Statistica Sinica **7**(1997), 739-753

COMPUTING OPTIMAL DESIGNS BY BUNDLE TRUST METHODS

Adalbert Wilhelm

Universität Augsburg

Abstract: Essentially two classes of iterative procedures have been proposed in the literature to solve optimal design problems in linear regression: exchange algorithms devoted to the construction of optimal exact designs in a finite design space and methods from convex programming yielding optimal moment matrices only. By simultaneously taking weights and support points as variables the design problem represents a nonconcave, not necessarily differentiable, but Lipschitz continuous maximization problem. We, therefore, adapt bundle trust methods from nondifferentiable optimization to the design problem and show their numerical behaviour. Explicit efficiency bounds for the numerical solutions can be given in the case of a regression range with finitely many elements.

Key words and phrases: Approximate designs, bundle trust methods, nondifferentiability, pth matrix means, support points and weights.

1. Introduction

Algorithms for computing optimal experimental designs traditionally fall in one of the following two classes: the class of exchange algorithms for constructing optimal designs for a finite design space and the class of convex programming methods that only yield optimum moment matrices leaving the problem how to reconstruct a design from a given moment matrix unsolved.

The first versions of exchange algorithms were addressed to D-optimality and go back to Wynn (1970) and Fedorov (1972). Reviews of these and other exchange schemes are given by Cook and Nachtsheim (1980) and, more recently, by Nguyen and Miller (1992). A new approach to the construction of optimal exact designs using pattern search was proposed by Hardin and Sloane (1993). Gaffke and Mathar (1992) gave an excellent review of convex programming algorithms used in the design context.

In the present article we introduce a new algorithm for computing optimal or nearly optimal approximate experimental designs. This algorithm is based on the bundle trust method developed by Schramm and Zowe (1988). Bundle methods have become popular in nondifferentiable convex optimization in the 1980's (see Hiriart-Urruty and Lemaréchal (1993) for a thorough introduction).

The use of nondifferentiable optimization techniques in a general purpose routine for optimal experimental design is motivated by the following facts: there are only few restrictions in the choice of optimality criteria under consideration, so in particular, nondifferentiable techniques allow the investigation of E-optimality; nondifferentiability is connected with singular moment matrices that often arise when only subsystems of the unknown modelling parameter vector are of interest and the use of nondifferentiable methods allows us to keep the number of support points variable.

Whereas for theoretical investigations it is useful to formulate the design problem in terms of moment matrices or probability measures, the numerical treatment of the problem is best done by viewing designs as an array of support points and corresponding weights. This transition, however, calls for application of methods from nonconvex optimization since the moment matrix mapping is convex in the support points whereas the information function is a concave function of the moment matrix.

Our computer code OPTDES is based on the bundle trust codes BTCLC and BTNCLC of Schramm (1991); BTCLC is designed for the minimization of a convex objective function subject to linear constraints, BTNCLC for minimization of a nonconvex, but locally Lipschitz continuous objective function subject to linear constraints. In the convex case knowledge of one arbitrary subgradient at each iteration point is required, in the nonconvex case the concept of subdifferentials is replaced by the concept of generalized gradients in the sense of Rockafellar (1980) and Clarke (1983). The generalized gradient calculus for the design problem is given in Wilhelm (1995). There chain rule arguments are used to characterize the set of generalized gradients for various design optimality criteria.

In the implemented version OPTDES we restrict ourselves to the class of matrix means ϕ_p, the concave analogues to Kiefer's φ_p-criteria, introduced in Kiefer (1974). The subgradient representation for this class of information functions was found by Gaffke (1985) and is also given in Lemma 6.16 of Pukelsheim (1993). To obtain subgradient representations for other optimality criteria the results in Hoang and Seeger (1991) could be related with Wilhelm (1995).

After reviewing the design problem in Section 2, we give a short description of the bundle trust methodology and present its transformation to the design context in Section 3. Section 4 deals with the computation of Kiefer's φ_p-criteria and of the corresponding subgradients. In Section 5 we describe the special features of our approach available in the case of finitely many regression vectors. Although affirmative results on the convergence of bundle trust methods are available only for concave maximization problems, the numerical results given in Section 6 show that the algorithm is of practical interest.

2. The Design Problem

We consider the design problem in classical linear models; more details can be found in Pukelsheim (1993) for example. For each regression vector x in the regression range $\mathcal{X} \subseteq \mathbb{R}^k$ a random variable Y is observed, whose expectation depends linearly on x and on an unknown parameter vector $\theta \in \mathbb{R}^k$,

$$\mathrm{E}[Y] = x'\theta.$$

An (approximate) experimental design is a probability measure on the regression range $\mathcal{X}$ with finite support. It can be identified with its support points $x_1, \ldots, x_\ell \in \mathcal{X}$ and corresponding weights $\omega_1, \ldots, \omega_\ell > 0$ that sum to 1, where ℓ varies over the set of positive integers. We use the notations $(X; \omega)$ and $\left\{ \begin{smallmatrix} x_1, \ldots, x_\ell \\ \omega_1, \ldots, \omega_\ell \end{smallmatrix} \right\}$ for a design, where the $k \times \ell$ matrix X contains as ith column the regression vector x_i, while the ℓ dimensional weight vector ω contains the weights ω_i.

We assume that for the experiment given by a design $(X; \omega)$ we can take uncorrelated and homoscedastic observations of Y. Let $K'\theta$ be the parameter system of interest, described by a $k \times s$ matrix K with full column rank s. In this setup, the choice of $(X; \omega)$ is based on its information matrix for $K'\theta$,

$$C_K(M(X; \omega)) = \min_{\{L \in \mathbb{R}^{s \times k} : LK = I_s\}} LM(X; \omega)L'.$$

Here the minimum refers to the Loewner partial ordering on the set of nonnegative definite $s \times s$ matrices, and the moment matrix $M(X; \omega)$ is defined by

$$M(X; \omega) = \sum_{i=1}^{\ell} \omega_i x_i x_i' = X \mathrm{diag}(\omega) X'.$$

The notation $\mathrm{diag}(a)$ is used for a diagonal matrix that has the entries of the vector a as its diagonal elements.

For some regression setups the number ℓ of support points of at least one optimal design is known. In general, estimability of the interesting parameter system requires at least s support points. The Carathéodory theorem gives an upper bound for ℓ, whence we get the following box constraints

$$s \leq \ell \leq s(s+1)/2 + s(k-s). \tag{1}$$

The number of support points can be eliminated from the variable list by using the upper bound in (1) and extending the domain of ω to the ℓ dimensional probability simplex $S^\ell = \{\alpha \in \mathbb{R}^\ell : \alpha_i \geq 0 \text{ for } i = 1, \ldots, \ell, \sum_{i=1}^{\ell} \alpha_i = 1\}$. By $\mathcal{X}^\ell$ we denote the ℓ fold cartesian product of the regression space $\mathcal{X}$.

The main problem thus consists in solving

$$\text{maximize} \quad \psi(X; \omega) \text{ subject to } X \in \mathcal{X}^\ell \text{ and } \omega \in S^\ell. \tag{2}$$

Here ψ is used as abbreviation of the composition $\phi \circ C_K \circ M$ for a given information function ϕ. The notion of information functions was introduced by Pukelsheim (1980) (Definition 3) and is thoroughly discussed in Chapter 5 in Pukelsheim (1993).

A popular class of information functions are the concave matrix means ϕ_p for $-\infty \leq p \leq 1$ (see Chapter 6 in Pukelsheim (1993)). They form the concave analogues to Kiefer's φ_p-criteria, introduced in Kiefer (1974).

For a positive definite $s \times s$ matrix A the pth matrix means are defined in terms of the ordered eigenvalues $\lambda_1, \ldots, \lambda_s$ of A by

$$\phi_p(A) = (s^{-1} \sum \lambda_i^p)^{1/p} \qquad \text{for } -\infty < p \leq 1, \quad p \neq 0.$$

The widely used D-criterion is obtained as a limit for $p = 0$, i.e. $\phi_0(A) = (detA)^{1/s}$. Further as limit for $p = -\infty$ one gets the well-known E-criterion which calls for maximization of the smallest eigenvalue λ_s of A. This criterion is differentiable if and only if the smallest eigenvalue is simple. Differentiability also breaks down for the other extremum $p = 1$ (T-optimality), since the optimum moment matrix tends to be singular in this case. Singular moment matrices may even arise for $0.5 < p < 1$ due to rounding errors and they also come with parameter subsystems.

In addition, differentiability may fail if the weight vector comes to lie on the boundary of the probability simplex (see Fellman (1980)). Therefore, a general algorithm for solving problem (2) should rely on nondifferentiable optimization techniques. Moreover, we must be able to treat nonconvex problems due to the following fact. The composition $\phi \circ C_K$ is an information function on the set of nonnegative definite $k \times k$ matrices and hence a concave function of the moment matrix M. The moment matrix mapping is a linear function of the weight vector ω, but it is convex considered as a function of the support points x_i. So we can neither ensure convexity nor concavity for our objective function ψ.

A special situation arises when the regression range consists of finitely many regression vectors. The design problem then can be formulated as a problem in the weights only, yielding a concave maximization problem. In this case we write ψ_X for $\psi(X, \cdot)$ and with this notation the problem reads as:

$$\text{maximize} \quad \psi_X(\omega) \text{ subject to } \omega \in S^\ell. \tag{3}$$

To solve this problem we adapt a bundle trust version for convex minimization problems.

3. The BT-Methodology

In Wilhelm (1995) it is shown that the objective function ψ in problem (2) is directionally Lipschitz continuous at all designs $(X;\omega)$ with range$K \subseteq$ range$X$. ψ is even locally Lipschitz continuous on the set of designs with feasible moment matrix, where 'feasible' means that the range of the moment matrix contains the range of the parameter matrix K. It is well known that a locally Lipschitz continuous function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is differentiable almost everywhere. The generalized gradient (see Clarke (1983)) is defined by

$$\begin{aligned}\partial f[x] = conv\{g : &\text{ there exists a sequence } (x_i)_{i\in\mathbb{N}} \text{ such that } \lim_{i\to\infty} x_i = x,\\ &f \text{ is differentiable at } x_i, i \in \mathbb{N}, \text{ and } \lim_{i\to\infty} \nabla f(x_i) = g\},\end{aligned}$$

where *conv* denotes the convex hull and ∇ the gradient mapping.

For a convex function f the generalized gradient coincides with the subdifferential

$$\partial f[x] = \{g \in \mathbb{R}^m : f(y) \geq f(x) + \langle g, y - x\rangle \text{ for all } y \in \mathbb{R}^m\}.$$

Therefore, the elements of $\partial f[x]$ are called subgradients of f at x, even for nonconvex f.

Computing the whole subdifferential is usually very time consuming and for many problems it is not possible to get more than one subgradient of f at x. Bundle methods require only the knowledge of one arbitrary subgradient at each iteration point. Subgradients evaluated during the process are stored in the so-called bundle and a convex combination of the stored subgradients is used to compute a search direction. The next iterate is derived through a line search along this direction. The next iteration point will be accepted if there is a sufficient increase of the function value. The calculated subgradient at this point will be added to the bundle if it enriches the subgradient information no matter whether the corresponding point is accepted or not. (For more details see Hiriart-Urruty and Lemaréchal (1993)). The bundle trust methods avoid the usual line search by using a cutting plane model with an additional trust region term. The next candidate is then obtained as solution of a quadratic program.

Transforming this methodology to the design problem (2) results in the following algorithm:

Step 1. Choose a starting design $(X^0;\omega^0)$, an accuracy parameter $\epsilon > 0$, a small positive constant c_0, two positive numbers $0 < s_1 < s_2 < 1$ and an upper bound $J_{\max} \geq 3$ for the number of subgradients to be stored in the bundle. Let $X_L(X_U)$ be a lower (upper) bound for the competing support matrices $X \in \mathcal{X}^\ell$.

Step 2. Put $m = 0$, $J_0 = \{0\}$, $Z^0 = X^0$, $\eta^0 = \omega^0$, and compute $\psi(X^0;\omega^0)$ and $g^0 \in \partial\psi[(X^0;\omega^0)]$.

Step 3. *Inner Iteration*
For all $j \in J_m$ compute the linearization errors

$$\alpha_j^m = \psi(Z^j;\eta^j) - \psi(X^m;\omega^m) + \langle g^j, (X^m;\omega^m) - (Z^j;\eta^j)\rangle$$

and put

$$\beta_j^m = \max\{\alpha_j^m, c_o\|(X^m;\omega^m) - (Z^j;\eta^j)\|^2\}.$$

For fixed $t^m > 0$ solve the following quadratic program in (v, D, δ) with $v \in \mathbb{R}$, $D \in \mathbb{R}^{k\times\ell}$, $\delta \in \mathbb{R}^\ell$:

$$\begin{aligned} &\text{maximize} \quad v - \frac{1}{2t^m}\|(D;\delta)\|^2 && (4)\\ &\text{subject to } v \le \beta_j^m + \langle g^j, (D;\delta)\rangle, \qquad j \in J_m && (5)\\ &\qquad -\omega_i^m \le \delta_i \le 1 - \omega_i^m, \qquad i = 1,\dots,\ell &&\\ &\qquad \sum_{i=1}^{\ell} \delta_i = 0 &&\\ &\qquad X_L - X^m \preceq D \preceq X_U - X^m. && \end{aligned}$$

(Here $\preceq$ denotes componentwise ordering of matrices.)

Let (v^m, D^m, δ^m) denote the attained solution and λ_j^m, $j \in J_m$, the corresponding Lagrange multipliers. Put $Z^{m+1} = X^m + D^m, \eta^{m+1} = \omega^m + \delta^m$ and compute $\psi(Z^{m+1};\eta^{m+1})$, one subgradient g^{m+1} of ψ at $(Z^{m+1};\eta^{m+1})$ and β_{m+1}^m.
If $\psi(Z^{m+1};\eta^{m+1}) - \psi(X^m;\omega^m) \ge s_1 v^m$ we do a serious step
else if $\beta_{m+1}^m + \langle g^{m+1}, (D^m;\delta^m)\rangle < s_2 v^m$ and $\beta_{m+1}^m \le \sum_{j\in J_m} \lambda_j^m \beta_j^m$ then we do a null step, else we modify t^m and try again solving problem (4).

Step 4. *Serious Step / Null Step*
If the serious criterion is fulfilled, we put $X^{m+1} = Z^{m+1}$, $\omega^{m+1} = \eta^{m+1}$, $J_{m+1} = J_m \cup \{m+1\}$, $m = m+1$, and go to Step 5, else if the null step criterion is fulfilled, we put $X^{m+1} = X^m$, $\omega^{m+1} = \omega^m$, $J_{m+1} = J_m \cup \{m+1\}$, $m = m+1$, and go to Step 5.

Step 5. *Stopping criterion*
If $\| \sum_{j\in J_m} \lambda_j^m g^j \| \le \epsilon$ and $\sum_{j\in J_m} \lambda_j^m \beta_j^m \le \epsilon$ then STOP else go to Step 6.

Step 6. *Reset*
If $\mid J_m \mid = J_{\max}$ then gather all active subgradients in the set $J \subset J_m$, that is put $J = \{j : j \in J_m, \lambda_j^m > 0\}$.
If $\mid J \mid \le J_{\max} - 1$ then put $J_{m+1} = J$ and go to Step 3, else introduce some additional index $\tilde{m}$, put $g^{\tilde{m}} = \sum_{j\in J_m} \lambda_j^m g^j$, $\beta_{\tilde{m}}^m = \sum_{j\in J_m} \lambda_j^m \alpha_j^m$ and $J_{m+1} = \{m\} \cup \{\tilde{m}\}$ and go to Step 3.
In the following paragraphs more detailed information with regard to the above algorithm is given.

Choice of initial parameters

In the concave case all linearization errors α_j^m are nonnegative and, therefore, in this case the parameter c_0 is chosen to be zero. In general, our objective function ψ is not concave, whence the subgradient inequality is no longer valid and the α_j^m's may become negative. It is common in nonsmooth optimization to replace the linearization errors under this circumstances by the "weights" β_j^m. The only aim for the parameter c_0 lies in preventing negative linearization errors, whence, usually a small value for c_0 is chosen. In our implementation we had good experiences by setting $c_0 = 0.01$, $s_1 = 0.1$, $s_2 = 0.2$ and $J_{\max} = 20$. (For the meaning of s_1 and s_2 see below.)

Upper and lower bounds for the competing support matrices are usually given by the special problem under investigation. Up to now we only use the constraints induced by the bounds of the design region.

Inner Iteration

The inner iteration aims at finding a suitable steplength t^m. The initial value and adaptation of the parameter t^m can be controlled by providing an estimate for the optimum value of ψ. As a default we take $|\psi(X^0;\omega^0)| * 10^3$ as optimum value of ψ, as recommended by Schramm (1991), p. 13.

The maximization problem (4) of the inner iteration is a reformulation of the following trust region problem:

$$\begin{aligned} &\text{maximize} \quad \widehat{\psi}_m((X^m;\omega^m) + (D;\delta)) \\ &\text{subject to } \frac{1}{2}\|(D;\delta)\|^2 \le \rho^m \\ &\qquad\qquad -\omega_i^m \le \delta_i \le 1 - \omega_i^m, \qquad i = 1,\dots,\ell \\ &\qquad\qquad \sum_{i=1}^{\ell} \delta_i = 0 \\ &\qquad\qquad X_L - X^m \preceq D \preceq X_U - X^m. \end{aligned} \tag{6}$$

The cutting plane model $\widehat{\psi}_m$ of ψ at $(X^m;\omega^m)$ is thereby obtained using the stored subgradients via

$$\widehat{\psi}_m((X^m;\omega^m) + (D;\delta)) = \psi(X^m;\omega^m) + \max_{j\in J_m}\{\langle g^j,(D;\delta)\rangle - \beta_j^m\}.$$

The quadratic constraint in problem (6) is resolved by a penalty approach and the trust region parameter ρ^m is substituted by the parameter t^m (see Schramm (1991) for justification).

Since v^m represents the increase of the function value predicted by the cutting plane model, the serious step criterion decides whether we can trust our model or not. The parameter s_1 indicates what fraction of the model increase is at least required in a serious step.

If a sufficient increase in the function value cannot be achieved, we then check in the null step criterion whether addition of the actual subgradient leads to a model change and improves information. The first inequality of the null step criterion says that the actual solution v^m of the quadratic program (4) violates constraint (5) for the new subgradient g^{m+1} and, hence, the model changes. If in the concave case the serious criterion is not fulfilled, then the first inequality in the null step criterion holds automatically even with parameter s_1. In the nonconcave case we add the mentioned inequality to the null step criterion with a weaker condition controlled by the parameter s_2.

If neither the serious nor the null step criterion is fulfilled the parameter t^m is adapted by bisection. If it is impossible to diminish t^m, a line search is performed, that sometimes can cause numerical troubles.

Stopping criterion and convergence

A necessary but not sufficient condition for $(X^m; \omega^m)$ to be optimal is the nullvector being an element of the subdifferential of ψ at $(X^m; \omega^m)$. The stopping rule merely says that 0 "lies up to ϵ" in the convex hull of certain subgradients g^j of ψ at designs $(Z^j; \eta^j)$ which are "not far away from $(X^m; \omega^m)$" (see Schramm (1991)).

It can be shown that there exists a cluster point $(\bar{X}; \bar{\omega})$ of the sequence $\{(X^m; \omega^m)\}_{m \in \mathbb{N}}$ generated by the BT-algorithm if ψ is weakly semismooth, bounded below and if the sequence of iterates is bounded (see Schramm (1991)). A function $f : \mathbb{R}^m \to \mathbb{R}$ is called weakly semismooth if its directional derivative $f'(x; d)$ exists for all x and d and can be represented as $f'(x; d) = \lim_{t \downarrow 0} g(x + td)'d$ where $g(x + td)$ is a generalized gradient of f at $(x + td)$.

4. Computing Function Values and Subgradients

In our implementation OPTDES we allow as information functions all pth matrix means, $-\infty \leq p \leq 1$. These information functions are completely determined by the eigenvalues of the information matrix. A spectral decomposition of the information matrix may cause numerical problems, whence a singular value decomposition of a matrix root of the information matrix is preferable. We calculate the function value and one subgradient of ψ at $(X; \omega)$ in the following way.

1. Solve the linear equation $(K'K)L = K'$ yielding a left inverse L of the coefficient matrix K. Let $R = I_k - KL$. This step only has to be worked out once.

2. Let $M = M(X;\omega) = X\mathrm{diag}(\omega)X'$. The minimizing left inverse L_M of K used in the defining relation of the information matrix $C = C_K(M)$ is obtained as solution of $\begin{pmatrix} K' \\ RM \end{pmatrix} L_M' = \begin{pmatrix} I_s \\ 0 \end{pmatrix}$.
3. Computing $\tilde{C} = L_M X\mathrm{diag}(\omega_1^{1/2},\ldots,\omega_\ell^{1/2})$ and running a singular value decomposition $\tilde{C} = U\mathrm{diag}(\sigma_1,\ldots,\sigma_s)V'$ yield the roots $\sigma_1,\ldots,\sigma_s$ of the eigenvalues of C.
4. From the defining relation of pth matrix means we derive the following formulae for the function value $\psi(X;\omega)$

$$\psi(X;\omega) = \begin{cases} (\frac{1}{s}\sum_{j=1}^{s}(\sigma_j)^{2p})^{\frac{1}{p}} & \text{for } p \in (-\infty,1],\ p \neq 0, \\ (\prod_{j=1}^{s}(\sigma_j)^2)^{\frac{1}{s}} & \text{for } p = 0, \\ (\sigma_s)^2 & \text{for } p = -\infty. \end{cases}$$

5. According to Lemma 6.16 in Pukelsheim (1993) and Theorem 2 in Wilhelm (1995), some simple vector and matrix operations easily gives us one subgradient of $\phi \circ C_K$ at M at hand:

$$B = \begin{cases} (\sum_{j=1}^{s}(\sigma_j)^{2p})^{-1} L_M{}'U\mathrm{diag}((\sigma)^{2p-2})U'L_M & \text{for } p \in (-\infty,1], \\ \frac{1}{\#\sigma_{\min}(\tilde{C})} \sum_{i=0}^{\#\sigma_{\min}(\tilde{C})-1} L_M{}'u_{s-i}u_{s-i}{}'L_M & \text{for } p = -\infty. \end{cases}$$

Here $\#\sigma_{\min}(\tilde{C})$ denotes multiplicity of the smallest eigenvalue of the information matrix C and u_j denotes a normalized eigenvector corresponding to the jth eigenvalue of C stored in the jth column of U.

A subgradient $(\tilde{B};\beta)$ of ψ at $(X;\omega)$ is obtained by calculating $\tilde{B} = 2BX\mathrm{diag}(\omega)$ and $\beta = (x_1{}'Bx_1,\ldots,x_\ell{}'Bx_\ell)'$.

The implementation is done in FORTRAN and for most of the matrix routines the NAG Fortran Library is called.

5. Finitely Many Regression Vectors

In the case of finitely many regression vectors the objective function is concave. We then put $c_0 = 0$ and, therefore, β_j^m equals α_j^m. If in the mth iteration the stopping criterion is fulfilled, we have at hand an efficiency bound for the optimal function value. Since all convex combinations $g = \sum_{j\in J_m}\mu_j g^j$ with $\mu_j \geq 0$, $\sum \mu_j = 1$ lie in the ϵ-subdifferential of $\psi_X(\omega^k)$ for $\epsilon = \sum_{j\in J_m}\mu_j\alpha_j^m$, the subgradient inequality implies for all $\gamma \in S^\ell$

$$\begin{aligned} \psi_X(\gamma) &\leq \psi_X(\omega^m) + \langle g, \gamma - \omega^m\rangle + \epsilon \\ &\leq \psi_X(\omega^m) + \|g\|\|\gamma - \omega^m\| + \epsilon \leq \psi_X(\omega^m) + \epsilon(1+\sqrt{2}). \end{aligned}$$

Thus, putting $\tilde{\epsilon} = \epsilon\psi_X(\omega^0)/(1+\sqrt{2})$ we get the following efficiency bound

$$\frac{\psi_X(\omega^m)}{\max_{\gamma\in S^\ell}\psi_X(\gamma)} \geq 1-\tilde{\epsilon}.$$

On the other hand we get an additional stopping criterion by using duality arguments.

Theorem 1. *A weight vector $\omega \in S^\ell$ with $\psi_X(\omega) > 0$ is ψ_X-optimal in S^ℓ for estimating $K'\theta$ if and only if there exists a subgradient $g \in \partial\psi_X[\omega]$ such that*

$$\begin{aligned} g_i &= \psi_X(\omega), && \textit{for all } \ i \textit{ with } \omega_i > 0 \\ g_i &\geq \psi_X(\omega), && \textit{for all } \ i \textit{ with } \omega_i = 0. \end{aligned}$$

Proof. From the Duality Theorem 7.12 in Pukelsheim (1993) we know that a moment matrix M^* maximizes $\phi \circ C_K$ over a set $\mathcal{M}$ if and only if there exists a matrix $N \in \mathcal{N} = \{N \in \mathrm{NND}(k) : \mathrm{tr}\, AN \leq 1 \text{ for all } A \in \mathcal{M}\}$ such that, upon setting $C = C_K(M^*)$, we have

$$\mathrm{tr}\, M^*N = 1 \tag{7}$$

$$M^*N = KCK'N \tag{8}$$

$$\phi(C)\phi^\infty(K'NK) = \mathrm{tr}\, CK'NK. \tag{9}$$

Here ϕ^∞ denotes the polar function to ϕ (see Chapter 5 in Pukelsheim (1993)). Denote $M = M(X;\omega)$ and let B be an arbitrary subgradient of $\phi \circ C_K$ at M. Obviously, $N = (\max_{i=1,\ldots,\ell} x_i'Bx_i)^{-1}B$ is a member of $\mathcal{N}$ and it follows from the subdifferential decomposition 7.8 and 7.9 in Pukelsheim (1993) that M and N fulfill conditions 8 and 9.

Condition 7 writes as $x_i'Bx_i = \max_{i=1,\ldots,\ell} x_i'Bx_i$ for all i with positive weight ω_i. Since N is a member of $\mathcal{N}$, Theorem 7.11 in Pukelsheim (1993) says that $\phi(C_K(M)) \leq (\phi^\infty(K'NK))^{-1}$. For the above choice of N the right hand side equals $\max_{i=1,\ldots,\ell} x_i'\, Bx_i$. The assertion now follows from the subgradient characterization in Corollary 2 in Wilhelm (1995).

According to Theorem 1 we can stop the optimization process if in the actual iteration point the quotient $(\max_{i=1,\ldots,\ell} g_i^m)^{-1}\psi_X(\omega^m)$ exceeds $1-\tilde{\epsilon}$.

6. Numerical Results

Convergence results for the BT-methodology are only available for the concave case, in the design context this means for problems with fixed support points or in the case of finitely many regression vectors. To show that the method proposed in this article ends in reliable results we give some examples for the numerical behaviour of our implementation. All computations were carried out on a HP 9000/400s Apollo Workstation.

6.1. Accuracy

First we give a comparison between the optimal function values obtained by OPTDES and those obtained by POLYPLAN. The latter program was written by Preitschopf (1989) and is particularly designed for differentiable experimental design problems in polynomial regression models over the interval $[-1,1]$. POLYPLAN requires that the number of support points of the optimal design is known in advance. The results obtained by the optimization routine are checked in POLYPLAN via the General Equivalence Theorem and bear, therefore, high accuracy. In Table 1 the optimal function values are given for the D- and A-criterion in polynomial regression models of degree d,

$$y(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \cdots + \theta_d t^d + \epsilon,$$

for $d = 2,\ldots,12$. As starting designs we used uniform distribution on $d+1$ equidistant points. Symmetry is used in POLYPLAN to reduce dimensionality, so POLYPLAN works with $d-1$ variables, whereas OPTDES works with $2d+2$ variables.

Table 1. A- and D-optimal function values for polynomial regression models obtained by OPTDES and POLYPLAN

	A-optimality		D-optimality	
d	OPTDES	POLYPLAN	OPTDES	POLYPLAN
2	.37500000	.37500000	.52913368	.52913368
3	.10660907	.10660907	.26749612	.26749612
4	.26497896E-01	.26497897E-01	.13385589	.13385589
5	.61067953E-02	.61067953E-02	.66785544E-01	.66785544E-01
6	.13399177E-02	.13399177E-02	.33293682E-01	.33293682E-01
7	.28390598E-03	.28390598E-03	.16595215E-01	.16595215E-01
8	.58600445E-04	.58600445E-04	.82728583E-02	.82728583E-02
9	.11851683E-04	.11851683E-04	.41249350E-02	.41249350E-02
10	.23581719E-05	.23581719E-05	.20571972E-02	.20571972E-02
11	.46298770E-06	.46298770E-06	.10261932E-02	.10261932E-02
12	.89892637E-07	.89892637E-07	.51199949E-03	.51199949E-03

The results of OPTDES differ from those of POLYPLAN only in the last significant digit of the optimal value for the A-criterion in polynomial regression of degree $d = 4$.

As a second example we consider linear regression over the cube $[-1.00, 1.00]^3$,

$$y(t_1,t_2,t_3) = \theta_0 + \theta_1 t_1 + \theta_2 t_2 + \theta_3 t_3 + \epsilon.$$

Let us try to find a D-optimal design by starting with the uniform distribution on the points of a full 3^3 factorial design with levels -1, 0 and 1. The D-optimal design obtained by OPTDES puts mass 1/4 on each of the following four corners, $x_1 = (-1.0, -1.0, 1.0)'$, $x_2 = (-1.0, 1.0, -1.0)'$, $x_3 = (1.0, -1.0, -1.0)'$ and $x_4 = (1.0, 1.0, 1.0)'$ reducing the number of support points from 27 to 4. In contrast, the D-optimal design obtained by ACED of Welch (1984) puts weight 0.065 to these four corners and weight 0.185 to the other four, see Figure 1. The value of the D-criterion for both designs is the same and equals 1.

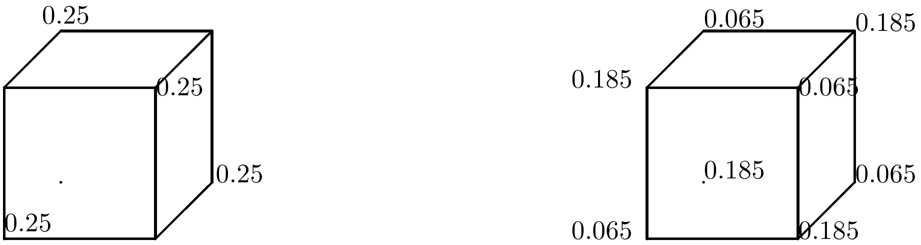

Figure 1. D-optimal designs for θ in linear regression over three-dimensional cube $[-1, 1]^3$. On the left we show the design found by OPTDES, on the right the one found by ACED.

6.2. Initial values

The quality of an algorithm can be judged by its robust behaviour under different starting points. Since the function value of the actual iteration point appears in the subgradient representation as a multiplicative factor, it is necessary that the starting design has positive information. A quite natural choice as starting design is uniform distribution on equidistant points, but the examples given below indicate that poor starting designs often perform better. Usually, in the neighbourhood of the optimum the objective function is very flat and the subgradients only offer a small amount of new information. Starting with a poor design opens the possibility of getting a good global model of our objective function and, hence, this can lead to fast convergence.

In Table 2 below, we show the results for different starting designs for E-optimality in polynomial regression of degree $d = 3$ over the interval $[-1, 1]$. The optimal value 0.040000 is attained at the design $\left\{\begin{array}{cccc} -1.0 & -0.5 & 0.5 & 1.0 \\ 0.126667 & 0.373333 & 0.373333 & 0.126667 \end{array}\right\}$. The last column in Table 2 gives the number of iterations that has been performed by OPTDES to obtain the optimum with an error bound of 10^{-6}. We failed to converge to the optimum with the last starting design. In this case we got stuck in a local optimum. The procedure ended in the optimal design for the reduced experimental domain $[0.0, 1.0]$. Numerical evidence suggests that we get stuck in a local optimum if and only if zero is included as support point in the starting design and the other support points are all positive or all negative.

Table 2. Results for E-optimality for different starting designs

starting design	starting value	# iterations
$\left\{\begin{array}{cccc} -1.0 & -1/3 & 1/3 & 1.0 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array}\right\}$	0.021205	73
$\left\{\begin{array}{ccccccccccc} -1.0 & -0.8 & -0.6 & -0.4 & -0.2 & 0.0 & 0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\ 1/11 & 1/11 & 1/11 & 1/11 & 1/11 & 1/11 & 1/11 & 1/11 & 1/11 & 1/11 & 1/11 \end{array}\right\}$	0.023364	108
$\left\{\begin{array}{cccc} -0.000001 & 0.3 & 0.6 & 1.0 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array}\right\}$	0.000146	138
$\left\{\begin{array}{cccc} -1.0 & -0.6 & -0.3 & 0.000001 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array}\right\}$	0.000146	155
$\left\{\begin{array}{cccc} 0.000001 & 0.3 & 0.6 & 1.0 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{array}\right\}$	0.000146	151
$\left\{\begin{array}{cccc} -0.8 & -0.2 & 0.1 & 0.6 \\ 0.0001 & 0.9997 & 0.0001 & 0.0001 \end{array}\right\}$	0.000001	66
$\left\{\begin{array}{cccc} 0.9 & 0.95 & 0.99 & 1.0 \\ 0.0001 & 0.0001 & 0.0001 & 0.9997 \end{array}\right\}$	1.140E-11	108
$\left\{\begin{array}{cccc} 0.0 & 0.5 & 0.75 & 1.0 \\ 0.000001 & 0.373333 & 0.5 & 0.1266666 \end{array}\right\}$	1.589E-08	∞

6.3. Further examples

New insight results from our method especially for E-optimality in polynomial regression models of degree d over a real compact interval $[a, b]$. Heiligers (1994) showed that simplicity of the smallest eigenvalue $\lambda_{\min}(C_K(M))$ depends on the interval width. For symmetric intervals $[-b, b]$ where b is small enough (e.g. if $b \leq 1$) the smallest eigenvalue is simple and the E-optimal design for all $d+1$ parameters is supported by the extremum points $x_{i,d}^{CH} = \cos[\pi(d-i)/d]$, for $i = 0, 1, \ldots, d$, of the dth degree Chebychev polynomial (see Heiligers (1994) and Pukelsheim and Studden (1993)). These points coincide with the dth degree quantiles of the arcsin distribution and designs having these points for its support are called arcsin support designs. However, for large b arcsin support designs may become inadequate for solving the E-optimal design problem. For cubic regression the efficiency of the E-optimal arcsin support design for θ compared to the E-optimal one among all designs decreases from 1.0 for $b = 1.6$ to 0.543 for $b = 3.0$. Table 3 shows the E-optimal designs for θ found by OPTDES and their information gain compared to the arcsin support designs.

Table 3. E-optimal designs for θ in cubic regression over $[-b, b]$ and their information gain compared to the arcsin support designs.

	support points				weights				percent
b	x_1^*	x_2^*	x_3^*	x_4^*	ω_1^*	ω_2^*	ω_3^*	ω_4^*	info gain
1.6	−1.60	−0.80	0.80	1.6	0.127	0.373	0.373	0.127	0.0
1.7	−1.700	−0.826	0.826	1.700	0.0779	0.4221	0.4221	0.0779	0.59
1.8	−1.800	−0.844	0.844	1.800	0.0764	0.4236	0.4236	0.0764	4.73
1.9	−1.900	−0.859	0.859	1.900	0.0742	0.4258	0.4258	0.0742	13.30
2.0	−2.000	−0.873	0.873	2.000	0.0715	0.4285	0.4285	0.0715	21.55
3.0	−3.000	−0.943	0.943	3.000	0.0439	0.4561	0.4561	0.0439	84.29

Acknowledgement

We thank H. Schramm for providing us with her BT-codes and the referees for valuable comments, especially, on the organizational concept of the paper.

References

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis.* Wiley, New York.

Cook, R. D. and Nachtsheim, C. J. (1980). A comparison of algorithms for constructing exact D-optimal designs. *Technometrics* **22**, 315-324.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Transl. and ed. by W.J. Studden and E. M. Klimko. Academic Press, New York.

Fellman, J. (1980). On the behaviour of the optimality criterion in the neighborhood of the optimal point. Swedish School of Economics and Business Administration, Helsinki. Working Paper 49, 15 pages.

Gaffke, N. (1985). Directional derivatives of optimality criteria at singular matrices in convex design theory. *Statistics* **16***,* 373-388.

Gaffke, N. and Mathar, R. (1992). On a class of algorithms from experimental design theory. *Optimization* **24***,* 91-126.

Hardin, R. H. and Sloane, N. J. A. (1993). A new approach to the construction of optimal designs. *J. Statist. Plann. Inference* **37***,* 339-369.

Heiligers, B. (1994). E-optimal designs in weighted polynomial regression. *Ann. Statist.* **22***,* 917-929.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (1993). *Convex Analysis and Minimization Algorithms I, II.* Springer, New York.

Hoang, T. and Seeger, A. (1991). On conjugate functions, subgradients, and directional derivatives of a class of optimality criteria in experimental design. *Statistics* **22***,* 349-368.

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

Nguyen, N. K. and Miller, A. J. (1992). A review of some exchange algorithms for constructing discrete D-optimal designs. *Comput. Statist. Data Anal.* **14***,* 489-498.

Preitschopf, F. (1989). *Bestimmung Optimaler Versuchspläne in der Polynomialen Regression.* Dissertation, Universität Augsburg, 763-770.

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

Pukelsheim, F. (1993). *Optimal Design of Experiments.* Wiley, New York.

Pukelsheim, F. and Studden, W. J. (1993). E-optimal designs for polynomial regression. *Ann. Statist.* **21***,* 402-415.

Rockafellar, R. T. (1980). Generalized directional derivatives and subgradients of nonconvex funtions. *Canad. J. Math.* **23***,* 257-280.

Schramm, H. (1991). Bundle trust methods: Fortran codes for nondifferentiable optimization: User's guide. Report No. 269, Schwerpunktprogramm der Deutschen Forschungsgemeinschaft *Anwendungsbezogene Optimierung und Steuerung*, Universität Bayreuth, 17 pages.

Schramm, H. and Zowe, J. (1988). A combination of the bundle approach and the trust region concept. In *Mathematical Research* 45*, Advances in Mathematical Optimization* (Edited by J. Guddat, B. Blank, H. Hollatz, P. Kall, D. Klatte, B. Kummer, K. Lommatzsch, K. Tammer, M. Vlach, K. Zimmermann), 196-209. Akademie Verlag, Berlin.

Welch, W. J. (1984). ACED: Algorithms for the construction of experimental designs: User's guide.

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

Wilhelm, A. (1995). Subdifferentiability and Lipschitz continuity in experimental design problems. *Metrika* **42**, 365-377.

Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse, Universität Augsburg, D-86135 Augsburg, Germany.

E-mail: adalbert.wilhelm@math.uni-augsburg.de

(Received November 1994; accepted August 1996)