The Annals of Statistics,
1983, Vol. 11, No. 4, 1060–1068

GENERAL DIFFERENTIAL AND LAGRANGIAN THEORY FOR OPTIMAL EXPERIMENTAL DESIGN

BY F. PUKELSHEIM AND D. M. TITTERINGTON

Universitat Freiburg im Breisgau and University of Glasgow

The problem of optimal experimental design for estimating parameters in linear regression models is placed in a general convex analysis setting. Duality results are obtained using two approaches, one based on subgradients and the other on Lagrangian theory. The subgradient concept is also used to derive a potentially useful equivalence theorm for establishing the optimality of a singular design and, finally, general versions of the original equivalence theorems of Kiefer and Wolfowitz (1960) are obtained.

1. Introduction. Theoretical work on the optimal design problem has taken two approaches. The original work of Kiefer and Wolfowitz (1960) was later revealed (Whittle, 1973, Kiefer, 1974, Silvey, 1980) to have appealing formulation in terms of optimality conditions based on directional derivatives. The other approach has involved duality theorems, based on Lagrangian theory (Silvey, 1972, Sibson, 1972, 1974, Silvey and Titterington, 1973) or on Fenchel's Duality Theorem (Malyutov, 1975, Pukelsheim, 1980).

It is in this last paper that the optimal design problem in its most general setting has been treated but it is certainly the case that the approach taken there is less intuitively appealing than the geometrically flavoured arguments based on directional derivatives and less familiar than that based on Lagrangian multipliers. The present paper provides a treatment of the general problem using the differential theory of convex analysis with, as the corner-stone, the concept of the *subgradient*, the direct generalization of directional derivative. A Lagrangian-based treatment of the general problem is also included.

In Section 2, the optimal design problem is posed as the maximization, over a compact convex subset of nonnegative definite matrices, of a real-valued function having certain properties, of which a crucial one is a type of concavity. Section 3 derives the key result alluded to above, the duality theorem of Pukelsheim (1980), in terms of subgradients. This new formulation more clearly brings out the interrelations between the duality theory and the original equivalence theorem of Kiefer and Wolfowitz. Introduction of subgradients can also be justified by considerations of numerical computation of optimal designs, because most algorithms in some way or other use gradients or directional derivatives. Section 4 provides an analysis of similar generality by the Lagrangian approach, which hitherto has only been worked through in detail for special cases.

A major advantage of the differential theory approaches is that they provide the formulation of a test of optimality. This test is not easy to apply when the optimal nonnegative definite matrix is singular, which can quite easily happen. Silvey (1978) has conjectured the equivalence of an alternative test of much more practical promise and he established one half of the equivalence. Section 5 uses the subgradient tool to give a complete proof of the equivalence.

Finally, in Section 6, general forms of the original Kiefer and Wolfowitz equivalence theorems are established, in terms of subgradients and in terms of directional derivatives. The generality of the results greatly helps one to avoid certain confusion experienced in a special case by Karlin and Studden (1966); see also Atwood (1969).

Received January 1983; revised May 1983.

AMS 1980 *subject classifications.* 62K05, 90C25

Key words and phrases. Optimal design, convex analysis, subgradient, Lagrange multipliers, duality.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

Two references, Rockafellar (1970) and Pukelsheim (1980), will be quoted often. For brevity, they will be referred to by [R] and [P]. Recently a different approach to the problem has been developed by Gaffke (1983).

2. The optimal design problem and its closure. Let $\mathcal{M}$ be a compact convex subset of NND(k), the set of all symmetric non-negative definite $k \times k$ matrices, and assume that $\mathcal{M}$ intersects PD(k), the set of all symmetric positive definite $k \times k$ matrices. Suppose K is a given $k \times s$ matrix of rank s, and denote by $\mathcal{A}(K)$ the set of those matrices A in NND(k) whose range (column space) contains the range of K. Define J to be the function from NND(k) into NND(s) which maps A into $(K'A^{-}K)^{-1}$ if A lies in $\mathcal{A}(K)$, and into 0 otherwise. Finally let j be an *information functional* on NND(s); that is, j is non-negative on NND(s) and positive on PD(s), positively homogeneous, and concave; see [P]. As a consequence, j is isotone with respect to the Loewner ordering $C \geq D \Leftrightarrow C - D \in$ NND(s). If j is strictly concave then it is also strictly isotone. The converse is not true, as illustrated by $j_I(C) = \operatorname{trace} C$. Moreover, not even $j_L(C)$ is strictly isotone, where $j_L(C) = \operatorname{trace} CL$, for $L \in$ NND(s), $L \neq 0$, rank $L < s$. Such functionals will be excluded from our discussion in Section 4.

What we shall call the optimal design problem is the following.

$$\text{(P)} \qquad \text{Maximize } j \circ J(M), \quad \text{subject to} \quad M \in \mathcal{M}.$$

Any optimal solution M of the problem (P) will be said to have $\mathcal{M}$-maximal j-information for $K'\beta$. The number $v = \sup_{M \in \mathcal{M}} j \circ J(M)$ will be called the optimal value of (P). This set-up has a familiar interpretation in terms of linear model theory. There, β is the k-dimensional parameter in the model, $K'\beta$ denotes the s linear combinations of β of interest, $\mathcal{M}$ will contain the information matrices corresponding to the designs under investigation, and $\mathcal{A}(K)$ will contain those members of $\mathcal{M}$ for which $K'\beta$ is identifiable under the corresponding design. The information functional identifies the optimal design criterion of interest.

There exist examples, however, where problem (P) does not admit any optimal solution. We shall consider, therefore, as well as problem (P), "its closure" in which the objective function is $\operatorname{cl}(j \circ J)$, the *closure of* $j \circ J$, which is the upper semi-continuous function given by

$$\operatorname{cl}(j \circ J)(A) = \lim_{\varepsilon \searrow 0} j \circ J(A + \varepsilon I_k),$$

([R], page 57); in particular, we have $\operatorname{cl}(j \circ J)(A) \geq j \circ J(A)$. Since on a compact set any upper semi-continuous function attains its maximum, the optimal value v satisfies

$$v = \max_{M \in \mathcal{M}} \operatorname{cl}(j \circ J)(M). \tag{1}$$

For $A \in \mathcal{A}(K)$ the matrix $J(A)$ is positive definite and hence lies in the interior of the effective domain of j. Continuity then implies $j \circ J(A) = \operatorname{cl}(j \circ J)(A)$. Hence the functions $j \circ J$ and $\operatorname{cl}(j \circ J)$ differ at most on the set NND(k) $- \mathcal{A}(K)$. Recall that $j \circ J$ itself is closed, i.e., $j \circ J = \operatorname{cl}(j \circ J)$, if and only if j vanishes outside PD(s) ([P], page 342).

Since problem (P) remains essentially unchanged under monotone transformation, we are at liberty to work with the more convenient concave functions

$$g = \log \circ j \circ J, \quad \operatorname{cl} g = \log \circ \operatorname{cl}(j \circ J).$$

In the sequel, let $R^{k \times k}$ and Sym(k) denote the spaces of all $k \times k$ matrices and of all symmetric $k \times k$ matrices, respectively.

3. Subgradient theory. A matrix $B \in R^{k \times k}$ will be called a *subgradient* of g at M if

$$g(A) \leq g(M) + \langle A - M, B \rangle, \quad \text{for all} \quad A \in R^{k \times k}, \tag{2}$$

where $\langle A, B \rangle = \operatorname{tr}(AB)$; ([R], pages 214, 308). Since g takes the value $-\infty$ outside Sym(k)

we need consider only matrices A and M, and hence also B, which are symmetric. If g is differentiable at M, the set $\partial g(M)$ of subgradients consists of a single member, the gradient vector $\nabla g(M)$ of g at M([R], page 242).

Now a major theorem of convex analysis provides the result that M has $\mathscr{M}$-maximal j-information for $K'\beta$ if and only if there exists a subgradient B of g at M such that

$$\langle A, B\rangle \le \langle M, B\rangle, \quad \text{for all} \quad A \in \mathscr{M}. \tag{3}$$

This theorem will take the place of the Fenchel Duality Theorem employed in ([P], page 346). That (3) is sufficient for optimality is almost self-evident by rearrangement of (2). A simple proof of necessity is outlined in ([R], page 271), and carried through in Bazaraa and Shetty (1979, page 95).

Next we obtain a useful characterization of the subgradients of the present problem. We work, however, with cl g and not with g. In addition we require the concept of the so-called *polar function* j^0 of the information functional j, defined on NND(s) by

$$j^0(D) = \inf\{\langle C, D\rangle / j(C) \mid C \in \mathrm{PD}(s)\}.$$

In fact, j^0 again is an information functional ([P], page 342) and $(j^0)^0 = \mathrm{cl}\, j$.

LEMMA 1. *Let $M \in \mathrm{Sym}(k)$ be such that $(\mathrm{cl}\, g)(M) > -\infty$, and let $B \in \mathrm{Sym}(k)$ be arbitrary. Then B is a subgradient of* cl g *at M if and only if $B \in \mathrm{NND}(k)$ and $1 = \langle M, B\rangle = \mathrm{cl}(j \circ J)(M) \cdot j^0(K'BK)$.*

PROOF. First we show that if $B \in \partial(\mathrm{cl}\, g)$ then $\langle M, B\rangle = 1$. For when in (2) we insert $A = \alpha M$, $\alpha > 0$, then

$$\langle M, B\rangle \le \inf_{\alpha>0}\{\alpha\langle M, B\rangle - \log \alpha\}.$$

Here the infimum is finite only if $\langle M, B\rangle > 0$, and then equals $1 + \log\langle M, B\rangle$. But $0 < \langle M, B\rangle \le 1 + \log\langle M, B\rangle$ forces $\langle M, B\rangle = 1$.

Now let $\mathscr{D}$ be the set of those matrices A with $(\mathrm{cl}\, g)(A) > -\infty$. Then (2) may be rewritten as

$$\langle M, B\rangle - (\mathrm{cl}\, g)(M) \le (\mathrm{cl}\, g)^*(B),$$

where $(\mathrm{cl}\, g)^*(B) = \inf_{A\in\mathscr{D}}\{\langle A, B\rangle - (\mathrm{cl}\, g)(A)\}$ is called the *conjugate function* of cl g. But $(\mathrm{cl}\, g)^* = g^*$, see ([R], pages 104, 308), while in ([P], page 346) it was shown that $g^*(B)$ equals $1 + \log j^0(K'BK)$ if $B \in \mathrm{NND}(k)$, and $-\infty$ otherwise. Thus $B \in \partial(\mathrm{cl}\, g)(M)$ if and only if $B \in \mathrm{NND}(k)$, $\langle M, B\rangle = 1$ and $-\log \mathrm{cl}(j \circ J)(M) \le \log j^0(K'BK)$. Since $\langle M, B\rangle = 1$, the last inequality is equivalent to $\langle M, B\rangle \le \mathrm{cl}(j \circ J)(M) \cdot j^0(K'BK)$. Since the reverse inequality holds quite generally, by the same arguments as in the proof of Theorem 3 in ([P], page 345), the lemma is proved.

We are now in a position to give an alternative proof of the central Theorem 4 on duality of ([P], page 345), in terms of subgradients.

THEOREM 1. *Let $\mathscr{N}$ be the set of those matrices $N \in \mathrm{NND}(k)$ which for all $M \in \mathscr{M}$ satisfy $\langle M, N\rangle \le 1$. Then*

$$\sup_{M\in\mathscr{M}} j \circ J(M) = \min_{N\in\mathscr{N}} 1/j^0(K'NK).$$

PROOF. Because of (1) we can replace the left side by $\max_{M\in\mathscr{M}} \mathrm{cl}(j \circ J)(M)$. But $M \in \mathscr{M}$ maximizes $\mathrm{cl}(j \circ J)$, or, equivalently, cl g, if and only if there exists a subgradient B of cl g at M satisfying (3). This B lies in $\mathscr{N}$, since $B \in \mathrm{NND}(k)$ and $\langle M, B\rangle = 1$, by Lemma 1. Thus $1/j^0(K'BK) = \mathrm{cl}(j \circ J)(M) \le 1/j^0(K'NK)$, for all $N \in \mathscr{N}$, where the equality is taken from Lemma 1 while the inequality follows as in Theorem 3 of ([P], 345). This completes the proof.

Thus subgradient theory allows all the results that were derived in [P] from Fenchel's Duality Theorem. Yet another approach is provided by Lagrangian duality and we turn to this next.

4. A general Lagrangian duality proof. The object of this section is to obtain the general result of Section 3 as a direct generalization of the Lagrangian theorems of Sibson (1972, 1974), and Silvey and Titterington (1973). It will be necessary to make the mild demand that the polar function j^0 be strictly isotone. As indicated in Section 2, this does rule out some information functionals, but few of much interest.

Our starting point is the second problem in the statement of Theorem 1, written in the form

$$\text{Minimize} \quad -\log j^0(K'NK),$$

$$\text{subject to} \quad N \in \mathrm{NND}(k), \quad \text{and} \quad \langle M, N\rangle \leq 1 \quad \text{for all} \quad M \in \mathit{M}.$$

Moreover, we shall replace M by a finite subset $\{A_1, \cdots, A_l\}$ of members of M. Generalization then goes exactly as in Sibson (1972, page 183) because, since M is convex, any $M \in \mathit{M}$ can be expressed as a convex combination of finitely many points of M. Thus we can express the problem as

$$\text{(4)} \qquad \text{Minimize} \quad -\log j^0(K'NK),$$

$$\text{subject to} \quad N \in \mathrm{NND}(k), \quad \text{and} \quad \langle A_i, N\rangle \leq 1 \quad \text{for all} \quad i = 1, \cdots, l.$$

As in Sibson (1972, page 183) the Lagrangian form associated with this problem turns out to be

$$\text{(5)} \qquad L(N, y) = -\log j^0(K'NK) + \langle M, N\rangle - \Sigma y_i, \quad \text{where} \quad M = \Sigma y_i A_i,$$

for $N \in \mathrm{NND}(k)$ and $y \in R_+^l$, i.e., $y = (y_1, \cdots, y_l)$ with $y_i \geq 0$ for all i. If, for fixed $N \in \mathrm{NND}(k)$, we take the supremum of $L(N, y)$ over $y \in R_+^l$ we obviously reproduce the objective function $-\log j^0(K'NK)$ of (4). The following lemma solves the dual task of minimizing over N with y fixed.

LEMMA 2. *Let j^0 be strictly isotone, and fix $y \in R_+^l$. Then* $\inf_{N\in\mathrm{NND}(k)} L(N, y) = \log \circ j \circ J(\Sigma y_i A_i) + 1 - \Sigma y_i$.

PROOF. Again set $M = \Sigma y_i A_i$. First consider the case $M \notin \mathscr{A}(K)$. Equivalently, e nullspace of M is not contained in the nullspace of K', whence there exists a vector u with $Mu = 0$ and $K'u \neq 0$. Observing that $K'K$ has full rank s and that $j^0(K'K) > 0$ we find $\inf_{N\geq 0} L(N, y) \leq \inf_{\alpha>0} L(I_k + \alpha uu', y) = -\infty$, since j^0 is assumed strictly isotone. On the other hand $M \notin \mathscr{A}(K)$ implies $J(M) = 0$ and $\log \circ j \circ J(M) = -\infty$. In the case $M \in \mathscr{A}(K)$ the matrix $C = (K'M^-K)^{-1}$ exists and, as in ([P], page 345), satisfies

$$\text{(6)} \qquad \langle M, N\rangle \geq \langle C, K'NK\rangle.$$

We can write (5) as $L(N, y) = L_1(N) + L_2(N) - \Sigma y_i$, where

$$L_1(N) = -\log j^0(K'NK) + \langle C, K'NK\rangle, \quad L_2(N) = -\langle C, K'NK\rangle + \langle M, N\rangle \geq 0.$$

We minimize L_1 and L_2 separately. Clearly we may restrict attention to those N for which $j^0(K'NK) > 0$, whence $0 < j(C) \cdot j^0(K'NK) \leq \langle C, K'NK\rangle$, by the definition of j^0. Then

$$\text{(7)} \qquad L_1(N) \geq \log j(C) - \log\langle C, K'NK\rangle + \langle C, K'NK\rangle \geq \log j(C) + 1,$$

where the second inequality follows from $\inf_{\alpha>0}\{\alpha - \log \alpha\} = 1$. Thus equality holds in (7) if and only if $j(C) \cdot j^0(K'NK) = \langle C, K'NK\rangle = 1$. While (7) proves $\inf_{N\geq 0} L_1(N)$

$\geq 1 + \log j(C)$, the converse inequality follows from

$$\begin{aligned}\inf_{N\geq 0} L_1(N) &= \inf_{N\geq 0} L_1(N/\langle C, K'NK\rangle)\\ &= 1 + \inf_{N\geq 0}\log\{\langle C, K'NK\rangle / j^0(K'NK)\}\\ &\leq 1 + \log \inf_{D\in \mathrm{PD}(s)}\{\langle C, D\rangle / j^0(D)\}.\end{aligned}$$

By definition, the last term equals $\log j^{00}(C)$, and since C is positive definite this is the same as $\log j(C)$. We have proved now that $\inf_{N\geq 0} L_1(N) = 1 + \log j(C)$, and that every minimizing sequence $(N_i)_{i=1,2,\ldots}$ forces $\lim_i \langle C, K'N_iK\rangle = 1$.

That it is possible to minimize, simultaneously, $L_2(N)$ and $L_1(N)$ is proved essentially by Sibson (1974). Let

$$L_3(D) = -\log j^0(D) + \langle C, D\rangle,$$

for $D \in \mathrm{PD}(S)$ and let D^* minimize $L_3(D)$. The expression at the foot of page 688 of Sibson (1974) is equivalent to $-L_2(N)$ and a matrix which achieves $L_2(N) = 0$, as well as minimizing $L_1(N)$, is given by Sibson's N_0 (see near the top of page 689) but with D^* in place of his $[\gamma_A(M)]^{-1}$.

Thus, altogether, $\inf_{N\geq 0} L(N, y) = \log j(C) + 1 - \Sigma y_i$, and the lemma is proved.

It follows that the Lagrangian dual problem of (4) is

$$\text{(8)} \qquad \text{Maximize} \quad \log \circ\, j \circ J(\Sigma y_i A_i) + 1 - \Sigma y_i, \qquad \text{subject to} \quad y \in R^l_+.$$

As a result of the homogeneity of j, this is equivalent to

$$\text{Maximize} \quad j \circ J(M), \qquad \text{subject to} \quad M \in \mathrm{conv}\{A_1, \cdots, A_l\}.$$

The fact that problems (4) and (8) have the same optimal values follows, by the concavity of $\log j^0(K' \cdot K)$ and the linearity of the constraints in (4), from the Strong Lagrangian Principle; see Whittle (1971, page 65).

The stages of the proof follow those of previous Lagrangian theorems. The decomposition of the Lagrangian into L_1 and L_2 and the inequality in (6) appear in the definition of the axis of the thinnest cylinder in Silvey and Titterington (1973, equation 3.8), of the centre of the minimal ellipsoid in Titterington (1975), and in equations (7) of Sibson (1974, page 689). The steps in which L_1 is minimized appear in slightly different form in ([P], page 346) and serve, as there, to avoid differentiation-based arguments specific to D-optimality in the other papers. A problem similar to (4) is discussed by Rosen (1965), without development of the duality.

The work in this and the previous section characterized the solution of the optimal design problem in terms of that of another optimization problem. We now turn to a related but slightly different sort of optimality characterization.

5. Optimality of singular information matrices using complementary subspaces. The familiar characterizations of optimal information matrices and designs as stated by Kiefer and Wolfowitz (1960), Kiefer (1974), Fedorov (1972), Whittle (1973) and Silvey (1980) can all be expressed as: $M \in \mathscr{M} \cap \mathscr{A}(K)$ is optimal if and only if the directional derivative of g at M is non-positive in every possible direction:

$$g'(M; A - M) \leq 0, \quad \text{for all } A \in \mathscr{M}.$$

Here $g'(M; A - M)$ is the derivative in the direction $A - M$, as defined in Rockafellar (1970), page 213); it is denoted by $F(M, A)$ and called the Fréchet (directional) derivative in Silvey (1980, page 18). If g is differentiable at M then $g'(M; A - M) = \langle \nabla g(M), A - M\rangle$ and it suffices to verify non-positivity for all A from a set S whose convex hull equals $\mathscr{M}$. As emphasized by Silvey (1978, 1980) this latter result greatly facilitates the checking of a design for optimality; he also points out that it is not applicable if, as may happen

with some criteria, the optimal information matrix M is singular, so that differentiability does not obtain.

However, many interesting information functionals fail to be differentiable only at singular information matrices M. On the other hand, singularity of $M \in \mathscr{A}(K)$ allows a wide choice of the generalized inverse in $J(M) = (K'M^{-}K)^{-1}$. Silvey (1978) used this observation to formulate an optimality condition in terms of a non-singular matrix of the form $M + HH'$ which satisfies $J(M) = J(M + HH')$; he proved his condition to be sufficient for optimality, and conjectured necessity. We shall now show, in terms of subgradients, that Silvey's condition is indeed both necessary and sufficient for optimality.

To be precise, let the set $\mathscr{H}(M)$ consist of all $k \times (k - \text{rank } M)$ matrices H such that the range of M and the range of H together span R^k; if rank $M = k$ take $\mathscr{H}(M) = \{0\}$. For $H \in \mathscr{H}(M)$, $(M + HH')^{-1}$ is a positive definite generalized inverse of M and, in particular, $J(M) = J(M + HH')$ for all $M \in \mathscr{A}(K)$ and $H \in \mathscr{H}(M)$. Given that $g(M)$ is differentiable at non-singular matrices M, Silvey's (1978; 1980, page 27) condition for optimality of $M \in \mathscr{M} \cap \mathscr{A}(K)$ is that, for some $H \in \mathscr{H}(M)$,

$$g'(M + HH'; A - M - HH') \leq 0, \quad \text{for all} \quad A \in \mathrm{S}.$$

We shall, in fact, consider the following more general formulation that, for some $H \in \mathscr{H}(\mathrm{M})$, and some $B \in \partial g(M + HH')$,

$$\langle A, B \rangle \leq \langle M + HH', B \rangle \quad \text{for all} \quad A \in \mathscr{M}.$$

Silvey's form of the condition is clearly a special case since, given differentiability at non-singular matrices A, $\nabla g(A)$ is the only subgradient, $\langle \nabla g(A), B \rangle = g'(A; B)$, and $\mathscr{M}$ may be replaced by a spanning set S.

It is easy to see that the general form of Silvey's condition is sufficient for optimality; indeed, inequality (2) entails $g(A) \leq g(M + HH') = g(M)$, whence M maximizes g over $\mathscr{M}$. Before we establish necessity of Silvey's condition we prove the following relation between the subgradients of g at M and at $M + HH'$.

LEMMA 3. *Suppose* $M \in \mathscr{A}(K)$ *and* $H \in \mathscr{H}(M)$. *Then* (i) $\partial g(M + HH') \neq \varnothing$, *and every* $B \in \partial g(M + HH')$ *satisfies*: (ii) $BH = 0$, *and* (iii) $B \in \partial g(M)$.

PROOF. The matrix $M + HH'$ is positive definite and hence lies in the relative interior of the effective domain of g; Theorem 23.4 of ([R], page 217) then proves (i). Now fix $B \in \partial g(M + HH')$, Since $M + HH'$ is non-singular we can effectively assume closedness and apply Lemma 1. Introducing $C = J(M + HH') = J(M)$, we have

$$\begin{aligned} \langle M + HH', B \rangle &= \mathrm{cl}(j \circ J)(M + HH') \cdot j^0(K'BK) \\ &= j(C) \circ j^0(K'BK) \leq \langle M, B \rangle, \end{aligned}$$

the last inequality following from (6) and the definition of j^0. Hence $\langle HH', B \rangle = 0$ and, since $B \in \mathrm{NND}(k)$, by Lemma 1, this proves (ii). Finally, since B is a subgradient at $M + HH'$, $g(A) \leq g(M + HH') + \langle A - M, B \rangle - \langle HH', B \rangle$; but the last term vanishes and $g(M + HH') = g(M)$, so that B is a subgradient also at M. The lemma is proved.

As a consequence of Lemma 3 the objective function g turns out to be subdifferentiable at every point M of its effective domain $\mathscr{A}(K)$: $\partial g(M) \supset \partial g(M + HH') \neq \varnothing$. Note also that in the general form of Silvey's condition the term $\langle M + HH', B \rangle$ could be replaced by $\langle M, B \rangle$, because of Lemma 3(ii), or simply by 1, because of Lemma 1. We shall now give a proof of sufficiency.

THEOREM 2. *Suppose the matrix* $M \in \mathscr{M}$ *lies in* $\mathscr{A}(K)$. *Then* M *has* $\mathscr{M}$*-maximal j-information for* $K'\beta$ *if and only if there exist matrices* $H \in \mathscr{H}(M)$ *and* $B_K \in \partial g(M + HH')$ *such that* $\langle A, B_K \rangle \leq \langle M + HH', B_K \rangle$, *for all* $A \in \mathscr{M}$.

PROOF. It remains to establish the direct part. In view of (1) we may consider maximization of cl g instead of g. As a result of optimality there exists a subgradient B of cl g at M satisfying (3). Define $B_K = BK(K'BK)^+K'B$, then $\langle M, B_K\rangle \le \langle M, B\rangle$ follows as in the proof of Lemma 2 in ([P], page 348), while conversely $\langle M, B\rangle = \mathrm{cl}(j \circ J)(M) \cdot j^0(K'BK) \le \langle C, K'BK\rangle$, by Lemma 1 and the definition of j^0, and $\langle C, K'BK\rangle = \langle C, K'B_KK\rangle \le \langle M, B_K\rangle$, by the same argument as in the proof of Theorem 3 in ([P], page 345). Hence $1 = \langle M, B_K\rangle = \mathrm{cl}(j \circ J)(M) \cdot j^0(K'B_KK)$, so that also B_K is a subgradient of cl g at M, by Lemma 1. Furthermore, $\langle A, B_K\rangle \le \langle A, B\rangle \le \langle M, B\rangle = \langle M, B_K\rangle$, for all $A \in \mathscr{M}$, whence B_K satisfies the required inequalities.

Now on $\mathscr{A}(K)$ the function g coincides with its closure, whence

$$(\mathrm{cl}\, g)(M) = g(M) = g(M + HH'), \quad \text{for all} \quad H \in \mathscr{A}(M).$$

The construction in the proof of Lemma 2 in ([P], page 348) shows that we can find some $H \in \mathscr{A}(M)$ such that $B_KH = 0$. For such H then, altogether,

$$g(A) \le (\mathrm{cl}\, g)(A) \le (\mathrm{cl}\, g)(M) + \langle A - M, B_K\rangle = g(M + HH') + \langle A - M - HH', B_K\rangle,$$

for all $A \in R^{k\times k}$, which means $B_K \in \partial g(M + HH')$.

The final section turns to directional derivatives proper and derives, without any differentiability assumptions, equivalence theorems of the Kiefer-Wolfowitz type, as discussed by Kiefer and Wolfowitz (1960), Kiefer (1974), and Whittle (1973).

6. Equivalence theorems of Kiefer-Wolfowitz type. The original Equivalence Theorem of Kiefer and Wolfowitz (1960) established equivalence of two optimality criteria which are, independently of each other, of statistical interest: D-, and G-optimality. Their result will be generalized to information functionals j, in that maximization of $j \circ J$ over $\mathscr{M}$ turns out to be equivalent to minimization of a certain function d over $\mathscr{M}$. While such equivalences are useful in inspiring numerical algorithms, and provide interesting aspects of the interrelations with subgradients and directional derivatives, the function d will, in general, lose its appealing statistical interpretation, as is manifest in the equivalence theorems of Fedorov (1972), for instance.

To be specific, define the function d_g by

$$d_g(M) = \inf_{B\in\partial g(M)} \max_{A\in\mathscr{M}} \langle A, B\rangle \tag{9}$$

if $\partial g(M) \neq \emptyset$, and $d_g(M) = +\infty$ otherwise; compare equations (3.12) and (4.3) in Silvey and Titterington (1973, pages 27–28). We first show that for closed functions g the infimum in (9) actually is a minimum.

LEMMA 4. *If $\partial(\mathrm{cl}\, g)(M) \neq \emptyset$ then there exists a matrix $B \in \partial(\mathrm{cl}\, g)(M)$ such that $d_{\mathrm{cl}g}(M) = \max_{A\in\mathscr{M}}\langle A, B\rangle$.*

PROOF. The assertion may be obtained from Theorem 27.3 in ([R], page 267); a simple direct proof follows. Choose a sequence $B_i \in \partial(\mathrm{cl}\, g)(M)$ such that $f^*(B_i) = \max_{A\in\mathscr{M}}\langle A, B_i\rangle$ tends to $d_{\mathrm{cl}g}(M)$ as $i \to \infty$. With some positive definite matrix $A \in \mathscr{M}$ then $\| B_i\| \le \| B_i^{1/2}\|^2 \le \| A^{-1/2}\|^2 \| A^{1/2}B_i^{1/2}\|^2 \le (\mathit{trace}\, A^{-1})f^*(B_i)$. Therefore $(B_i)_{i\in N}$ is bounded. Let B be one of its points of accumulation. Then $B \in \partial(\mathrm{cl}\, g)(M)$, see ([R], page 215), and $d_{\mathrm{cl}g}(M) \le f^*(B) = f^*(B - B_i + B_i) \le \max_{A\in\mathscr{M}}\| A\| \| B - B_i\| + f^*(B_i) \to d_{\mathrm{cl}g}(M)$. It follows that $f^*(B) = d_{\mathrm{cl}g}(M)$.

THEOREM 3. *Suppose the matrix $M \in \mathscr{M}$ lies in $\mathscr{A}(K)$. Then the following three statements are equivalent:*

(a) *M has $\mathscr{M}$-maximal j-information for $K'\beta$.*

(b) *M minimizes $d_{\mathrm{cl}g}$ over $\mathscr{M}$.*

(c) $d_g(M) = 1$.

PROOF. Note first that $d_{\mathrm{cl}\,g}(A) = d_g(A)$, for all $A \in \mathscr{A}(K)$, since then $\partial(\mathrm{cl}\ g)(A) = \partial g(A)$. Secondly observe that $d_g(M) = d_{\mathrm{cl}\,g}(M) \geq 1$, as a consequence of Lemma 1.

(a) $\Leftrightarrow$ (c): M is optimal, that is, (a), if and only if there exists a subgradient of g at M which lies in the set $\mathscr{A}$ of Theorem 1, and this is (c).

(c) $\Rightarrow$ (b): Obvious from the second remark.

(b) $\Rightarrow$ (c): The closure of problem (P), i.e., maximization of cl g over $\mathscr{M}$, always admits an optimal solution M^*. For this matrix M^* the implication (a) $\Rightarrow$ (c) shows $d_{\mathrm{cl}\,g}(M^*) = 1$. Therefore the minimum value of $d_{\mathrm{cl}\,g}$ over $\mathscr{M}$ equals 1, and (b) implies (c). The proof is complete.

Now fix $M \in \mathscr{M} \cap \mathscr{A}(K)$. For all matrices $A \in \mathscr{A}(K)$ and $B \in \partial g(M)$ the same arguments as in the proof of Theorem 3 in ([P], page 345) show that $\langle A, B\rangle \geq j \circ J(A) \cdot j^0(K'BK)$. By Lemma 1, the last expression equals $j \circ J(A)/j \circ J(M)$ and does not depend on B. Hence if we first take the maximum over $A \in \mathscr{M}$ and then the infimum over $B \in \partial g(M)$ we obtain $d_g(M) \geq v/j \circ J(M) \geq 1$, or equivalently,

$$1/d_g(M) \leq j \circ J(M)/v \leq 1. \tag{10}$$

Thus $1/d_g(M)$ provides a lower bound for the efficiency factor $j \circ J(M)/v$ of M. This generalizes Theorem 4.3 of Atwood (1969, page 1596). Furthermore, it follows that for every sequence $(M_i)_{i\in N}$ in $\mathscr{M} \cap \mathscr{A}(K)$,

$$\lim_{i\to\infty} d_g(M_i) = 1 \Rightarrow \lim_{i\to\infty} j \circ J(M_i) = v.$$

The converse of this does not always hold. For instance if, in Example 2 of Pukelsheim (1981), we take $M_i = M + (1/i)I_2$, then $\lim j \circ J(M_i) = v$, but $\lim d_g(M_i) = 36/25 > 1$. Thus any algorithm based on minimizing d_g will produce optimizing sequences for problem (P), but may not catch all of them; compare the remark on G- and D-efficiency in Kiefer and Studden (1976, page 1118).

Inclusion of the closure of g in statement (b) of Theorem 3 relates to the problem of whether it is possible that

$$\inf_{M\in\mathscr{M}} d_g(M) > 1 = \min_{M\in\mathscr{M}} d_{\mathrm{cl}\,g}(M).$$

Obviously this cannot happen when g is closed, nor when problem (P) admits an optimal solution, as is shown by the proof of Theorem 3. Whittle's (1973) equivalence theorem does not answer this question either, since its proof does not cover the case when an optimal solution for (P) fails to exist.

Finally we note that Theorem 3 may be reformulated in terms of directional derivatives. Define $D_g(M) = \sup_{A\in\mathscr{M}} g'(M; A - M)$.

COROLLARY 3.1. *Suppose the matrix $M \in \mathscr{M}$ lies in $\mathscr{A}(K)$. Then the following three statements are equivalent:*

(a) *M has $\mathscr{M}$-maximal j-information for $K'\beta$.*

(b) *M minimizes $D_{\mathrm{cl}\,g}$ over $\mathscr{M}$.*

(c) $D_g(M) = 0$.

Its proof is quite similar to that of Theorem 3 and is therefore omitted.

REFERENCES

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

BAZARAA, M. S. and SHETTY, C. M. (1979). *Nonlinear Programming.* Wiley, New York.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments.* Academic, New York.

GAFFKE, N. (1983). Directional derivatives of optimality criteria at singular matrices in convex design theory. Unpublished manuscript.

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

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

KIEFER, J. and STUDDEN, W. J. (1976). Optimal designs for large degree polynomial regression. *Ann. Statist.* **4** 1113–1123.

KIEFER, J. and WOLFOWITZ, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.* **12** 363–366.

MALYUTOV, M. B. (1975). Note on the equivalence theorem. In *Optimum Design of Experiments*, Ed. M. B. Malyutov, 161–163. State University Press, Moscow.

PUKELSHEIM, F. (1980). On linear regression designs which maximize information. *J. Statist. Plann. Inference* **4** 339–364.

PUKELSHEIM, F. (1981). On c-optimal design measures. *Math. Operationsforsch. Statist., Ser. Statist.* **12** 13–20.

ROCKAFELLER, R. T. (1970). *Convex Analysis*. Princeton University Press.

ROSEN, J. B. (1965). Pattern recognition by convex programming. *J. Math. Anal. Appl.* **10** 123–134.

SIBSON R. (1972). Discussion of a paper by H. P. Wynn. *J. Roy. Statist. Soc. Ser. B* **34** 181–183.

SIBSON, R. (1974). D_A-optimality and duality. *Colloq. Math. Soc. Janos Bolyai* **9** *II*, 677–692. North Holland, Amsterdam.

SILVEY, S. D. (1972). Discussion of a paper by H. P. Wynn. *J. Roy. Statist. Soc. Ser. B* **34** 174–175.

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

SILVEY, S. D. (1980). *Optimal Design*. Chapman Hall, London.

SILVEY, S. D. and TITTERINGTON, D. M. (1973). A geometric approach to optimal design theory. *Biometrika* **60** 21–32.

TITTERINGTON, D. M. (1975). Optimal design: some geometrical aspects of D-optimality. *Biometrika* **62** 313–320.

WHITTLE, P. (1971). *Optimization under Constraints*. Wiley, New York.

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

LEHRSTUHL FÜR STOCHASTIK
UND IHRE ANWENDUNGEN
UNIVERSITÄT AUGSBURG
MEMMINGER STRASSE 6
D-8900 AUGSBURG
FEDERAL REPUBLIC OF GERMANY

DEPARTMENT OF STATISTICS
UNIVERSITY OF GLASGOW
UNIVERSITY GARDENS
GLASGOW G12 8QW
SCOTLAND