\mathbf{M}^*|$$

and L-optimality:

$$\Phi(\mathbf{M}) = \operatorname{tr} \mathbf{C}\mathbf{M}^{-1}$$

where $\mathbf{C}$ is given, symmetric and nonnegative definite. If $\mathbf{M}$ is singular, define the L-optimality criterion as

$$\Phi(\mathbf{M}) = \lim_{\alpha \to 0^+} \operatorname{tr} \mathbf{C}[\mathbf{M} + \alpha\mathbf{A}]^{-1}$$

where $\mathbf{M} + \alpha\mathbf{A}$ is symmetric and positive definite for small $\alpha > 0$.

There is a representation which is useful for checking that L-optimality satisfies Assumptions 4 and 5 below. If $\mathbf{C}$ has rank $s < k$, let $\mathbf{P}$ be orthogonal such that

$$\mathbf{PCP}^T = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

for some $s \times s$ matrix $\mathbf{D}$. Then

$$\Phi(\mathbf{M}) = \lim_{\alpha \to 0^+} \operatorname{tr} \mathbf{C}[\mathbf{M} + \alpha\mathbf{A}]^{-1} = \lim_{\alpha \to 0^+} \operatorname{tr} \mathbf{PCP}^T\mathbf{P}[\mathbf{M} + \alpha\mathbf{A}]^{-1}\mathbf{P}^T$$

$$= \operatorname{tr} \mathbf{D}[(\mathbf{PMP}^T)^*]^{-1} \tag{1.5}$$

if $(\mathbf{PMP}^T)^*$ is nonsingular, and $\Phi(\mathbf{M}) = \infty$ otherwise. So the L-optimality criterion can be expressed in terms of the $s \times s$ information matrix of $(\mathbf{P}\hat{\boldsymbol{\theta}})_1$.

2. Main results. Let $\mathbf{f}$, $\mathcal{X}$, ξ, and $\mathbf{M}$ be as defined above. Let Ξ denote the class of designs, and Ξ^+ the (nonempty) class of nonsingular designs, i.e., designs with $\mathbf{M}(\xi)$ nonsingular. Let $\mathfrak{M} = \{\mathbf{M}(\xi) | \xi \in \Xi\}$, let $\mathfrak{M}^+ = \{\mathbf{M}(\xi) | \xi \in \Xi^+\}$, and let $\mathfrak{R}$ be the set of all $k \times k$ matrices, with the Euclidean topology and some norm $\| \ \|$. We will use the convention that $\mathbf{M}$ denotes an element of $\mathfrak{M}$ while $\mathbf{A}$ and $\mathbf{B}$ are arbitrary elements of $\mathfrak{R}$, and that ξ denotes an element of Ξ while η is a signed measure. Let Φ be the criterion function to be minimized.

ASSUMPTIONS.

1. $\mathbf{f}$ is continuous and $\mathcal{X}$ is compact.
2. Φ is defined and real valued on $\mathfrak{N}$, a neighborhood in $\mathfrak{R}$ of $\mathfrak{M}^+$.
3. The first two partial derivatives of Φ exist and are continuous in $\mathfrak{N}$, with

$$\partial\Phi(\mathbf{M} + \alpha\mathbf{A})/\partial\alpha|_{\alpha=0}$$

and

$$\partial^2\Phi(\mathbf{M} + \alpha\mathbf{A} + \beta\mathbf{B})/\partial\alpha\,\partial\beta|_{\alpha=0,\beta=0}$$

linear in $\mathbf{A}$ and $\mathbf{B}$ for all $\mathbf{M} \in \mathfrak{N}$, $\mathbf{A} \in \mathfrak{R}$, $\mathbf{B} \in \mathfrak{R}$.

4. Φ is convex on $\mathfrak{M}^+$.
5. $\partial^2\Phi(\mathbf{M} + \alpha\mathbf{A})/\partial\alpha^2|_0 > 0$ for all $\mathbf{M} \in \mathfrak{M}^+$ and symmetric $\mathbf{A}$ such that $\partial\Phi(\mathbf{M} + \alpha\mathbf{A})/\partial\alpha|_0 \neq 0$.

These assumptions do not require $\Phi(\mathbf{M})$ to be defined for singular $\mathbf{M}$. One approach, followed by Gribik and Kortanek [8], is to call ξ^* "limit optimal" if there is a sequence $\{\mathbf{M}_n\}$ in $\mathfrak{M}^+$ such that $\mathbf{M}_n \to \mathbf{M}(\xi^*)$ and $\Phi(\mathbf{M}_n) \to \inf_{\mathfrak{M}^+}\Phi(\mathbf{M})$. Gribik and Kortanek define $\Phi(\mathbf{M}) = \infty$ for singular $\mathbf{M}$, but they could have defined it arbitrarily or left it undefined, since limit optimality does not depend on the value of $\Phi(\mathbf{M})$ at singular $\mathbf{M}$.

Another approach when $\mathbf{M}$ is singular is to define

$$\Phi(\mathbf{M}) = \lim_{\varepsilon \to 0^+} \inf_{\|M' - M\| < \varepsilon} \Phi(\mathbf{M}') \tag{2.1}$$

for $\mathbf{M}'$ in $\mathfrak{M}^+$. It is easy to show that with this definition Φ is convex on all of $\mathfrak{M}$. Convexity and (2.1) imply that for $(1-\alpha)\mathbf{M} + \alpha\mathbf{M}' \in \mathfrak{M}^+$,

$$\lim_{\alpha \to 0} \Phi\left[(1-\alpha)\mathbf{M} + \alpha\mathbf{M}'\right] = \Phi(\mathbf{M}).$$

Moreover, ξ^* is Φ-optimal, i.e., $\mathbf{M}(\xi^*)$ minimizes $\Phi(\mathbf{M})$ over $\mathfrak{M}$, if and only if ξ^* is limit optimal. This approach fits our earlier definitions of D_s- and L-optimality, and is the one we will follow.

Now for some further comments about the assumptions. Assumption 1 implies that Ξ and $\mathfrak{M}$ are compact, so an optimal $\mathbf{M}(\xi)$ exists. Also, $\inf \Phi(\mathbf{M})$ is finite, since Φ is convex on a compact space. The linearity part of Assumption 3 gives the usual relations between directional derivatives and partial derivatives. Assumption 5 is not required to prove the main convergence theorem, but it is required to construct the quadratic and diagonalized quadratic increments of Section 3. Assumptions 4 and 5 cannot be strengthened in any obvious way which is valid for D_s-optimality, since, if $\mathbf{M}$ moves along a straight path where $\mathbf{M}^*$ is constant, $\Phi(\mathbf{M})$ is constant and so not strictly convex.

A criterion function Φ which may be finite at a singular $\mathbf{M}$ will be called a singular criterion. Singularity of Φ is allowed by the assumptions but not required, so the results which follow also apply to nonsingular criteria such as D-optimality and L-optimality with nonsingular $\mathbf{C}$.

A condition which implies Assumptions 4 and 5 is given by the following lemma. Call Ψ an order reversing function of symmetric matrices of a given size if $\Psi(\mathbf{M}_1) \leqslant \Psi(\mathbf{M}_2)$ when $\mathbf{M}_1 \geqslant \mathbf{M}_2$.

LEMMA 2.1. *If Φ satisfies Assumptions* 1–3, *and $\Phi(\mathbf{M}) = \Psi(\mathbf{M}^*)$ for some order reversing Ψ with*

$$\partial^2\Psi(\mathbf{M}^* + \alpha\mathbf{A})/\partial\alpha^2|_0 > 0 \tag{2.2}$$

for all $s \times s$ symmetric $\mathbf{A} \neq 0$ and all nonsingular $\mathbf{M}^*$, then Φ satisfies Assumptions 4 *and* 5.

PROOF. As shown by Kiefer [10], $\mathbf{M}^*$ is concave in $\mathbf{M}$, with respect to the ordering (1.3), so if $\mathbf{M}_1$ and $\mathbf{M}_2$ are in $\mathfrak{M}$,

$$\begin{aligned}\Phi\left[(1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2\right] &= \Psi\{\left[(1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2\right]^*\} \\ &\leqslant \Psi\left[(1-\alpha)\mathbf{M}_1^* + \alpha\mathbf{M}_2^*\right] \leqslant (1-\alpha)\Psi(\mathbf{M}_1^*) + \alpha\Psi(\mathbf{M}_2^*) \\ &= (1-\alpha)\Phi(\mathbf{M}_1) + \alpha\Phi(\mathbf{M}_2)\end{aligned}$$

by the order reversing property and convexity of Ψ respectively. So Φ is convex. In fact, if $\mathbf{M}$ moves in a direction such that the first directional derivative of $\Phi(\mathbf{M})$ is not zero, then $\mathbf{M}^*$ must initially move in some direction. By (2.2), the directional second derivative of Φ must therefore be nonzero. $\square$

This proof imitates Kiefer's proof [10] of convexity for D_s-optimality. The lemma applies to D_s-optimality immediately, and to L-optimality after the reparametrization (1.5).

The algorithm below gives a class of sequences $\{\xi_n\}$, and the accompanying theorem proves that $\Phi[\mathbf{M}(\xi_n)]$ converges monotonically to the optimal value. First come the necessary definitions and notations.

Let ξ_x denote the design supported at a point x, and for $\mathbf{M}(\xi)$ nonsingular define

$$\varphi(x, \xi) = -\partial\Phi\{\mathbf{M}[(1-\alpha)\xi + \alpha\xi_x]\}/\partial\alpha|_0$$
$$\bar{\varphi}(\xi) = \max_x \varphi(x, \xi).$$

If the criterion is D_s-optimality, then $\varphi(x, \xi) = d_s(x, \xi) - s$. Assumptions 1 and 3 imply that $\varphi(x, \xi)$ attains its maximum at some x. Since $\int \varphi(x, \xi)\xi(dx) = 0$, we see that $\bar{\varphi}(x) \geqslant 0$.

Fix ε_1, ε_2, ε_3, a_1, and a_2 with

$$0 < \varepsilon_1 < \varepsilon_2 < 1$$

and ε_3, a_1 and a_2 all > 0 and < 1. In practice ε_3 will be near 0 and a_1 near 1. These constants determine how fast the design sequence may approach a singular design, how much time should be spent on each iteration searching for the best improvement in a given direction, etc.

Let h be any matrix- or real-valued function defined on $\mathfrak{M}^+$ such that for $a > 0$,

$$\mathbf{M}_1 \geqslant a\mathbf{M}_2 \Rightarrow h(\mathbf{M}_1) \geqslant ah(\mathbf{M}_2)$$

and such that for $\mathbf{M}_1$ nonsingular and $a > 0$,

$$\{\mathbf{M} \in \mathfrak{M} \,|\, h(\mathbf{M}) \geqslant ah(\mathbf{M}_1)\}$$

is a compact convex subset of $\mathfrak{M}^+$. Examples are $h(\mathbf{M}) = \mathbf{M}$, $h(\mathbf{M}) = |\mathbf{M}|^{1/k}$, and $h(\mathbf{M}) = 1/\operatorname{tr}(\mathbf{M}^{-1})$.

Let ξ_0 have finite support with $\mathbf{M}(\xi_0)$ nonsingular. At every iteration, a nonsingular design ξ_n and a number $c_n > 0$ are generated. Write $\bar{\varphi}_n$ for $\bar{\varphi}(\xi_n)$ and $\mathbf{M}_n$ for $\mathbf{M}(\xi_n)$.

ALGORITHM 1.

Step 1. For $n \geqslant 0$, let

$$\eta_n = \xi' - \xi_n$$

for some design ξ' such that

$$\partial\Phi[\mathbf{M}(\xi_n + \alpha\eta_n)]/\partial\alpha|_0 \leqslant -\varepsilon_2\bar{\varphi}_n. \tag{2.3}$$

Then let

$$A_n = \{\alpha | 0 \leqslant \alpha \leqslant 1 \text{ and } \Phi[\mathbf{M}(\xi_n + \alpha\eta_n)] - \Phi(\mathbf{M}_n) \leqslant -\alpha\varepsilon_1\bar{\varphi}_n\}. \tag{2.4}$$

Let $\alpha_n \in A_n$ be such that

$$(2.5)\quad \Phi[\mathbf{M}(\xi_n + \alpha_n\eta_n)] - \Phi(\mathbf{M}_n) \leq a_1\{\min_{\alpha\in A_n} \Phi[\mathbf{M}(\xi_n + \alpha\eta_n)] - \Phi(\mathbf{M}_n)\}.$$

Step 2. If $\mathbf{M}(\xi_n + \alpha_n\eta_n)$ is singular, so necessarily $\alpha_n = 1$, redefine $\alpha_n = a_2$. (This increases $\Phi[\mathbf{M}(\xi_n + \alpha_n\eta_n)] - \Phi(\mathbf{M}_n)$ to at most a_2 times its original value, by the convexity of Φ.) Define $\xi_{n+} = \xi_n + \alpha_n\eta_n$.

Step 3. If desired, replace ξ_{n+} by some other ξ_{n+1}, as long as $\Phi[\mathbf{M}(\xi_{n+1})] \leq \Phi[\mathbf{M}(\xi_{n+})]$ and

$$(2.6)\qquad h[\mathbf{M}(\xi_{n+1})] \geq c_n h[\mathbf{M}(\xi_{n+})]$$

for some c_n with $0 < c_n \leq 1$ and $\Pi_0^n c_i > \varepsilon_3$. If no modification of ξ_{n+} is desired, let $\xi_{n+1} = \xi_{n+}$ and $c_n = 1$.

Step 4. Find $\bar{\varphi}(\xi_{n+1})$. If $\bar{\varphi}(\xi_{n+1}) = 0$, stop. Otherwise increase n by 1 and go to Step 1.

The use of a two stage algorithm, where different kinds of design changes are made in Steps 1 and 3, follows Wu [20], [21].

In Section 3, ways will be given for explicitly obtaining a number of η_n and α_n satisfying Step 1. For now, note that such η_n and $\alpha_n > 0$ exist. For if $\varphi(x, \xi_n)$ is maximized at x_n, then the increment $\eta_n = \xi_{x_n} - \xi_n$ satisfies (2.3). And if η_n satisfies (2.3) then A_n is a nondegenerate interval containing 0; this follows from the behavior of the defining inequality in (2.4) at $\alpha = 0$, and the convexity of Φ.

In Step 1, using $\varepsilon_2 < 1$ allows η_n other than $\xi_{x_n} - \xi_n$. Also, the numbers ε_2 and a_1 reflect the fact that in practice neither $\bar{\varphi}$ nor the minimizing α are known exactly, but both can be known accurately enough so that (2.3) and (2.4) are achievable.

Step 3 might be used to improve the weights on the support points under consideration, before taking the trouble to search for a new x which maximizes $\varphi(x, \xi_{n+1})$. As such it can itself encompass a finite sequence of iterations. Step 3 can also be used to move the support points or to collapse clusters of support points. If Φ is a nonsingular criterion, the restriction (2.6) is generally no problem. For example, if $\Phi(\mathbf{M}) = -\log|\mathbf{M}|$, let $h(\mathbf{M}) = |\mathbf{M}|^{1/k}$. If $\Phi(\mathbf{M}) = \operatorname{tr} \mathbf{C}\mathbf{M}^{-1}$ for some positive definite $\mathbf{C}$, let $h(\mathbf{M}) = 1/\Phi(\mathbf{M})$. Then (2.6) is automatic with $c_n = 1$. In the singular case (2.6) is necessary to keep $\mathbf{M}$ from becoming nearly singular before it is nearly optimal. Ways to use Step 3 will be considered in Section 3.

THEOREM 2.1. *Under Assumptions* 1–4 *the sequence constructed by Algorithm* 1 *satisfies* $\Phi(\mathbf{M}_n) \to \inf_{\mathfrak{M}^+} \Phi(\mathbf{M})$ *monotonically, and some subsequence of* $\{\bar{\varphi}_n\}$ *converges to* 0.

If for singular $\mathbf{M}$, $\Phi(\mathbf{M})$ is defined either by (2.1) or as $+\infty$, then $\inf_{\mathfrak{M}^+}\Phi(\mathbf{M}) = \inf_{\mathfrak{M}} \Phi(\mathbf{M})$.

PROOF. Monotonicity is clear from (2.4). Inequality (6.5) of Kiefer [12] says that for nonsingular ξ

$$(2.7)\qquad \Phi[\mathbf{M}(\xi)] - \inf \Phi(\mathbf{M}) \leq \bar{\varphi}(\xi).$$

Suppose that $\Phi(\mathbf{M}_n) \geqslant \inf \Phi(\mathbf{M}) + \delta$ for all n and some $\delta > 0$. Then

$$\bar{\varphi}_n \geqslant \delta \qquad \text{for all} \quad n. \tag{2.8}$$

This will lead to a contradiction, proving both assertions of the theorem.

By (2.4), $\lim \Phi(\mathbf{M}_n) - \Phi(\mathbf{M}_0) \leqslant -\varepsilon_1 \Sigma \alpha_n \bar{\varphi}_n$. Since $\Phi(\mathbf{M})$ is bounded below, $\Sigma \alpha_n \bar{\varphi}_n$ must be finite. From (2.8), therefore, $\Sigma \alpha_n$ is finite. So, for some n_0 (with $n_0 = 0$ if $\alpha_i < 1$ for all i),

$$\Pi_{n_0}^{\infty}(1 - \alpha_i) > 0.$$

From now on, $n > n_0$ will be understood. Now

$$\begin{aligned} h(\mathbf{M}_{n+1}) &\geqslant c_n h(\mathbf{M}_{n+}) \\ &\geqslant c_n(1 - \alpha_n) h(\mathbf{M}_n) \\ &\geqslant \left[\Pi_{n_0}^{n} c_i\right]\left[\Pi_{n_0}^{n}(1 - \alpha_i)\right] h(\mathbf{M}_{n_0}) \\ &> \varepsilon_3\left[\Pi_0^{n_0 - 1} c_i\right]^{-1}\left[\Pi_{n_0}^{\infty}(1 - \alpha_i)\right] h(\mathbf{M}_{n_0}) \\ &= b h(\mathbf{M}_{n_0}) \end{aligned}$$

defining b. Note $b > 0$. Since $\mathbf{M}_{n_0}$ is nonsingular, we have that $\mathbf{M}_{n+1}$ and $\mathbf{M}_{n+}$ are in $\mathfrak{M}_1$, a convex compact subset of $\mathfrak{M}^+$ defined by $\mathfrak{M}_1 = \{\mathbf{M} \in \mathfrak{M}^+ | h(\mathbf{M}) > bh(\mathbf{M}_{n_0})\}$. Furthermore, there is a positive number C such that

$$\|\mathbf{M}' - \mathbf{M}\| \leqslant C, \quad \mathbf{M} \in \mathfrak{M}_1, \quad \mathbf{M}' \in \mathfrak{M} \Rightarrow \mathbf{M}' \in \mathfrak{M}^+.$$

Define

$$\mathfrak{M}_0 = \{\mathbf{M}' \in \mathfrak{M} \mid \|\mathbf{M}' - \mathbf{M}\| \leqslant C \text{ for some } \mathbf{M} \in \mathfrak{M}_1\}. \tag{2.9}$$

Then $\mathfrak{M}_0$ is a compact convex subset of $\mathfrak{M}^+$. By Lemma 2.1 of [5], there is a constant $U' > 0$ such that for all $\mathbf{M} \in \mathfrak{M}_0$ and all $\eta = \xi' - \xi$ for designs ξ' and ξ,

$$\begin{aligned} \partial^2 \Phi[\mathbf{M} + \alpha \mathbf{M}(\eta)]/\partial \alpha^2|_0 &\leqslant \|\mathbf{M}(\eta)\|^2 U' \\ &\leqslant (\operatorname{diam} \mathfrak{M})^2 U' \\ &= U, \end{aligned} \tag{2.10}$$

defining U.

Since $\mathbf{M}_n$ is in $\mathfrak{M}_1$, a sufficient condition for $\mathbf{M}_n + \alpha \mathbf{M}(\eta) \in \mathfrak{M}_0$ is that $\alpha \in A_0 = [0, \alpha_0]$, where α_0 is defined as the smaller of 1 and $C/(\operatorname{diam} \mathfrak{M})$. Therefore

$$\begin{aligned} \Phi(\mathbf{M}_{n+1}) - \Phi(\mathbf{M}_n) &\leqslant \Phi(\mathbf{M}_{n+}) - \Phi(\mathbf{M}_n) \\ &\leqslant a_1 a_2 \min_{\alpha \in A_n}\left[\Phi(\mathbf{M}_n + \alpha \mathbf{M}(\eta_n)) - \Phi(\mathbf{M}_n)\right] \\ &< a_1 a_2 \min_{\alpha \in A_n}\left[\Phi(\mathbf{M}_n + \alpha \mathbf{M}(\eta_n)) + \alpha \varepsilon_1 \bar{\varphi}_n - \Phi(\mathbf{M}_n)\right] \\ &= a_1 a_2 \min_{0 \leqslant \alpha \leqslant 1}\left[\Phi(\mathbf{M}_n + \alpha \mathbf{M}(\eta_n)) + \alpha \varepsilon_1 \bar{\varphi}_n - \Phi(\mathbf{M}_n)\right] \\ &\leqslant a_1 a_2 \min_{\alpha \in A_0}\left[\Phi(\mathbf{M}_n + \alpha \mathbf{M}(\eta_n)) + \alpha \varepsilon_1 \bar{\varphi}_n - \Phi(\mathbf{M}_n)\right] \end{aligned}$$

$$\leq a_1 a_2 \min_{\alpha \in A_0}\left[-\alpha(\varepsilon_2 - \varepsilon_1)\bar{\varphi}_n + \tfrac{1}{2}\alpha^2 U\right]$$

$$\leq a_1 a_2 \min_{\alpha \in A_0}\left[-\alpha(\varepsilon_2 - \varepsilon_1)\delta + \tfrac{1}{2}\alpha^2 U\right] \tag{2.11}$$

by (2.5), (2.4), (2.3), (2.10), and (2.8). But in (2.8), δ may be taken small enough so that $(\varepsilon_2 - \varepsilon_1)\delta/U$ is in A_0. In that case the minimum of the expression in brackets in (2.11) is $-(\varepsilon_2 - \varepsilon_1)^2\delta^2/2U < 0$. This implies that $\Phi(\mathbf{M}_n) \to -\infty$, which is impossible. □

COROLLARY 2.1. *Under Assumptions* 1–4, $\inf_{\Xi^+}\bar{\varphi}(\xi) = 0$.

Of course, for specific singular optimality criteria, the corollary follows from any of the iterative results for those criteria, e.g., [22], [3], [18]. Without a sequence of designs with $\bar{\varphi}(\xi_n) \to 0$, the corollary is not obvious even for D_s-optimality.

Corollary 2.1 allows us to prove an equivalence theorem for singular criteria.

THEOREM 2.2. *Under Assumptions* 1–4, *if* ξ^* *is nonsingular, then the three conditions below are equivalent.*

(i) ξ^* *is* Φ-*optimal*;

(ii) $\bar{\varphi}(\xi^*) = \inf_{\Xi^+}\bar{\varphi}(\xi)$;

(iii) $\bar{\varphi}(\xi^*) = 0$.

PROOF. Corollary 2.1 gives (ii) $\Leftrightarrow$ (iii). Inequality (2.7) shows (iii) $\Rightarrow$ (i). Since $\bar{\varphi}(\xi) \geq 0$, (i) $\Rightarrow$ (iii). □

Although the most useful part of the theorem, namely (i) $\Leftrightarrow$ (iii), is trivial, the implication (ii) $\Rightarrow$ (iii), is not self-evident without Corollary 2.1. The proofs of the D_s-optimality equivalence theorem given in [10] and [7] are incorrect at that point, since they assume, explicitly or implicitly, that $\inf_{\Xi^+}\bar{\varphi}(\xi)$ is attained in Ξ^+, which is false if there is no nonsingular optimal design.

Theorem 2.2 is an analogue of Kiefer's result [10] for D_s-optimality. One might attempt to define φ at singular ξ so that (iii) $\Leftrightarrow$ (i) without the restriction that ξ^* be nonsingular. Such an extension based simply on defining $\varphi(x, \xi)$ as a derivative of $\Phi\{\mathbf{M}[(1 - \alpha)\xi + \alpha\xi_x]\}$ is impossible, as is shown by Kiefer's example on page 309 of [10]. In that example $\mathbf{M}^*(\xi)$ is 1×1. If $\Phi(\mathbf{M})$ is any differentiable monotone decreasing function of $\mathbf{M}^*$, and the parameter b^2 is greater than 4, then ξ^* satisfies (iii) but not (i). This example, but now with $1 < b^2 < 4$, also provides a counterexample to the L-optimality equivalence theorem as stated in [7], Theorem 2.9.2, which incorrectly omits the restriction of nonsingularity. In that theorem φ is defined at singular ξ by formal analogy to the expression when ξ is nonsingular, using the Moore-Penrose inverse, rather than as a directional derivative.

To date, correct theorems which do not need the restriction that ξ^* be nonsingular ([11], and [9] as modified in [2]) have defined the needed function φ at singular ξ in complicated ways which are not written explicitly. For related further comments see Kiefer [12] Sections 3K and 7, and Silvey [13a].

3. Techniques for changing design weights and support points. Theorem 2.1 gave a class of sequences such that $\Phi[\mathbf{M}(\xi_n)] \to \inf_{\mathfrak{M}^+} \Phi(\mathbf{M})$. The present section gives explicitly some sequences in the class.

A. *Vertex direction methods.* As was mentioned just after Algorithm 1, $\{\eta_n\}$ satisfies the necessary conditions if $\eta_n = \xi_{x_n} - \xi_n$, where x_n maximizes $\varphi(x, \xi_n)$. This has been called the steepest descent method ([19] and [5]). Wu has pointed out that φ is a derivative w.r.t. α rather than w.r.t. actual distance traveled in $\mathfrak{M}$, so he proposes "vertex direction" as a more appropriate term than "steepest descent".

As in [3], an elementary modification is to consider removing mass at that y_n which minimizes $\varphi(x, \xi_n)$ over the support of ξ_n. This can be fitted into Algorithm 1, with $\alpha > 0$, by letting

$$\begin{aligned} \xi'(x) &= \xi_n(x)/[1 - \xi_n(y_n)] \qquad \text{for} \quad x \neq y_n \\ \xi'(y_n) &= 0. \end{aligned}$$

Then define $\eta_n' = \xi' - \xi_n$. It is not hard to verify that

$$\partial\Phi[\mathbf{M}(\xi_n + \alpha\eta_n')]/\partial\alpha|_0 = -\varphi(y_n, \xi_n)\xi(y_n)/[1 - \xi_n(y_n)]. \tag{3.1}$$

Condition (2.3) is not automatically satisfied by η_n'; it must be checked. If (2.3) is satisfied then either η_n or η_n' may be used. One may choose between them by computing both improvements, or by estimating these improvements with second derivatives. If the decision is not too time consuming, this may accelerate convergence.

If $\xi_n(y_n)$ is very small, so that η_n' does not satisfy (2.3), the mass at y_n can still be removed under Step 3 of Algorithm 1, as long as (2.6) is satisfied. However, it is undesirable to do this often, since $\Phi(\mathbf{M})$ only improves slightly each time.

B. *Gradient and conjugate gradient methods.* From now on let $\mathfrak{X}_n$ be a finite set which contains x_n and the support of ξ_n. For ξ supported on $\mathfrak{X}_n$, treat $\Phi[\mathbf{M}(\xi)]$ as a function of m variables, the m weights $w_i = \xi(y_i)$ for $y_i \in \mathfrak{X}_n$. Let $\mathbf{w}$ and $\mathbf{u}$ denote the m-dimensional vectors with components $\xi(y_i)$ and $\eta(y_i)$. (We use the notation $\mathbf{w}$, rather than ξ, to emphasize that only the weights are considered variable. In Section 3E $\mathbf{x}$, the vector of support points of ξ, will also be allowed to vary.) Let $\mathbf{g}$ be the gradient vector, defined as the m-dimensional vector with ith element

$$\partial\Phi[\mathbf{M}(\xi_n + \alpha\xi_{y_i})]/\partial\alpha|_0.$$

One cannot use $-\mathbf{g}$ as an increment, since it does not satisfy the constraint that its components sum to zero. However, the projection onto the constrained space could be used, letting $-\mathbf{u}$ equal

$$\mathbf{g} - (m^{-1}\mathbf{g}^T\mathbf{e})\mathbf{e} \tag{3.2}$$

where $\mathbf{e}$ consists of m 1's. This is similar to an algorithm in [17].

Wu [20] considers only nonsingular criteria. Having defined $\mathfrak{X}_n$, he uses a vertex direction η in Step 1 of Algorithm 1. Then he suggests using Step 3 to adjust further the weights on $\mathfrak{X}_n$, and advocates a conjugate gradient projection iterative technique to do this. Each $\mathbf{u}$ is a linear combination of the projected gradient (3.2) and the previous $\mathbf{u}$. He continues iterating within Step 3 until $u_0\mathbf{u}^T\mathbf{g}$, is small, where u_0 is defined as the largest number such that $\mathbf{w} + u_0\mathbf{u}$ has all m components nonnegative. See [20] for details.

We mention the method here in order to point out that a restriction of the method can be used in Theorem 2.1, with possibly singular Φ satisfying only Assumptions 1–4. The restriction is that ξ_{n+} can be modified by conjugate gradient iterations as long as the resulting ξ_{n+1} satisfies (2.6). Roughly, $\mathbf{M}_{n+1}$ must not get too close to singular before we move on to Step 4.

C. *Quadratic increment.* In Sections C and D, Assumption 5 will be used.

Let us follow the method of [5] and approximate $\Phi[\mathbf{M}(\xi)]$ by a quadratic Taylor expansion about $\mathbf{w}$, and choose $\mathbf{u}$ to minimize the expansion. Let $\mathcal{X}_n$ and $\mathbf{g}$ be as in 3B, and let the Hessian $\mathbf{H}$ be the $m \times m$ symmetric matrix with ijth element

$$\partial^2\Phi\left[\mathbf{M}(\xi_n + \alpha\xi_{y_i} + \beta\xi_{y_j})\right]/\partial\alpha\,\partial\beta|_{\alpha=0,\beta=0} \tag{3.3}$$

for y_i and y_j in $\mathcal{X}_n$. The Taylor expansion to be minimized is

$$\Phi\left[\mathbf{M}(\xi_n)\right] + \mathbf{u}^T\mathbf{g} + \tfrac{1}{2}\mathbf{u}^T\mathbf{H}\mathbf{u}. \tag{3.4}$$

Expression (3.4) is to be minimized subject to $\Sigma u_i = 0$, i.e., $\mathbf{u}^T\mathbf{e} = 0$, where $\mathbf{e}$ is the column vector consisting of m 1's. Let λ be a Lagrange multiplier and set the usual derivatives equal to zero.

$$\mathbf{g} + \mathbf{H}\mathbf{u} - \lambda\mathbf{e} = 0 \tag{3.5}$$

$$\mathbf{u}^T\mathbf{e} = 0. \tag{3.6}$$

Since $\mathbf{H}$ is not invertible we must invoke, for the first time, Assumption 5:

$$\mathbf{v}^T\mathbf{H}\mathbf{v} = 0 \Rightarrow \mathbf{v}^T\mathbf{g} = 0. \tag{3.7}$$

Therefore, if $C(\mathbf{H})$ is the column space of $\mathbf{H}$ and $\mathbf{v}$ is arbitrary

$$\begin{aligned} \mathbf{v} \perp C(\mathbf{H}) &\Rightarrow \mathbf{v}^T\mathbf{H} = \mathbf{0}^T \\ &\Rightarrow \mathbf{v} \perp \mathbf{g} \qquad \text{by (3.7).} \end{aligned}$$

Therefore, $\mathbf{g} \in C(\mathbf{H})$. So if λ satisfies (3.5), then

$$\lambda\mathbf{e} \in C(\mathbf{H}). \tag{3.8}$$

There are two cases.

CASE 1. $\mathbf{e} \in C(\mathbf{H})$. The solution is formally written as

$$\begin{aligned} \mathbf{u} &= \mathbf{H}^-(-\mathbf{g} + \lambda\mathbf{e}) \\ \lambda &= (\mathbf{e}^T\mathbf{H}^-\mathbf{g})/(\mathbf{e}^T\mathbf{H}^-\mathbf{e}) \end{aligned}$$

for any g-inverse $\mathbf{H}^-$. (See Rao [13] 1b.5.ii.) The component of $\mathbf{u}$ which is in $C(\mathbf{H})$ is unique.

CASE 2. $\mathbf{e} \notin C(\mathbf{H})$. By (3.8), λ must be 0. The resulting (3.5) and (3.6) have the solution

$$\mathbf{u} = -\mathbf{H}^{-}\mathbf{g} + (\mathbf{H}^{-}\mathbf{H} - \mathbf{I})\mathbf{v}$$

where $\mathbf{v}$ is chosen such that $\mathbf{u}^T\mathbf{e} = 0$.

In [5], page 1127 below (2.4), it was incorrectly claimed that if Φ is nonsingular, $\mathbf{H}$ must be nonsingular. (The notation of [5] uses $\boldsymbol{\eta}$ for our $\mathbf{u}$.) If $\mathbf{H}$ is singular in [5], then the explicit formula for $\boldsymbol{\eta}$ must be changed to that given here for $\mathbf{u}$. This does not affect the remainder of [5].

If a quadratic increment $\mathbf{u}$ is to be used in Step 1 of Algorithm 1, it must be modified slightly to satisfy the required conditions. Let u_0 be the largest number such that $\mathbf{w} + u_0\mathbf{u}$ has all m components nonnegative, where $\mathbf{w}$ is the vector of weights of ξ_n, and $\mathbf{u}$ is the solution of (3.5) and (3.6). Define $\mathbf{u}_n = u_0\mathbf{u}$ and let η_n have weights $\mathbf{u}_n$. Then $\eta_n = \xi' - \xi_n$ for some design ξ', as assumed in Step 1 of the algorithm. Condition (2.3) must now be checked. If (2.3) is not satisfied then use of the quadratic increment in Step 1 is not covered by Theorem 1. Some other increment must be used, such as a vertex direction increment or a weighted average of a vertex direction increment and a quadratic increment.

Quadratic increments may also be used in Step 3 rather than Step 1. It may then happen that the ith component u_i is negative when w_i is very small (or zero). In this case one could minimize (3.4) subject to the constraints $\mathbf{u}^T\mathbf{e} = 0$ and $u_i = -w_i$. The mechanics are similar to what has been done above, but with one fewer variable. If then a search for suitable α is performed, it should be among increments with components αu_j for $j \neq i$, $u_i = -w_i$. In all of this, the conditions (2.6) and $\Phi[\mathbf{M}(\xi_{n+1})] \leqslant \Phi[\mathbf{M}(\xi_{n+})]$ must be checked. If they cannot be satisfied, we must leave Step 3.

D. *Diagonalized quadratic increment.* Computation of the quadratic increment may be so time consuming that the method is not economical. A modification is now given which is much simpler to compute but which still changes ξ_n at many points simultaneously.

Rather than using the full Hessian matrix $\mathbf{H}$, use only the diagonal terms. That is, define the $m \times m$ matrix $\mathbf{J}$ with $J_{ij} = 0$ for $i \neq j$ and

$$J_{ii} = \partial^2\Phi\left[\mathbf{M}(\xi_n + \alpha\xi_{y_i})\right]/\partial\alpha^2|_0.$$

Then find $\mathbf{u}$ to minimize

$$\Phi\left[\mathbf{M}(\xi_n)\right] + \mathbf{u}^T\mathbf{g} + \tfrac{1}{2}\mathbf{u}^T\mathbf{J}\mathbf{u}. \tag{3.9}$$

By Assumption 4, $\mathbf{J}$ is nonnegative definite. If $J_{ii} = 0$, consider $\mathbf{v}$ which has all components zero except for v_i. For this $\mathbf{v}$, $\mathbf{v}^T\mathbf{H}\mathbf{v} = 0$, so Assumption 5 implies that $\mathbf{v}^T\mathbf{g} = 0$, i.e., $g_i = 0$. Therefore, $\mathbf{g} \in C(\mathbf{J})$. Therefore (3.9) is minimized just as (3.4) is, but now the solution can be written explicitly. Note that $\mathbf{e} \in C(\mathbf{J})$ iff $\mathbf{J}$ is nonsingular. So if $\mathbf{J}$ is nonsingular the solution is

$$\eta_i = (-g_i + \lambda)/J_{ii}$$
$$\lambda = (\Sigma g_i J_{ii}^{-1})/\Sigma J_{ii}^{-1}.$$

If $\mathbf{J}$ is singular the solution is

$$\eta_i = -g_i/J_{ii} \quad \text{if} \quad J_{ii} \neq 0$$

and the remaining components of $\boldsymbol{\eta}$ are adjusted so that $\Sigma\eta_i = 0$.

The same restrictions as discussed in the last two paragraphs of Section 3C must be checked and obeyed.

Call this the "diagonalized quadratic" method. It is similar in appearance to Algorithm II suggested by Titterington [17], and to others suggested there and in [14], although the motivation seems to be different.

It is instructive to compare the gradient projection, quadratic and diagonalized quadratic increments from a geometric viewpoint. The gradient projection (3.2) reflects the geometry of the simplex rather than that of $\mathfrak{M}$. The vector $\mathbf{w}$ lies in a simplex, and every support point y_i corresponds to a vertex of this simplex, regardless of how close $\mathbf{M}(\xi_{y_i})$ is to $\mathbf{M}(\xi)$ or to other $\mathbf{M}(\xi_{y_j})$. The quadratic method reflects the geometry of $\mathfrak{M}$ much more closely. The mixed partial derivatives take into account the interaction of an $\mathbf{M}(\xi_{y_i})$ and $\mathbf{M}(\xi_{y_j})$ which are close to each other, while the second derivatives w.r.t. each variable help correct for the varying distances between $\mathbf{M}(\xi)$ and the different $\mathbf{M}(\xi_{y_i})$. The diagonalized quadratic method corrects for the distances between $\mathbf{M}(\xi)$ and $\mathbf{M}(\xi_{y_i})$ but not for the relationship between neighboring $\mathbf{M}(\xi_{y_i})$ and $\mathbf{M}(\xi_{y_j})$. Wu [20] uses gradient projections other than (3.2) which apparently are like the diagonalized quadratic method in this respect.

A comparison of these methods and the vertex direction method in terms of improvement in $\Phi(\mathbf{M})$ is not easy. In particular, it is not clear that the use of second derivatives buys greater improvements in $\Phi(\mathbf{M})$ than the use of first derivatives alone would give. The intuitive geometric argument of the preceding paragraph suggests that (at the cost of more computation time) the quadratic method should give a greater single step improvement in $\Phi(\mathbf{M})$, *if* $\Phi(\mathbf{M})$ behaves well. However, in a neighborhood of a singular optimal ξ^*, $\Phi[\mathbf{M}(\xi)]$ is not necessarily even continuous, much less quadratic.

E. *Sliding support points.* For simplicity of exposition, we assume here that $\mathfrak{X}$ is one-dimensional. In higher dimensions, appropriate generalizations will be apparent.

Let $x_1, \cdots, x_m$ be the support points of ξ. Let $z_1, \cdots, z_m$ be arbitrary numbers such that $x_i + tz_i \in \mathfrak{X}$ for t small and positive. Assume that $\mathbf{f}(x_i + tz_i)$ is twice continuously differentiable for t small and positive, twice differentiable from the right at $t = 0$, and that these derivatives are continuous at $t = 0$. Let ξ_t be the design obtained by putting (fixed) weight $\xi(x_i) = w_i$ at $x_i + tz_i$ rather than at x_i.

Then

$$\mathbf{M}(\xi_t) = \Sigma_i \mathbf{f}(x_i + tz_i)\mathbf{f}^T(x_i + tz_i)w_i.$$

Let $\dot{\varphi}(x, \xi)$, $\ddot{\varphi}(x, \xi)$, $\dot{f}_j(x)$ and $\ddot{f}_j(x)$ denote the first and second derivatives of φ and f_j w.r.t. x. If the derivatives are one-sided, the side will be understood from context. Let m_{ij} be the ijth element of $\mathbf{M}$.

THEOREM 3.1. *Under Assumptions* 1–4 *and the above assumptions on* $\{z_i\}$ *and* $\mathbf{f}$,

$$\partial\Phi[\mathbf{M}(\xi_t)]/\partial t|_0 = -\Sigma_i \dot{\varphi}(x_i, \xi)z_i w_i,$$

$$\partial^2\Phi[\mathbf{M}(\xi_t)]/\partial t^2|_0 \geq -\Sigma_i \ddot{\varphi}(x_i, \xi)z_i^2 w_i.$$

PROOF. Define

$$\Phi_{jk}(\mathbf{M}) = \partial\Phi(\mathbf{M})/\partial m_{jk},$$

$$\Phi_{jk,\,uv}(\mathbf{M}) = \partial^2\Phi(\mathbf{M})/\partial m_{jk}\,\partial m_{uv},$$

$$a_{jk}(t) = \Sigma_i f_j(x_i + tz_i)\dot{f}_k(x_i + tz_i)z_i w_i.$$

It is direct to show that

$$\varphi(x, \xi) = -\Sigma_{jk}\Phi_{jk}[\mathbf{M}(\xi)][f_j(x)f_k(x) - m_{jk}(\xi)],$$

from which it follows that

$$\begin{aligned}\partial\Phi[\mathbf{M}(\xi_t)]/\partial t &= 2\Sigma_{jk}\Phi_{jk}[\mathbf{M}(\xi_t)]a_{jk}(t)\\ &= -\Sigma_i\dot{\varphi}(x_i + tz_i, \xi_t)z_i w_i.\end{aligned}$$

Also,

$$\partial^2\Phi[\mathbf{M}(\xi_t)]/\partial t^2 = 4\Sigma_{jk,\,uv}\Phi_{jk,\,uv}[\mathbf{M}(\xi_t)]a_{jk}a_{uv} + 2\Sigma_{jk}\Phi_{jk}[\mathbf{M}(\xi_t)]\ \partial a_{jk}(t)/\partial t.$$

The first term on the right is nonnegative, by the convexity of Φ in $\mathbf{M}$, and the second term equals

$$-\ \Sigma_i\ddot{\varphi}(x_i + tz_i, \xi_t)z_i^2 w_i. \qquad \square$$

This theorem says that if $\varphi(x, \xi)$ is not level at an interior support point, then ξ can be improved by moving that support point in the direction in which φ increases. It would be natural to conjecture the stronger result that if $\varphi(x, \xi)$ is not locally maximized at a support point, then ξ can be improved by moving that support point in a direction of increasing φ. A counterexample to this conjecture is given by linear regression on $\mathcal{X} = [0, 1]$, with $s = 1$ and D_s-optimality. If ξ puts fixed weight α at some $x > 0$ and weight $1 - \alpha$ at 0, then $\varphi(y, \xi) = d_s(y, \xi) - 1 = (1 - \alpha)^{-1}x^{-2}(y - x)^2 - 1$, which is minimized at x. However, as mentioned in Section 1, moving x slightly in either direction does not change $\Phi[\mathbf{M}(\xi)]$ at all, since $\mathbf{M}^*(\xi) = 1 - \alpha$ for all $x > 0$. Whether the conjecture is true for nonsingular criteria remains an open question.

Theorem 3.1 may be used in Step 3 of Algorithm 1. Having chosen weights $w_1, \cdots, w_m$ in Step 1, now consider $\Phi[\mathbf{M}(\xi)]$ as a function of the support points

$x_1, \cdots, x_m$. The only constraint on the x_i's is that they must remain in $\mathcal{X}$. Then all the techniques of optimization theory are available to suggest how to move the support points. As a function of the x_i's, $\Phi(\mathbf{M})$ is not convex, but if $\{x_i, \cdots, x_{m'}\}$ is the set of points which are actually moved, e.g., the interior points, and if $\varphi(x_i, \xi)$ has negative second derivative at these support points, then $\Phi(\mathbf{M})$ is locally convex as the points move in some neighborhood.

If, in Step 3, ξ_{n+1} is obtained by moving the support points of ξ_{n+} so that each support point approximately gives either a boundary point local maximum or an interior level point of $\varphi(x, \xi_{n+1})$, then on the next use of Step 1 no new support points need to be added which are close to current support points, and clusters of support points do not form.

As always, adjustments of ξ_{n+} in Step 3 must satisfy the conditions $\Phi[\mathbf{M}(\xi_{n+1})] \leqslant \Phi[\mathbf{M}(\xi_{n+})]$ and (2.6). The first condition will generally not cause much trouble, but the second condition must be checked. Without (2.6) it would be possible to "improve" ξ_{n+} by sliding two support points into each other, thereby perhaps making ξ_{n+1} singular and terminating the iterations too soon.

As described above, changes in weights and changes in support points are done separately, in different steps of Algorithm 1. We now present one way to vary the weights and support points simultaneously.

On any iteration, consider all designs supported on some m-element set, and regard the resulting $\Phi[\mathbf{M}(\xi)]$ as a differentiable function of $2m$ variables, namely the weights and the support points. The tools of optimization theory may now be used to reduce $\Phi[\mathbf{M}(\xi)]$. In this form $\Phi[\mathbf{M}(\xi)]$ is not convex, so care is needed, and no specific details will be given here. However, the general method can be fitted into Algorithm 1 as follows.

Suppose some $\mathbf{M}_{n+1}$ has tentatively been obtained somehow from $\mathbf{M}_n$, e.g., by varying the weights and support points of ξ_n simultaneously (perhaps through a finite number of iterations). As the support points of ξ move, $\mathbf{M}(\xi)$ moves from $\mathbf{M}_n$ to $\mathbf{M}_{n+1}$ along a curved path, so the proof of Theorem 2.1 does not apply directly. But once $\mathbf{M}_{n+1}$ has been obtained, we can consider the straight path in $\mathfrak{M}$ from $\mathbf{M}_n$ to $\mathbf{M}_{n+1}$. That is, in Step 1 of the Algorithm set ξ' equal to the ξ_{n+1} which was tentatively obtained and set $\alpha_n = 1$. Then check conditions (2.3)–(2.5). If they are satisfied then ξ_{n+1} may be used for this iteration, and the sequence of such designs converges, by Theorem 2.1. If on any iteration the conditions (2.3)–(2.5) are not satisfied, then the ξ_{n+1} obtained by varying support points should not be used, and another ξ_{n+1}, e.g., obtained from a single vertex direction, should be used instead. In all of this, Step 3 can be skipped, i.e., set $\xi_{n+} = \xi_{n+1}$.

I think that if the optimal support points are not known, then any iterative algorithm must somehow treat the support points as movable if it is to be successful after the early iterations. The points must eventually be moved, and it seems inefficient to do this by adding new points and letting the changes in weights eventually eliminate the old points.

4. Barrier methods and compromise criteria. Nonsingular criteria are simpler than singular criteria, at least in theory, so a natural attack on the singular problem is to approximate the singular optimality criterion by a nonsingular criterion, minimize it, and show that the design obtained is approximately optimal for the original criterion. In fact, this did not work well in the one example I have tried; see Section 5, and for a fuller discussion see [4]. However, the theoretical basis is so simple that it is presented here, for possible consideration in other problems.

THEOREM 4.1. *Let Assumption* 1 *hold, let* Φ *be real valued on* $\mathfrak{M}^+$ *and convex and lower semicontinuous on* $\mathfrak{M}$. *Let* Φ_B *be real valued on* $\mathfrak{M}^+$ *and bounded below on* $\mathfrak{M}$. *Let* $\mathcal{G}$ *be any open set in* $\mathfrak{M}$ *which contains all* Φ*-optimal* $\mathbf{M}$. *Then for sufficiently small* $\varepsilon > 0$, $\Phi(\mathbf{M}) + \varepsilon\Phi_B(\mathbf{M})$ *is minimized only in* $\mathcal{G}$.

Usually Φ_B, the "barrier function", will be chosen so that $\Phi(\mathbf{M}) + \varepsilon\Phi_B(\mathbf{M}) \to \infty$ as $\mathbf{M}$ approaches a singular matrix in $\mathfrak{M}$. If Φ satisfies Assumptions 2–4, then (2.1) gives the required convexity and lower semicontinuity on $\mathfrak{M}$.

PROOF. To avoid trivialities, $\mathcal{G} \neq \mathfrak{M}$. Define $v = \inf \Phi(\mathbf{M})$. By the lower semicontinuity of Φ there is some $\delta > 0$ such that $\Phi(\mathbf{M}) \geq v + \delta$ on $\mathfrak{M} - \mathcal{G}$. And by the convexity of Φ there is some $\mathbf{M}' \in \mathfrak{M}^+$ with

$$\Phi(\mathbf{M}') < v + \delta/3.$$

Note that $M' \in \mathcal{G}$. Define $v_B = \inf \Phi_B(\mathbf{M})$. Let $\varepsilon > 0$ be small enough so that

$$\varepsilon\Phi_B(\mathbf{M}') \leq \delta/3$$

and

$$\varepsilon v_B \geq -\delta/3.$$

Then

$$\begin{aligned}\inf[\Phi(\mathbf{M}) + \varepsilon\Phi_B(\mathbf{M})] &\leq \Phi(\mathbf{M}') + \varepsilon\Phi_B(\mathbf{M}')\\ &< v + 2\delta/3.\end{aligned}$$

But for $\mathbf{M} \in \mathfrak{M} - \mathcal{G}$,

$$\begin{aligned}\Phi(\mathbf{M}) + \varepsilon\Phi_B(\mathbf{M}) &\geq v + \delta + \varepsilon v_B\\ &\geq v + 2\delta/3,\end{aligned}$$

yielding the conclusion. □

This is readily applied to D_s-optimality. It is well known that

$$|\mathbf{M}| = |\mathbf{M}^*| \cdot |\mathbf{M}_{22}|. \tag{4.1}$$

Suppose $\Phi(\mathbf{M}) = -\log|\mathbf{M}^*|$, and define $\Phi_B(\mathbf{M}) = -\log|\mathbf{M}_{22}|$. Then

$$\Phi_\varepsilon(\mathbf{M}) = -\log|\mathbf{M}^*| - \varepsilon \log|\mathbf{M}_{22}| \tag{4.2}$$

is finite only for nonsingular $\mathbf{M}$, by (4.1).

Suppose we want a design ξ^* which is close enough to D_s-optimal so that $\bar{d}_s(\xi^*) \leq s + \delta$, for some $\delta > 0$. The function φ_ε corresponding to Φ_ε is

$$\varphi_\varepsilon(x, \xi) = d_s(x, \xi) + \varepsilon d_r(x, \xi) - (s + \varepsilon r). \tag{4.3}$$

Here $d(x, \xi) = \mathbf{f}^T(x)\mathbf{M}^{-1}(\xi)\mathbf{f}(x)$, $d_r(x, \xi) = \mathbf{f}_2^T(x)\mathbf{M}_{22}^{-1}(\xi)\mathbf{f}_2(x)$ with $\mathbf{f}_2$ consisting of the last $r = k - s$ components of $\mathbf{f}$, and $d_s(x, \xi) = d(x, \xi) - d_r(x, \xi)$. If ε is chosen such that

$$0 < \varepsilon r \leqslant (1 - t)\delta$$

for some t between 0 and 1, and iterations to minimize Φ_ε are performed until

$$\bar{\varphi}_\varepsilon(\xi) = \max_x[d_s(x, \xi) + \varepsilon d_r(x, \xi) - (s + \varepsilon r)] \leqslant t\delta;$$

this guarantees that at all x

$$\begin{aligned} d_s(x, \xi) &\leqslant d_s(x, \xi) + \varepsilon d_r(x, \xi) \\ &\leqslant s + \varepsilon r + t\delta \\ &\leqslant s + \delta \end{aligned}$$

as desired. Moreover, for such a design and all x

$$d_r(x, \xi) \leqslant (s + \delta)/\varepsilon.$$

This gives an indication of how small $\mathbf{M}_{22}(\xi)$ can be, i.e., how ill conditioned $\mathbf{M}(\xi)$ can be.

Atkinson and Cox [1] propose Φ_ε as an optimality criterion in its own right, as a compromise between the goals of estimating $\boldsymbol{\theta}_1$ well and being able to estimate $\boldsymbol{\theta}_2$ at all.

Stigler [15] has considered constrained criteria, i.e., minimize $\Phi(\mathbf{M})$ subject to a constraint $\Phi_B(\mathbf{M}) \leqslant C$ for some function Φ_B and constant C. This is related to the above considerations by a trivial theorem:

THEOREM 4.2. *Let ε be positive. If $\mathbf{M}_0$ minimizes $\Phi(\mathbf{M}) + \varepsilon\Phi_B(\mathbf{M})$, and $\Phi_B(\mathbf{M}_0) = C$, then $\mathbf{M}_0$ minimizes $\Phi(\mathbf{M})$ subject to $\Phi_B(\mathbf{M}) \leqslant C$.*

5. Example. Consider the problem of estimating the first degree coefficient in fourth degree polynomial regression on $[-1, 1]$. Since the parameter to be estimated is a scalar, criteria based on minimizing the variance of the estimator coincide. We formulate the problem in terms of D_s-optimality with $s = 1$. Studden [16] showed that the optimal design is supported on ± 1 and $\pm .5$. For the optimal weights, see Table I.

TABLE I
Optimal design and design obtained by iteration

Design	Support Points	Corresponding Weights	$\|\mathbf{M}^*\|$	$\bar{d}_s$
ξ_0	± 1	.2	.05635	2.4359
	$\pm .6$	.2		
	0	.2		
ξ_4	± 1	.05749	.11109	1.0089
	$\pm .49987$	.44348		
	0	1.3×10^{-8}		
ξ_8	± 1	.05552	.11111	1.00009
	$\pm .49996$	.44448		
	0	1.3×10^{-8}		
Optimal	± 1	.0555 · · ·	.11111 · · ·	1
	$\pm .5$	.4444 · · ·		

A method of iteration was used, as described below, which resulted in always having designs ξ_n which were supported on exactly five points: 0, ± 1 and $\pm x_n$ for some x_n, $0 < x_n < 1$. As a result, clusters of support points did not form, and the dimensions of the Hessian matrix **H** remained small, only 5×5. Moreover, ξ_n had symmetrical weights, i.e., $\xi_n(x) = \xi_n(-x)$. This symmetry caused **H** to contain a row of zeros and a column of zeros, corresponding to the point $x = 0$. So in the terminology of Sec. 3C, $\mathbf{e} \notin C(\mathbf{M})$ and Case 2 always applied. Deletion of the zero row and column of **H** left a 4×4 nonsingular matrix, so (3.9) could be solved without g-inverses.

In this example the optimal design does not need more points than the initial design. Of course we cannot expect to be so fortunate in every example. However, it may often prove to be convenient to add extra support points only reluctantly.

Several initial designs were tried, all uniform on five symmetric points: 0, ± 1, and $\pm x_0$, where on the different runs x_0 was set to .3, .4, .6 and .7.

Step 1 of Algorithm 1 was used only to change the weights in the current design, not to move support points or add new support points. Three types of increments were considered: (1) add equal mass at the two symmetrically placed points where $d_s(x, \xi_n)$ is maximized (always either ± 1 or $\pm x_n$) and subtract mass proportionately from the other three points; (2) subtract mass at the point where $d_s(x, \xi_n)$ is minimized (this point is always 0) and add mass proportionately at the other four points; (3) use the quadratic increment of Section 3C. The improvement in $\Phi[\mathbf{M}(\xi)]$ was estimated for each of these three directions, based on the first two directional derivatives of $\Phi[\mathbf{M}(\xi)]$ at ξ_n, and the apparently best increment was used. On most iterations, this was type (1). If there had been more support points, then type (1) might not have looked as good compared to type (3).

The α_n to use with the chosen direction η_n was determined by a Newton's method line search, usually lasting two or three iterations.

On these test runs, constants ε_1, ε_2, ε_3, and a_1 were ignored, but examination of the output can give information on what values would have been appropriate for strict application of Algorithm 1. All the iterations of types (1) and (3) would have been performed as they in fact were if $\varepsilon_1 \leqslant .22$ and $\varepsilon_2 \leqslant .30$. The iterations of type (2), subtraction of mass at 0, would have been performed as they were if ε_1 and ε_2 were less than .0036. The line searches all would have been sufficient if $a_1 \leqslant .82$.

In Step 2, $a_2 = .9$ was used.

In Step 3, the points $\pm x_n$ were moved symmetrically to the points which the second derivative approximation indicated was optimal. The relevant formulas for the calculation of the derivatives are

$$
\begin{aligned}
(\partial/\partial x_i)\log|\mathbf{M}(\xi)| &= \xi(x_i)\dot{d} \\
(\partial^2/\partial x_i^2)\log|\mathbf{M}(\xi)| &= -2[\xi(x_i)]^2\left[(\dot{d}/2)^2 + \mathbf{f}^T\mathbf{M}^{-1}\ddot{\mathbf{f}}d\right] + \xi(x_i)\ddot{d} \\
(\partial/\partial x_i)(\partial/\partial x_j)\log|\mathbf{M}(\xi)| &= -\xi(x_i)\xi(x_j)\left[2d_{01}(x_i, x_j)d_{10}(x_i, x_j)\right. \\
&\qquad \left. + 2d_{11}(x_i, x_j)d_{00}(x_i, x_j)\right]
\end{aligned}
$$

where throughout

$$\begin{aligned}
\mathbf{M} &= \mathbf{M}(\xi) \\
\dot{d} &= (\partial/\partial x)|_{x_i}\, d(x, \xi) \\
\ddot{d} &= (\partial^2/\partial x^2)|_{x_i}\, d(x, \xi) \\
\dot{\mathbf{f}}(x) &= (\partial/\partial x)\mathbf{f}(x) \\
\dot{\mathbf{f}} &= \dot{\mathbf{f}}(x_i) \\
d_{00}(x_i, x_j) &= \mathbf{f}^T(x_i)\mathbf{M}^{-1}\mathbf{f}(x_j) \\
d_{01}(x_i, x_j) &= \mathbf{f}^T(x_i)\mathbf{M}^{-1}\dot{\mathbf{f}}(x_j) \\
d_{10}(x_i, x_j) &= \dot{\mathbf{f}}^T(x_i)\mathbf{M}^{-1}\mathbf{f}(x_j) \\
d_{11}(x_i, x_j) &= \dot{\mathbf{f}}^T(x_i)\mathbf{M}^{-1}\dot{\mathbf{f}}(x_j).
\end{aligned}$$

(To reduce clutter, the dependence of d, $\dot{d}$, and $\ddot{d}$ on x_i is not shown explicitly in the notation.) Therefore, if x_n is to be moved to $x_n + t$, and if $d_s(x, \xi_n)$ is concave at x_n, then the second derivative approximation of $\Phi[\mathbf{M}(\xi)]$ says, after some calculation, to use $t = -\dot{d}_s(x_n, \xi_n)/\ddot{d}_s(x_n, \xi_n)$, where dots denote differentiation of $d_s(x, \xi)$ w.r.t. x. To protect against possible wild changes, $|t|$ was never allowed to be greater than the arbitrary value .15.

The function $h(\mathbf{M})$ was taken to be $|\mathbf{M}|^{1/k}$. Moving the support points in Step 3 increased $h(\mathbf{M})$ as often as it reduced it, and never reduced $h(\mathbf{M})$ to less than .89 times its former value. In the four test runs performed, the reduction was appreciable only on the first few iterations.

Iterations were performed until $\bar{d}_s(\xi_n) < 1.00005$ or until the effect of numerical inaccuracy became noticeable on a CDC 7600—e.g., when a matrix could not be inverted or the calculated Φ no longer was convex. The process ran for from six to twelve iterations, depending on the initial design used. The sequence starting with support points 0, ± 1, and $\pm .6$ is typical. This sequence is summarized in Table I.

The method described near the end of Section 3E, whereby the locations of the support points and the corresponding weights are changed simultaneously, was also tried. The design sequences approached the optimal design, but in this example changing the weights and support points in separate steps seemed to work more efficiently.

Simple use of the barrier method does not work well in this example, for reasons discussed in [4]. One important reason is that φ_e, given by (4.3), has a local maximum at 0 which is very sensitive to changes in $\xi(0)$. Therefore, much effort is spent adjusting $\xi_n(0)$ up or down slightly. This is wasted effort, since the D_s-optimal design puts no mass at 0.

Acknowledgments. I am indebted to D. M. Titterington for helpful discussion from which the idea of a diagonalized quadratic increment arose, and to K. O.

Kortanek for organizing the conference session where we met. I also am very grateful to J. Kiefer, C. F. Wu and a referee for their perceptive, thorough and helpful comments on the first draft of this paper.

REFERENCES

[1] ATKINSON, A. C., and COX, D. R. (1974). Planning experiments for discriminating between models. *J. Roy. Statist. Soc. Ser. B* **36** 321–334.

[2] ATWOOD, C. L. (1969). Optimal and efficient designs of experiments. *Ann. Math. Statist.* **40** 1570–1602.

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

[4] ATWOOD, C. L. (1976). Computational considerations for convergence to an optimal design. In *Proc. 1976 Conf. Information Sciences and Systems* 345–349. Dept. Elec. Eng., Johns Hopkins Univ.

[5] ATWOOD, C. L. (1976). Convergent design sequences for sufficiently regular optimality criteria. *Ann. Statist.* **4** 1124–1138.

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

[7] FEDOROV, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

[8] GRIBIK, P. R., and KORTANEK, K. O. (1976). Equivalence theorems and cutting plane algorithms for a class of experimental design problems. *SIAM J. Appl. Math.* **32** 232–259.

[9] KARLIN, S. and STUDDEN, W. J. (1966). Optimal experimental designs. *Ann. Math. Statist.* **37** 783–815.

[10] KIEFER, J. (1961). Optimum designs in regression problems, II. *Ann. Math. Statist.* **32** 298–325.

[11] KIEFER, J. (1962). An extremum result. *Canad. J. Math.* **14** 597–601.

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

[13] RAO, C. R. (1973). *Linear Statistical Inference and Its Applications, 2nd Ed.* Wiley, New York.

[13a] SILVEY, S. D. (1978). Optimal design measures with singular information matrices. *Biometrika* **65** 553–559.

[14] SILVEY, S. D., TITTERINGTON, D. M., and TORSNEY, B. (1978). An algorithm for optimal designs on finite design space. *Comm. Statist. A* **7** 1379–1389.

[15] STIGLER, S. M. (1971). Optimal experimental design for polynomial regression. *J. Amer. Statist. Assoc.* **66** 311–318.

[16] STUDDEN, W. J. (1968). Optimal designs on Tchebyscheff points. *Ann. Math. Statist.* **39** 1435–1447.

[17] TITTERINGTON, D. M. (1976). Algorithms for computing D-optimal designs on a finite design space. In *Proc. 1976 Conf. Information Sciences and Systems*. 213–216. Dept. Elec. Eng., Johns Hopkins Univ.

[18] TSAY, J. Y. (1977). A convergence theorem in L-optimal design theory. *Ann. Statist.* **4** 790–794.

[19] WHITTLE, P. (1973). Some general points in the theory of optimal experimental design. *J. Roy. Statist. Soc. Ser. B* **35** 123–130.

[20] WU, C. F. (1978). Some iterative procedures for generating nonsingular optimal designs. *Comm. Statist. A* **7** 1399–1412.

[21] WU, C. F. (1978). Some algorithmic aspects of the theory of optimal designs. *Ann. Statist.* **6** 1286–1301.

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

EG & G IDAHO, INC.
P.O. BOX 1625
IDAHO FALLS, IDAHO
83415