Improvements on removing non-optimal support points in D-optimum design algorithms

Radoslav Harman [a,1] and Luc Pronzato [b,*,2]

[a] *Department of Applied Mathematics and Statistics*
Faculty of Mathematics, Physics & Informatics
Comenius University, Mlynská Dolina, 84248 Bratislava, Slovakia

[b] *Laboratoire I3S, CNRS/UNSA, Bât. Euclide, Les Algorithmes,*
2000 route des Lucioles, BP 121, 06903 Sophia-Antipolis Cedex, France

Abstract

We improve the inequality used in (Pronzato, 2003) to remove points from the design space during the search for a D-optimum design. Let ξ be any design on a compact space $\mathcal{X} \subset \mathbb{R}^m$ with a nonsingular information matrix, and let $m + \epsilon$ be the maximum of the variance function $d(\xi, \mathbf{x})$ over all $\mathbf{x} \in \mathcal{X}$. We prove that any support point $\mathbf{x}_*$ of a D-optimum design on $\mathcal{X}$ must satisfy the inequality $d(\xi, \mathbf{x}_*) \geq m(1 + \epsilon/2 - \sqrt{\epsilon(4 + \epsilon - 4/m)}/2)$. We show that this new lower bound on $d(\xi, \mathbf{x}_*)$ is, in a sense, the best possible, and how it can be used to accelerate algorithms for D-optimum design.

Key words: D-optimum design, design algorithm, support points
PACS: 62K05, 90C46

* author for correspondence
Email addresses: `harman@fmph.uniba.sk` (Radoslav Harman), `pronzato@i3s.unice.fr` (Luc Pronzato).

[1] The work of the first author was supported by the VEGA grant No. 1/0264/03 of the Slovak Agency.

[2] The work of the second author was partially supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors view.

1 Introduction

Let $\mathcal{X} \subseteq \mathbb{R}^m$ be a compact design space and let Ξ be the set of all designs (i.e., finitely supported probability measures) on $\mathcal{X}$. For any $\xi \in \Xi$, let

$$\mathbf{M}(\xi) = \int_{\mathcal{X}} \mathbf{x}\mathbf{x}^\top \, \xi(\mathbf{dx})$$

denote the information matrix. Suppose that there exists a design with nonsingular information matrix and let Ξ^+ be the set of such designs. Let ξ^* denote a D-optimum design, that is, a measure in Ξ that maximizes $\det \mathbf{M}(\xi)$, see, e.g., (Fedorov, 1972). Note that a D-optimum design always exists and that the D-optimum information matrix $\mathbf{M}_* = \mathbf{M}(\xi^*)$ is unique. For any $\xi \in \Xi^+$ denote $d(\xi, \cdot) : \mathcal{X} \to [0, \infty)$ the variance function defined by

$$d(\xi, \mathbf{x}) = \mathbf{x}^\top \mathbf{M}^{-1}(\xi)\mathbf{x}\,.$$

The celebrated Kiefer-Wolfowitz Equivalence Theorem (1960) writes as follows.

Theorem 1 *The following three statements are equivalent:*

(i) ξ^* *is D-optimum;*
(ii) $\max_{\mathbf{x}\in\mathcal{X}} d(\xi^*, \mathbf{x}) = m$*;*
(iii) ξ^* *minimizes* $\max_{\mathbf{x}\in\mathcal{X}} d(\xi, \mathbf{x})$*,* $\xi \in \Xi^+$*.*

Notice that

$$\int_{\mathcal{X}} d(\xi^*, \mathbf{x}) \, \xi^*(d\mathbf{x}) = \int_{\mathcal{X}} \mathbf{x}^\top \mathbf{M}_*^{-1} \mathbf{x} \, \xi^*(d\mathbf{x}) = \mathrm{trace}(\mathbf{M}_* \mathbf{M}_*^{-1}) = m\,.$$

Hence, (ii) of Theorem 1 implies that for any support point $\mathbf{x}_*$ of the design ξ^* (i.e., for a point satisfying $\xi^*(\mathbf{x}_*) > 0$), we have

$$d(\xi^*, \mathbf{x}_*) = m\,. \tag{1}$$

In the next section we show that the equality (1) can be used to prove that

$$\forall \xi \in \Xi^+\,,\ d(\xi, \mathbf{x}_*) \geq m\lambda_1^*(\xi)$$

where λ_1^* depends on ξ only via the maximum of $d(\xi, \cdot)$ over the design space $\mathcal{X}$. Hence, we can test candidate support points by using any finite number of design measures $\xi \in \Xi^+$, e.g., those that are generated by a design algorithm on its way towards the optimum: any point that does not pass the test defined by ξ^k of iteration k need not be considered for further investigations and can thus be removed from the design space.

2 A necessary condition for candidate support points

For ξ a design in Ξ^+ denote $\mathbf{M} = \mathbf{M}(\xi)$,

$$\mathbf{H} = \mathbf{M}^{-1/2}\mathbf{M}_*\mathbf{M}^{-1/2}$$

and $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$ the eigenvalues of $\mathbf{H}$. Notice that $\lambda_1 > 0$ and that the eigenvalues depend on the design ξ as well as on the D-optimum information matrix $\mathbf{M}_*$. Let $\mathbf{x}_*$ be a support point of a D-optimum design and let $\mathbf{y}_* = \mathbf{H}^{-1/2}\mathbf{M}^{-1/2}\mathbf{x}_*$. The equality (1) can be written in the form $\mathbf{y}_*^\top\mathbf{y}_* = m$ which implies:

$$d(\xi, \mathbf{x}_*) = \mathbf{x}_*^\top\mathbf{M}^{-1}\mathbf{x}_* = \mathbf{y}_*^\top\mathbf{H}\mathbf{y}_* \geq \lambda_1\mathbf{y}_*^\top\mathbf{y}_* = m\lambda_1\,. \tag{2}$$

To be able to use the inequality (2), we need to derive a lower bound λ_1^* on λ_1 that does not depend on the unknown matrix $\mathbf{M}_*$.

Theorem 1-(ii) implies

$$\begin{aligned}\sum_{i=1}^m \lambda_i^{-1} &= \mathrm{trace}(\mathbf{H}^{-1})\\ &= \mathrm{trace}(\mathbf{M}_*^{-1}\mathbf{M}) = \int_{\mathcal{X}} \mathbf{x}^\top\mathbf{M}_*^{-1}\mathbf{x}\,\xi(d\mathbf{x}) = \int_{\mathcal{X}} d(\xi^*, \mathbf{x})\,\xi(d\mathbf{x}) \leq m\,.\end{aligned}$$

Also,

$$\begin{aligned}\sum_{i=1}^m \lambda_i &= \mathrm{trace}(\mathbf{H})\\ &= \mathrm{trace}(\mathbf{M}_*\mathbf{M}^{-1}) = \int_{\mathcal{X}} \mathbf{x}^\top\mathbf{M}^{-1}\mathbf{x}\,\xi^*(d\mathbf{x}) \leq \max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top\mathbf{M}^{-1}\mathbf{x} = m + \epsilon\,,\end{aligned}$$

where we used the notation

$$\epsilon = \epsilon(\xi) = \max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top\mathbf{M}^{-1}\mathbf{x} - m \geq 0\,. \tag{3}$$

For $m = 1$ we directly obtain the lower bound $\lambda_1 \geq \lambda_1^* = 1$. For $m > 1$, the Lagrangian for the minimisation of λ_1 subject to $\sum_{i=1}^m \lambda_i^{-1} \leq m$ and $\sum_{i=1}^m \lambda_i \leq m + \epsilon$ is given by

$$\mathcal{L}(\lambda, \mu_1, \mu_2) = \lambda_1 + \mu_1\left(\sum_{i=1}^m \lambda_i^{-1} - m\right) + \mu_2\left(\sum_{i=1}^m \lambda_i - m - \epsilon\right)$$

with $\lambda = (\lambda_1, ..., \lambda_m)^\top$, $\mu_1, \mu_2 \geq 0$. The stationarity of $\mathcal{L}(\lambda, \mu_1, \mu_2)$ with respect to the λ_i's and the Kuhn-Tucker conditions

$$\mu_1 \left(\sum_{i=1}^{m} \lambda_i^{-1} - m \right) = 0, \; \mu_2 \left(\sum_{i=1}^{m} \lambda_i - m - \epsilon \right) = 0$$

give $\lambda_i = L$ for $i = 2, ..., m$, with λ_1 and L satisfying

$$\begin{aligned} &\lambda_1^{-1} + (m-1)L^{-1} = m \\ &\lambda_1 + (m-1)L = m + \epsilon \,. \end{aligned}$$

The solution is thus

$$\lambda_1^* = 1 + \frac{\epsilon}{2} - \frac{\sqrt{\epsilon(4 + \epsilon - 4/m)}}{2} \leq 1 \tag{4}$$

and $\lambda_i^* = L^* = (m-1)/(m - 1/\lambda_1^*) \geq 1$, $i = 2, ..., m$. Notice that the bound (4) gives $\lambda_1^* = 1$ when $m = 1$ and can thus be used for any dimension $m \geq 1$. By substituting λ_1^* for λ_1 in (2) we obtain the following result.

Theorem 2 *For any design $\xi \in \Xi^+$, any point $\mathbf{x}_* \in \mathcal{X}$ such that*

$$d(\xi, \mathbf{x}_*) < h_m(\epsilon) = m \left[1 + \frac{\epsilon}{2} - \frac{\sqrt{\epsilon(4 + \epsilon - 4/m)}}{2} \right] \tag{5}$$

where $\epsilon = \max_{\mathbf{x} \in \mathcal{X}} d(\xi, \mathbf{x}) - m$, cannot be a support point of a D-optimum design measure.

The inequality in (Pronzato, 2003) uses

$$\tilde{h}_m(\epsilon) = m \left[1 + \frac{\epsilon}{2} - \frac{\sqrt{\epsilon(4 + \epsilon)}}{2} \right] . \tag{6}$$

Notice, that $m \geq h_m(\epsilon) > \tilde{h}_m(\epsilon)$ for all integer $m \geq 1$ and all $\epsilon > 0$, and that $\lim_{\epsilon \to \infty} h_m(\epsilon) = 1$ while $\lim_{\epsilon \to \infty} \tilde{h}_m(\epsilon) = 0$. The new bound is thus always stronger, especially for large values of ϵ, i.e. when the design ξ is far from being optimum. Although in practice the improvement over (6) can be marginal, see the example below, the important result here is that the bound (5) cannot be improved. Indeed, when $m = 1$, $h_1(\epsilon) = 1$ for any $\epsilon > 0$ which is clearly the best possible bound. When $m \geq 2$, $h_m(\epsilon)$ is the tightest lower bound on the variance function $d(\xi, \mathbf{x}_*)$ at a D-optimal support point $\mathbf{x}_*$ that depends only on m and ϵ, in the sense of the following theorem.

Theorem 3 *For any integer $m \geq 2$ and any $\epsilon, \delta > 0$ there exist a compact design space $\mathcal{X} \subset \mathbb{R}^m$, a design ξ on $\mathcal{X}$ and a point $\mathbf{x}_* \in \mathcal{X}$ supporting a D-optimum design on $\mathcal{X}$ such that $\epsilon = \max_{\mathbf{x}\in\mathcal{X}} d(\xi, \mathbf{x}) - m$ and*

$$d(\xi, \mathbf{x}_*) < h_m(\epsilon) + \delta \, .$$

Proof. Denote $h = h_m(\epsilon)$ and $k = 2^{m-1}$. Let $\mathbf{x}_1, ..., \mathbf{x}_k$ correspond to the k vectors of $\mathbb{R}^m$ of the form

$$\left(\sqrt{\frac{1}{h}}, \pm\sqrt{\frac{h-1}{h(m-1)}}, \ldots, \pm\sqrt{\frac{h-1}{h(m-1)}} \right)^\top$$

and let $\mathbf{y}_1, \ldots, \mathbf{y}_k$ correspond to the k vectors $\left(\sqrt{1/m}, \pm\sqrt{1/m}, \ldots, \pm\sqrt{1/m}\right)^\top$. Take $\mathbf{x}_* = (\sqrt{b}, 0, \ldots, 0)^\top \in \mathbb{R}^m$ with $1 < b < \min\{(\epsilon + m)/h, (h+\delta)/h\}$, $\mathcal{X}$ as the finite set $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_k, \mathbf{y}_1, ..., \mathbf{y}_k, \mathbf{x}_*\}$ and let ξ be the uniform probability measure on $\mathbf{x}_1, ..., \mathbf{x}_k$. Note that $\mathbf{M}(\xi)$ is a diagonal matrix with diagonal elements $(1/h, (h-1)/[h(m-1)], \ldots, (h-1)/[h(m-1)])$. One can easily verify that

$$\max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top \mathbf{M}^{-1}(\xi)\mathbf{x} - m = \epsilon \text{ and } d(\xi, \mathbf{x}_*) = \mathbf{x}_*^\top \mathbf{M}^{-1}(\xi)\mathbf{x}_* = b\,h < h_m(\epsilon) + \delta \, .$$

The uniform probability measure η on $\mathbf{y}_1, ..., \mathbf{y}_k$ is D-optimum on $\mathcal{X}/\{\mathbf{x}_*\}$, as can be directly verified by checking (ii) of the Equivalence Theorem 1. On the other hand, η is not D-optimum on $\mathcal{X}$ since $\mathbf{x}_*^\top \mathbf{M}^{-1}(\eta)\mathbf{x}_* = b\,m > m$, which implies that $\mathbf{x}_*$ must support a D-optimum design on $\mathcal{X}$. ■

Example: We consider a series of problems defined by the construction of the minimum covering ellipse for an initial set of 1000 random points in the plane, i.i.d. $\mathcal{N}(0, \mathbf{I}_2)$. These problems correspond to D-optimum design problems for randomly generated $\mathcal{X} \subset \mathbb{R}^3$, see Titterington (1975, 1978). The following recursion can thus be used:

$$w_i^{k+1} = w_i^k \frac{d(\xi^k, \mathbf{x}_i)}{m} \, , \; i = 1, \ldots, q(k) \, , \tag{7}$$

where $k \geq 0$, $w_i^k = \xi^k(\mathbf{x}_i)$ is the weight given by the discrete design ξ^k to the point $\mathbf{x}_i$ and $q(k)$ is the cardinality of $\mathcal{X}$ at iteration k. In the original algorithm, $q(k) = q(0)$ for all k and, initialized at a ξ^0 that gives a positive weight at each point of $\mathcal{X}$, the algorithm converges monotonically to the optimum, see (Torsney, 1983) and (Titterington, 1976). The tests (5) and (6) can be used to decrease $q(k)$: at iteration k, any design point $\mathbf{x}_j$ satisfying $d(\xi^k, \mathbf{x}_j) < h_m[\epsilon(\xi^k)]$, see (3, 5), or $d(\xi^k, \mathbf{x}_j) < \tilde{h}_m[\epsilon(\xi^k)]$, see (3, 6), can be removed from $\mathcal{X}$. The total weight of the points that are cancelled is then reallocated to the $\mathbf{x}_i$'s that stay in $\mathcal{X}$ (e.g., proportionally to w_i^k).

Figure 1 presents a typical evolution of $q(k)$ as a function of $\log(k)$ for ξ^0 uniform on $\mathscr{X}$ and shows the superiority of the test (5) over (6). The improvement is especially important in the first iterations, when the design ξ^k is far from the optimum. Define $k^*(\delta)$ as the number of iterations required to reach a given precision δ,

$$k^*(\delta) = \min\left\{k \geq 0 : \epsilon(\xi^k) < \delta\right\} ,$$

with $\epsilon(\xi^k)$ defined by (3). Notice that from the concavity of $\log\det \mathbf{M}(\xi)$ we have

$$\begin{aligned}\log\det \mathbf{M}(\xi^*) - \log\det \mathbf{M}(\xi^{k^*(\delta)}) &\leq \frac{\partial \log\det \mathbf{M}[(1-\alpha)\xi^{k^*(\delta)} + \alpha\xi^*]}{\partial\alpha}_{|\alpha=0} \\ &= \int_{\mathscr{X}} d(\xi^{k^*(\delta)}, \mathbf{x})\, \xi^*(d\mathbf{x}) - m < \delta\,.\end{aligned}$$

Table 1 shows the influence on the algorithm (7) of the cancellation of points based on the tests (5) and (6), in terms of $k^*(\delta)$, of the corresponding computing time $T(\delta)$, the number of support points $n(\delta)$ of $\xi^{k^*(\delta)}$ and the first iteration k_{10} when ξ^k has 10 support points or less, with $\delta = 10^{-3}$. The results are averaged over 1000 independent problems. The values of $k^*(\delta)$ and k_{10} are rounded to the nearest larger integer, the computing time for the algorithm with the cancellation of points based on (5) is taken as reference and set to 1 (the algorithm without cancellation was at least 4.5 times slower in all the 1000 repetitions). Although cancelling points has little influence on the number of iterations $k^*(\delta)$, is renders the iterations simpler: on average the introduction of the test (5) in the algorithm (7) makes it about 30 times faster.

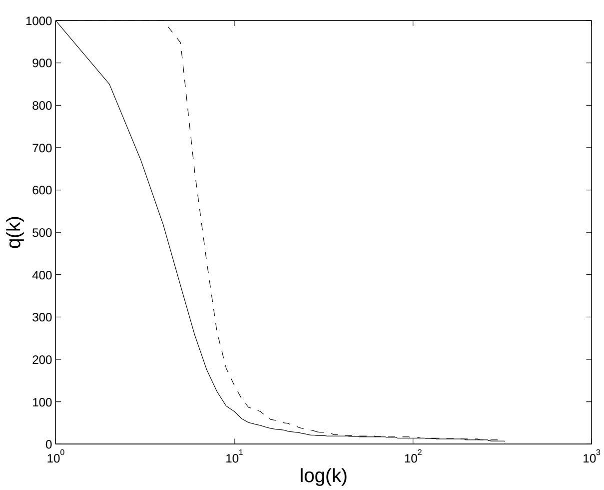

Fig. 1. $q(k)$ as a function of $\log(k)$: cancellation based on (5) in solid line, on (6) in dashed line.

The influence of the cancellation on the performance of the algorithm can be further improved as follows. Let $(k_j)_j$ denote the subsequence corresponding

Algorithm	$k^*(\delta)$	$T(\delta)$	$n(\delta)$	k_{10}
(7)	252	31.6	1000	–
(7) and (6)	248	1.4	5.8	82
(7) and (5)	247	1	5.5	66

Table 1
Influence of the tests (5) and (6) on the average performance of the algorithm (7) for the minimum covering ellipse problem (1000 repetitions, $\delta = 10^{-3}$).

to the iterations where some points are removed from $\mathcal{X}$. We have $j \leq q(0)$, the cardinality of the initial $\mathcal{X}$, and the convergence of the algorithm (7) is therefore maintained whatever the heuristic rule used at the iterations k_j for updating the weights of the points that stay in $\mathcal{X}$ (provided these weights remain strictly positive). The following one has been found particularly efficient on a series of examples: for all $t \in T_j$, the set of indices corresponding to the points that stay in $\mathcal{X}$ at iteration k_j, replace $w_t^{k_j}$ by

$$w_t'^{k_j} = \frac{z_t}{\sum_{s \in T_j} z_s} \quad \text{where } z_t = \begin{cases} A w_t^{k_j} & \text{if } d(\xi^{k_j}, \mathbf{x}_t) \geq m \\ w_t^{k_j} & \text{otherwise} \end{cases}$$

for some $A \geq 1$. A final remark is that by including the test (5) in the algorithm (7) one can in general quickly identify potential support points for an optimum design. When the number n of these points is small enough, switching to a more standard convex-programming algorithm for the optimization of the n associated weights might then form a very efficient strategy.

References

Fedorov, V., 1972. Theory of Optimal Experiments. Academic Press, New York.

Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canadian Journal of Mathematics 12, 363–366.

Pronzato, L., 2003. Removing non-optimal support points in D-optimum design algorithms. Statistics & Probability Letters 63, 223–228.

Titterington, D., 1975. Optimal design: some geometrical espects of D-optimality. Biometrika 62 (2), 313–320.

Titterington, D., 1976. Algorithms for computing D-optimal designs on a finite design space. In: Proc. of the 1976 Conference on Information Science and Systems. Dept. of Electronic Engineering, John Hopkins University, Baltimore, pp. 213–216.

Titterington, D., 1978. Estimation of correlation coefficients by ellipsoidal trimming. Journal of Royal Statistical Society C27 (3), 227–234.

Torsney, B., 1983. A moment inequality and monotonicity of an algorithm.

In: Kortanek, K., Fiacco, A. (Eds.), Proc. Int. Symp. on Semi-infinite Programming and Applications. Springer, Heidelberg, pp. 249–260.