Submitted to the Annals of Statistics

OPTIMAL DISCRIMINATING DESIGNS FOR SEVERAL COMPETING REGRESSION MODELS

By Dietrich Braess and Holger Dette

Ruhr-Universität Bochum

The problem of constructing optimal discriminating designs for a class of regression models is considered. We investigate a version of the T_p-optimality criterion as introduced by Atkinson and Fedorov (1975b). The numerical construction of optimal designs is very hard and challenging, if the number of pairwise comparisons is larger than 2. It is demonstrated that optimal designs with respect to this type of criteria can be obtained by solving (nonlinear) vector-valued approximation problems. We use a characterization of the best approximations to develop an efficient algorithm for the determination of the optimal discriminating designs. The new procedure is compared with the currently available methods in several numerical examples, and we demonstrate that the new method can find optimal discriminating designs in situations where the currently available procedures fail.

1. Introduction. An important problem in optimal design theory is the construction of efficient designs for model identification in a nonlinear relation of the form

$$Y = \eta(x, \theta) + \varepsilon. \tag{1.1}$$

In many cases there exist several plausible models which may be appropriate for a fit to the given data. A typical example are dose-finding studies, where various models have been developed for describing the dose-response relation [Pinheiro, Bretz and Branson (2006)]. Some of these models, which have also been discussed by Bretz, Pinheiro and Branson (2005), are listed in Table 1. In these and similar situations the first step of the data analysis consists in the identification of an appropriate model from a given class of competing regression models.

The optimal design problem for model identification has a long history. Early work can be found in Stigler (1971), who determined designs for discriminating between two nested univariate polynomials by minimizing the volume of the confidence ellipsoid for the parameters corresponding to the

AMS 2000 subject classifications: Primary 62K05; secondary 41A30, 41A50

Keywords and phrases: Optimal design, model discrimination, vector-valued approximation

extension of the smaller model. Several authors have worked on this approach in various other classes of nested models [Dette and Haller (1998) or Song and Wong (1999) among others].

Table 1
Candidate dose response models as a function of dose x.

Model	Full model specification
Linear	$\eta_1(x, \rho_{(1)}) = 60 + 0.56x$
Quadratic	$\eta_2(x, \rho_{(2)}) = 60 + (7/2250)x(600 - x)$
$\mathrm{E}_{\max}$	$\eta_3(x, \rho_{(3)}) = 60 + 294x/(25 + x)$
Logistic	$\eta_4(x, \rho_{(5)}) = 49.62 + 290.51/\{1 + \exp[(150 - x)/45.51]\}$

In a pioneering paper Atkinson and Fedorov (1975a) proposed the T-optimality criterion to construct designs for discriminating between two competing regression models. It provides a design such that the sum of squares for a lack of fit test is large. Atkinson and Fedorov (1975b) extended this approach later for discriminating a selected model η_1 from a class of other regression models, say $\{\eta_2, \ldots, \eta_k\}$, $k \geq 2$. This concept does not require competing nested models and has found considerable attention in the statistical literature [see e.g. Fedorov (1981), Fedorov and Khabarov (1986) for early and Ucinski and Bogacka (2005), López-Fidalgo, Tommasi and Trandafir (2007), Atkinson (2008b,a), Tommasi (2009), Wiens (2009) or Dette, Melas and Shpilev (2012) for some more recent references].

In general, the problem of finding T-optimal designs, either analytically orx numerically, is a very hard and challenging one. Although Atkinson and Fedorov (1975a) indicated some arguments for the convergence of their iterative procedure, there is no evidence that the convergence is sufficiently fast in cases with more than two pairwise comparisons of regression models such that the procedure can be used in those applications.

In the present paper we construct optimal discriminating designs for several competing regression models where none of the models is selected in advance to be tested against all other ones. Let d denote the number of pairwise comparison of interest. In Section 2 we introduce a T_p-optimality criterion, which is a weighted average of d different T-optimality criteria corresponding to these pairs. It is demonstrated in Section 3 that the corresponding optimal design problems are closely related to (nonlinear) vector-valued approximation problems. The support points of optimal discriminating designs are contained in the set of extreme points of a best approximation, and the optimal design can be determined with the knowledge of these points. Because we are only aware of the work of Brosowski (1968) on vector-valued

approximation, we consider this problem in Section 4.

Duality theory is then used to determine not only the points of the support, but also the masses. The theory shows that there exist optimal designs with a support of at most $n+1$ points, where n is the total number of parameters in the competing regression models. We will illustrate by a simple example that the number of support points is usually much smaller. It turns out that this fact occurs in particular for $d \geq 3$ comparisons, and therefore our investigations explain the difficulties in the computation of T-optimal discriminating designs. For this reason we find numerical results in the literature mainly for the cases $d = 1$ and $d = 2$, and advanced techniques are required for the determination of T_p-optimal discriminating designs if $d \geq 3$.

In Sections 5 and 6 we use the theoretical results to develop an efficient algorithm for calculating T_p-optimal discriminating designs. The main idea of the algorithm is very simple and essentially consists of two steps.

(1) The relation to the corresponding vector-valued approximation problem is used to identify a reference set which contains all support of the T_p-optimal discriminating design. This is done by linearizing the optimization problem. A combinatorial argument in connection with dual linear programs determines which points are included in the support of the optimal design.
(2) A linearization of a saddle point problem that is concealed behind the design problem is used for a simultaneous update of all weights.

The implementation of these two steps which are usually iterated is more complicated and described in Section 6. Some comments regarding the convergence and details for the main technical step of the algorithm are given in an online appendix to this paper. In Section 7 we provide several numerical examples and compare our approach with the currently available methods. In particular, we consider the problem of determining optimal discriminating designs for the dose response models specified in Table 1. Here the currently available procedure fails in the case of many pairwise comparisons, while the new method determines a design with high efficiency in less than 10 iteration steps.

2. Preliminaries. Following Kiefer (1974) we consider designs that are defined as probability measures with finite support on a compact design space $\mathcal{X}$. If the design ξ has masses $w_1, \ldots, w_\nu$ at the distinct points $x_1, \ldots, x_\nu$, then observations are taken at these points with the relative proportions given by the masses. Let $\mathcal{M} = \{\eta_1, \ldots, \eta_k\}$ denote a class of possible models for the regression function η in (1.1), where $\theta_{(j)}$ denotes the vector of parameters in model η_j that varies in the set $\Theta^{(j)}$ $(j = 1, \ldots, k)$.

Atkinson and Fedorov (1975b) proposed to select one model in $\mathcal{M}$, say η_1, to fix its vector of parameters $\rho_{(1)}$ and to determine a discriminating design by maximizing

$$\min_{2\le j\le k} T_{1,j}(\xi), \tag{2.1}$$

where

$$T_{1,j}(\xi) := \inf_{\theta_{(j)}\in\Theta^{(j)}} \int_{\mathcal{X}} [\eta_1(x,\rho_{(1)}) - \eta_j(x,\theta_{(j)})]^2 d\xi(x) \quad (2\le j\le k).$$

If the competing regression models $\eta_1,\dots,\eta_k$ are not nested (as in Table 1), it is not clear which model should be fixed in this approach, and it is useful to have more "symmetry" in this concept. For illustration consider the case of two competing non-nested models, say $\eta_i(x,\theta_{(i)}), \eta_j(x,\theta_{(j)})$, and assume that the experimenter can fix a parameter for each model, say $\rho_{(1)}$ and $\rho_{(2)}$. In this case for a given design ξ there exist two T-optimality criteria, say $T_{1,2}$ and $T_{2,1}$, corresponding to the specification of the model η_1 or η_2, respectively, where

$$T_{i,j}(\xi) := \inf_{\theta_{(j)}\in\Theta^{(j)}} \Delta_{i,j}(\theta_{(j)},\xi) = \inf_{\theta_{(j)}\in\Theta^{(j)}} \int_{\mathcal{X}} [\eta_i(x,\rho_{(i)}) - \eta_j(x,\theta_{(j)})]^2 d\xi(x) \tag{2.2}$$

$(i\neq j)$. The first index i in the term $\Delta_{i,j}$ corresponds to the fixed model $\eta_i(x,\rho_{(i)})$, while the minimum in (2.2) is taken with respect to the parameter of the model specified by the index j. The parameter associated to the minimum is denoted as

$$\theta^*_{(i,j)} := \operatorname*{argmin}_{\theta_{(j)}\in\Theta^{(j)}} \Delta_{i,j}(\theta_{(j)},\xi), \tag{2.3}$$

where we assume its existence and do not reflect its dependence on the design ξ and the parameter $\rho_{(i)}$ since this will always be clear from the context. Note that we use the notation $\theta^*_{(i,j)}$ for the parameter corresponding to the best approximation of the model η_i (with fixed paramater $\rho_{(i)}$) by the model η_j.

If a discriminating design has to be constructed for k competing models, there exist $k(k-1)$ expressions of the form (2.2). Let $p_{i,j}$ be given nonnegative weights satisfying $\sum_{i\neq j} p_{i,j} = 1$, then a design ξ^* is called *T_p-optimal discriminating* for the class of models $\mathcal{M} = \{\eta_1,\dots,\eta_k\}$ if it maximizes the functional

$$T(\xi) := \sum_{1\le i\neq j\le k} p_{i,j} T_{i,j}(\xi) = \sum_{1\le i\neq j\le k} p_{i,j} \inf_{\theta_{(j)}\in\Theta^{(j)}} \Delta_{i,j}(\theta_{(j)},\xi) \tag{2.4}$$

[see also Atkinson and Fedorov (1975b)]. Note that the special choice $p_{i,j} > 0$ $(j = 2, \ldots, k)$, $p_{i,j} = 0$ $(i = 2, \ldots, k,\ j = 2, \ldots, k;\ i \neq j)$, refers to the case where one model (namely η_1) has been fixed and is tested against all other ones. The criterion (2.4) provides a more symmetric formulation of the general discriminating design problem. It has also been investigated by Tommasi and López-Fidalgo (2010) among others for $k = 2$ competing regression models. They proposed to maximize a weighted mean of efficiencies which is equivalent to the criterion (2.4) if the weights $p_{i,j}$ are chosen appropriately.

In order to deal with the general case we denote the set of indices corresponding to the positive weights in (2.4) as

$$\mathcal{I} := \{(i,j) \mid p_{i,j} > 0;\ 1 \leq i \neq j \leq k\}.$$

We assume without loss of generality that the set $\mathcal{I}$ can be decomposed in $p \leq k$ subsets of the form $\{(i,j) \in \mathcal{I} \mid 1 \leq j \leq k\}$ and define $\mathcal{I}_i = \{j \in \{1, \ldots, k\} \mid (i,j) \in \mathcal{I}\}$ as the set of indices corresponding to those models which are used for a comparison with model η_i. For each model η_i, $(i = 1, 2, \ldots, p)$, a parameter, say $\rho_{(i)}$, is fixed due to prior information and the model $\eta_i(x, \rho_{(i)})$ has to be discriminated from the other ones in the set $\mathcal{I}_i$. Define

$$\lambda_i := \#\mathcal{I}_i, \quad d := \sum_{i=1}^{p} \lambda_i \tag{2.5}$$

as the cardinality of the sets $\mathcal{I}_i$ and $\mathcal{I}$, respectively. Note that d denotes the total number of pairwise comparisons included in the optimality criterion (2.4). Consider the space $\mathcal{F}_d = C(\mathcal{X})^d$ of continuous vector-valued functions defined on $\mathcal{X}$, and define for a function $g = (g_{ij})_{(i,j)\in\mathcal{I}} \in \mathcal{F}_d$ a norm by

$$\|g\| := \sup_{x\in\mathcal{X}} |g(x)|, \tag{2.6}$$

where $|g(x)|^2 := \sum_{(i,j)\in\mathcal{I}} p_{i,j}\, g_{ij}^2(x)$ denotes a *weighted Euclidean norm* on $\mathbb{R}^d$. In this framework the distance between two functions $f, g \in \mathcal{F}_d$ is given by $\|f - g\|$. Next, given the parameters $\rho_{(1)}, \ldots, \rho_{(p)}$ for the models $\eta_1, \ldots, \eta_p$,respectively, due to prior information, define the d-dimensional vector-valued function

$$\eta(x) := \Big(\underbrace{\eta_1(x, \rho_{(1)}), \ldots, \eta_1(x, \rho_{(1)})}_{\lambda_1 \text{times}}, \ldots, \underbrace{\eta_p(x, \rho_{(p)}), \ldots, \eta_p(x, \rho_{(p)})}_{\lambda_p \text{times}}\Big)^T, \tag{2.7}$$

where each function $\eta_j(x,\rho_{(j)})$ appears λ_j times in the vector $\eta(x)$. We also consider a vector of approximating functions

$$\eta(x,\theta) := \Big(\underbrace{(\eta_j(x,\theta_{(1,j)}))_{j\in\mathcal{I}_1}}_{\in\mathbb{R}^{\lambda_1}},\ldots,\underbrace{(\eta_j(x,\theta_{(p,j)}))_{j\in\mathcal{I}_p}}_{\in\mathbb{R}^{\lambda_p}}\Big)^T \in \mathcal{F}_d\,. \tag{2.8}$$

We emphasize again that we use the notation $\theta_{(i,j)}$ for the parameter in the model η_j. This means that different parameters $\theta_{(i,j)}$ and $\theta_{(k,j)}$ are used if the model η_j has to be discriminated from the models η_i and η_k $(i \neq k)$. The corresponding parameters are collected in the vector

$$\theta = \Big((\theta_{(1,j)})_{j\in\mathcal{I}_1},\ldots,(\theta_{(p,j)})_{j\in\mathcal{I}_p}\Big)^T \in \Theta = \otimes_{i=1}^p \otimes_{j\in\mathcal{I}_i} \Theta^{(j)}, \tag{2.9}$$

and we denote by $n := \dim\Theta = \sum_{i=1}^p \sum_{j\in\mathcal{I}_i} \dim\Theta^{(j)}$ the total number of all parameters involved in the T_p-optimal discriminating design problem. With this notation the optimal design problem can be rewritten as

$$\max_{\xi} \sum_{1\le i\neq j\le p} p_{i,j} \min_{\theta_{(j)}\in\Theta^{(j)}} \Delta_{i,j}(\theta_{(j)},\xi) \tag{2.10}$$

and the following examples illustrate this general setting.

Example 2.1 Consider the case $k = 3$ and assume that all weights $p_{i,j}$ in the criterion (2.4) are positive. Here no model is preferred, and there are 6 pairwise comparisons. This yields $p = k = 3$,

$$\begin{aligned} \mathcal{I} &= \{(1,2),(1,3),(2,1),(2,3),(3,1),(3,2)\},\\ \mathcal{I}_1 &= \{2,3\},\quad \mathcal{I}_2 = \{1,3\},\quad \mathcal{I}_3 = \{1,2\}, \end{aligned}$$

$\lambda_1 = \lambda_2 = \lambda_3 = 2$ and $d = 6$. We obtain for the vectors in (2.7) and (2.8)

$$\begin{aligned} \eta(x) &= \big(\eta_1(x,\rho_{(1)}),\eta_1(x,\rho_{(1)}),\eta_2(x,\rho_{(2)}),\eta_2(x,\rho_{(2)}),\\ &\qquad\qquad \eta_3(x,\rho_{(3)}),\eta_3(x,\rho_{(3)})\big)^T,\\ \eta(x,\theta) &= \big(\eta_2(x,\theta_{(1,2)}),\eta_3(x,\theta_{(1,3)}),\eta_1(x,\theta_{(2,1)}),\eta_3(x,\theta_{(2,3)}),\\ &\qquad\qquad \eta_1(x,\theta_{(3,1)}),\eta_2(x,\theta_{(3,2)})\big)^T, \end{aligned}$$

with

$$\theta = (\theta_{(1,2)},\theta_{(1,3)},\theta_{(2,1)},\theta_{(2,3)},\theta_{(3,1)},\theta_{(3,2)})^T \in \Theta^{(2)}\times\Theta^{(3)}\times\Theta^{(1)}\times\Theta^{(3)}\times\Theta^{(1)}\times\Theta^{(2)}.$$

Example 2.2 Consider the problem of discriminating between $k = 3$ nested polynomial models $\eta_1(x, \theta_{(1)}) = \theta_{10} + \theta_{11}x$, $\eta_2(x, \theta_{(2)}) = \theta_{20} + \theta_{21}x + \theta_{22}x^2$ and $\eta_3(x, \theta_{(3)}) = \theta_{30} + \theta_{31}x + \theta_{32}x^2 + \theta_{33}x^3$. A common strategy to identify the degree of the polynomial is to test a quadratic against a linear and a cubic against the quadratic model. In this case we choose only two positive weights $p_{2,1}$ and $p_{3,2}$ in the criterion (2.4) which yields $\mathcal{I} = \{(2,1),(3,2)\}$, $\mathcal{I}_1 = \{1\}$, $\mathcal{I}_2 = \{2\}$, $p = 2$, $\lambda_1 = \lambda_2 = 1$ and $d = 2$. The functions η and $\eta(\cdot, \theta)$ are given by $\eta(x) = (\eta_2(x, \rho_{(2)}), \eta_3(x, \rho_{(3)}))^T$,

$$\eta(x,\theta) \;=\; (\eta_1(x,\theta_{(2,1)}), \eta_2(x,\theta_{(3,2)}))^T = (\theta_{10} + \theta_{11}x,\, \theta_{20} + \theta_{21}x + \theta_{22}x^2)^T,$$

respectively, where $\theta = (\theta_{(2,1)}, \theta_{(3,2)})^T \in \mathbb{R}^5$.

3. Characterization of optimal designs. The T_p-optimality of a given design ξ can be checked by an equivalence theorem (Theorem 3.1) that can be proved by the same arguments as used by Atkinson and Fedorov (1975b). As usual, the following properties tacitly are assumed to hold.

A1. The regression functions $\eta_i(x, \theta_{(i)})$ are differentiable with respect to the parameter $\theta_{(i)}$ $(i = 1, \ldots, k)$

A2. Let ξ^* be a T_p-optimal discriminating design. The parameter $\theta^* = (\theta^*_{(i,j)})^T_{(i,j)\in\mathcal{I}}$ defined by (2.3) exists, is unique and an interior point of Θ.

Both assumptions are always satisfied in linear models. Moreover, assumption A1 is satisfied for many commonly used nonlinear regression models [see Seber and Wild (1989)]. It is usually harder to check Assumption A2 because it depends on the individual T_p-optimal design.

Theorem 3.1 *(Equivalence Theorem) A design ξ is a T_p-optimal discriminating design for the class of models $\mathcal{M}$ if and only if for all $x \in \mathcal{X}$*

$$\psi(x,\xi) := \sum_{(i,j)\in\mathcal{I}} p_{i,j}[\eta_i(x,\rho_i) - \eta_j(x,\theta^*_{(i,j)})]^2 \le T(\xi), \tag{3.1}$$

where $\theta^*_{(i,j)}$ is defined by (2.3)*. Moreover, if ξ is a T_p-optimal discriminating design, then equality holds in* (3.1) *for all support points of ξ.*

The Equivalence Theorem asserts that there is no gap between the solution of the max min problem (2.10) and the corresponding min max problem. The following result shows that the T_p-optimal design problem is intimately related to a nonlinear vector-valued approximation problem with respect to the norm (2.6).

Theorem 3.2 *Let η, $\eta(\cdot,\theta)$ and $T(\xi)$ be defined by (2.8), (2.9) and (2.4), respectively, then*

$$\sup_{\xi} T(\xi) = \inf_{\theta\in\Theta} \|\eta - \eta(\cdot,\theta)\|^2, \tag{3.2}$$

i.e., with ψ defined in the equivalence theorem

$$\sup_{\xi}\inf_{\theta\in\Theta}\int_{\mathcal{X}} \psi(x,\xi)d\xi = \inf_{\theta\in\Theta}\sup_{\xi}\int_{\mathcal{X}} \psi(x,\xi)d\xi. \tag{3.3}$$

If ξ^* maximizes $T(\xi)$, then the vector $\theta^* = (\theta^*_{(i,j)})_{(i,j)\in\mathcal{I}}$ defined in (2.3) satisfies

$$\|\eta(x) - \eta(x,\theta^*)\| = \inf_{\theta\in\Theta}\|\eta(x) - \eta(x,\theta)\| = T(\xi^*). \tag{3.4}$$

Moreover, the support of the T_p-optimal discriminating design ξ^* for the class $\mathcal{M}$ satisfies

$$\operatorname{supp}(\xi^*) \subset \mathcal{A} := \Big\{ x\in\mathcal{X} \;\Big|\; |\eta(x) - \eta(x,\theta^*)| = \|\eta - \eta(\cdot,\theta^*)\| \Big\}. \tag{3.5}$$

Proof. We have for any design $\tilde{\xi}$ the relation

$$\inf_{\theta\in\Theta}\int_{\mathcal{X}} \psi(x,\xi)d\tilde{\xi} \le \inf_{\theta\in\Theta}\sup_{\xi}\int_{\mathcal{X}} \psi(x,\xi)d\xi,$$

and the left-hand side of (3.3) cannot be larger than the right-hand side, i.e., $T(\tilde{\xi}) \le \inf_{\theta\in\Theta}\|\eta - \eta(\cdot,\theta)\|^2$. Since $\tilde{\xi}$ is an arbitrary design, the bound holds also for $\sup_\xi T(\xi)$. This means in terms of (3.2) $\sup_\xi T(\xi) \le \inf_{\theta\in\Theta}\|\eta - \eta(\cdot,\theta)\|^2$.

Now the characterization of T_p-optimality in Theorem 3.1 and the definition of $\theta^* = (\theta^*_{(i,j)})_{(i,j)\in\mathcal{I}}$ in Theorem 3.1 yield

$$\begin{aligned} T(\xi^*) &\le \inf_{\theta\in\Theta}\|\eta - \eta(\cdot,\theta)\|^2 \le \|\eta - \eta(\cdot,\theta^*)\|^2 \\ &= \sup_{x\in\mathcal{X}} \sum_{(i,j)\in\mathcal{I}} p_{i,j}[\eta_i(x,\rho_{(i)}) - \eta_j(x,\theta^*_{(i,j)})]^2 \le T(\xi^*) \end{aligned}$$

which proves the first part of Theorem 3.2. The statement on the support points of ξ^* follows directly from these considerations. □

Equality (3.4) means that the parameter θ^* defined in (2.3) corresponds to the best approximation of the function η in (2.7) by functions of the

form (2.8) with respect to the norm (2.6). If this nonlinear approximation problem has been solved and the parameter $\bar\theta = ((\bar\theta_{(i,j)})_{j\in\mathcal{I}_1}, \dots, (\bar\theta_{(p,j)})_{j\in\mathcal{I}_p})$ corresponds to a best approximation, i.e.,

$$\|\eta - \eta(\cdot,\bar\theta)\|^2 = \min_{\theta\in\Theta} \|\eta - \eta(\cdot,\theta)\|^2, \tag{3.6}$$

it follows from Theorem 3.2 that the support of the T_p-optimal discriminating design is contained in the set $\mathcal{A}$ defined in (3.5). In linear models and in many of the commonly used nonlinear regression models θ^* and $\bar\theta$ are uniquely determined.

Example 3.3 In Example 2.1 we considered discriminating design problems for 3 rival models η_1, η_2, η_3 and all weights in the optimality criterion are positive. By Theorem 3.2 the support of the T_p-optimal discriminating design problem can be found by solving the nonlinear vector-valued approximation problem

$$\inf_{\theta\in\Theta} \|\eta - \eta(\cdot,\theta)\|^2 = \inf\Big\{\sup_{x\in\mathcal{X}} \sum_{1\le i\ne j\le 3} p_{i,j} |\eta_i(x,\rho_{(i)}) - \eta_j(x,\theta_{(i,j)})|^2 \ \Big| \ \theta_{(i,j)} \in \Theta^{(j)};\ 1\le i\ne j\le 3\Big\}.$$

The following result is an approach in this framework for the calculation of the masses of the T_p-optimal discriminating design.

Corollary 3.4 *Assume that a parameter $\bar\theta$ defined in* (3.6) *exists and is an interior point of Θ, and let $\nabla_{\theta_{(i,j)}}$ denote the gradient of η_j with respect to $\theta_{(i,j)}$.*
(a) If a design ξ is a T_p-optimal discriminating design for the class $\mathcal{M}$, then

$$\int_{\mathcal{A}} \Big(\eta_i(x,\rho_{(i)}) - \eta_j(x,\bar\theta_{(i,j)})\Big) \nabla_{\theta_{(i,j)}} \eta_j(x,\theta_{(i,j)})\Big|_{\theta_{(i,j)}=\bar\theta_{(i,j)}} d\xi(x) = 0 \tag{3.7}$$

holds for all $(i,j)\in\mathcal{I}$.

(b) Conversely, if all competing models are linear, and the design ξ satisfies (3.7) *such that* $\operatorname{supp}(\xi)\subset\mathcal{A}$*, then ξ is a T_p-optimal discriminating design for the class $\mathcal{M}$.*

Proof. If condition (3.7) is not satisfied, there is a direction in the parameter space Θ in which the criterion decreases. Thus (3.7) is a necessary condition. From Theorem 3.2 we know that the best approximation gives rise to a T_p-optimal design, and it follows from a uniqueness argument that the condition is also sufficient in this case. □

4. Chebyshev approximation of d-variate functions. By Theorem 3.2 a T_p-optimal discriminating design is associated to an approximation problem in the space of continuous d-variate functions on the compact design space $\mathcal{X}$ where d is the number of comparisons as specified by (2.5). This relation can be used for the computation of T_p-optimal designs and for the evaluation of the efficiency of computed designs.

In this section we will investigate these approximation problems in more detail for the case of linear models. We restrict the presentation to linear models because we want to emphasize that the main difficulties already appear in linear models if $d \geq 3$. The extension to nonlinear regression models is straightforward and will be provided in Section 6.3.

The general theory here and in the previous section provides only the information that a T_p-optimal discriminating design exists with $n+1$ or less support points where $n = \dim \Theta$. We will demonstrate in Section 4.2 that the number of support points is often much smaller than $n+1$. This is the reason for the difficulties in the numerical construction, even if only linear models are involved. In contrast to other methods [see e.g. López-Fidalgo, Tommasi and Trandafir (2007)] the construction via the approximation problem has the advantage that the points of the support of the T_p-optimal discriminating design are directly calculated.

4.1. *Characterization of best approximations.* We will avoid double indices for vectors and vector-valued functions throughout this section in order to avoid confusion with matrices. We write $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ instead of $(\theta_{(i,j)})_{(i,j)\in\mathcal{I}}^T$ and $(f_1, \ldots, f_d)^T$ instead of the vector $\eta(x)$ defined in (2.7). The approximation problem is considered for a given d-variate function $f = (f_1, \ldots, f_d)^T \in \mathcal{F}_d = C(\mathcal{X})^d$. It is not necessary that some components of f are equal as it occurs in the function (2.7).

In the case of linear models, equation (2.8) defines an n-dimensional linear subspace

$$V = \Big\{ v = \sum_{m=1}^{n} \theta_m v_m \,\Big|\, \theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^n \Big\} \subset \mathcal{F}_d \,, \tag{4.1}$$

where $v_1, v_2, \ldots, v_n \in \mathcal{F}_d$ denotes a basis of V and n is the dimension of the parameter space Θ in (2.9). Note that $f(x)$ and $v(x)$ are d-dimensional vectors for $x \in \mathcal{X}$ and $v \in V$. Theorem 3.2 relates the T_p-optimal discriminating design problem to the problem of determining the best Chebyshev approximation u^* of the function f by elements of the subspace V, i.e.,

$$\|f - u^*\| = \min_{v\in V} \|f - v\|.$$

As stated in (2.6), the norm $\|\cdot\|$ refers to the maximum-norm on $C(\mathcal{X})^d$, $\|g\| := \sup_{x\in\mathcal{X}} |g(x)|$, where the weighted Euclidean norm $|\cdot|$ and the corresponding inner product in $\mathbb{R}^d$ are defined by

$$|r|^2 := \sum_{l=1}^{d} p_l\, |r_l|^2, \qquad \langle \tilde{r}, r\rangle := \sum_{l=1}^{d} p_l\, \tilde{r}_l r_l, \quad r, \tilde{r} \in \mathbb{R}^d \tag{4.2}$$

(here the weights p_l correspond to the weights $p_{i,j}$ used in the definition (2.6)). Because the family V defined in (4.1) is a linear space, the classical Kolmogorov criterion [see Meinardus (1967)] can be generalized to the problem of vector-valued approximation. The result is easily obtained from the cited literature if products of real or complex numbers in the proof of the classical theorem are replaced by the Euclidean inner products of d-vectors. The nonlinear character of the procedures for determining best approximations does not matter at this point.

Lemma 4.1 *(Kolmogorov criterion for vector-valued approximation) Let $u \in V$ and*

$$\mathcal{A} := \{x \in \mathcal{X} \mid |\varepsilon(x)| = \|\varepsilon\|\} \tag{4.3}$$

be the set of extreme points of the error function $\varepsilon := f - u$. The d-variate function u is a best approximation to f in V if and only if for all $v \in V$

$$\min_{x\in\mathcal{A}} \langle \varepsilon(x), v(x)\rangle \le 0. \tag{4.4}$$

Assume that u is a best approximation of the function f. Condition (4.4) in the Kolmogorov criterion means that the system of inequalities

$$\langle \varepsilon(x), v_0(x)\rangle > 0 \quad \forall\, x \in \mathcal{A}.$$

is not solvable. Let $v_1, v_2, \ldots, v_n$ be a basis of V. Using the representation

$$v(x) = \sum_{m=1}^{n} \alpha_m v_m(x) \tag{4.5}$$

and setting $r_m(x) := \langle \varepsilon(x), v_m(x)\rangle$ we obtain the unsolvable system

$$\sum_{m=1}^{n} \alpha_m r_m(x) > 0 \quad \forall\, x \in \mathcal{A}, \tag{4.6}$$

for the vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T \in \mathbb{R}^n$. The numbers $r_m(x)$ are considered as the components of a vector $r(x)$, and by the Theorem on linear

inequalities [see Cheney (1966), p. 19] it follows that the system (4.6) is not solvable if and only if the origin in $\mathbb{R}^n$ is contained in the convex hull of the vectors $\{r(x) = (r_1(x), \ldots, r_n(x))^T,\ x \in \mathcal{A}\}$. By Carathéodory's theorem there are $\nu \leq n+1$ points $x_1, \ldots, x_\nu \in \mathcal{A}$ and numbers $w_1, \ldots, w_\nu \geq 0$ such that $\sum_{i=1}^{\nu} w_i = 1$ and

$$\sum_{i=1}^{\nu} w_i r(x_i) = \sum_{i=1}^{\nu} w_i \langle \varepsilon(x_i), v(x_i) \rangle = 0 \quad \forall\ v \in V. \tag{4.7}$$

Theorem 4.2 *(Characterization Theorem) Let $u \in V$ and $\mathcal{A}$ be the set of extreme points of $\varepsilon = f - u$. The following statements are equivalent:*

(i) u is a best approximation to f in V.

(ii) There exist $\nu \leq n+1$ points $x_1, x_2, \ldots, x_\nu \in \mathcal{A}$ such that for all $v \in V$

$$\min_{1 \leq i \leq \nu} \langle \varepsilon(x_i), v(x_i) \rangle \leq 0. \tag{4.8}$$

(iii) There exist $\nu \leq n+1$ points $x_1, x_2, \ldots, x_\nu \in \mathcal{A}$ and ν weights w_1, $w_2, \ldots, w_\nu \geq 0$, $\sum_{i=1}^{\nu} w_i = 1$ such that the functional

$$\ell(g) := \frac{1}{\|\varepsilon\|} \sum_{i=1}^{\nu} w_i \langle \varepsilon(x_i), g(x_i) \rangle \tag{4.9}$$

satisfies

$$\ell(\varepsilon) = \|\varepsilon\|, \quad \|\ell\| = 1, \quad \textit{and} \quad V \subset \ker(\ell), \tag{4.10}$$

where $\ker(\ell) = \{v \in V \mid \ell(v) = 0\}$ denotes the kernel of the linear functional ℓ.

Proof. The equivalence of (i) and (ii) follows from the Kolmogorov criterion. To verify the equivalence with condition (iii), let u^* be a best approximation and $\varepsilon^* = f - u^*$. Define the functional (4.9) with the parameters x_i and w_i from (4.7). By the Cauchy–Schwarz inequality we obtain $\langle \varepsilon^*(x_i), g(x_i) \rangle \leq |\varepsilon^*(x_i)|\, |g(x_i)| \leq \|\varepsilon^*\|\, \|g\|$ with equality if $g = \varepsilon^*$. Since $\sum_i w_i = 1$, it follows that $\ell(g) \leq \|g\|$, again with equality if $g = \varepsilon^*$, and the properties in (4.10) are verified.

Finally, assume that $u \in V$ and a functional with the properties (4.10) exists. We have for any $v \in V$

$$\|f - v\| = \|\ell\|\, \|f - v\| \geq \ell(f - v) = \ell(f - u) + \ell(u - v) = \|f - u\| + 0,$$

and u is a best approximation. □

The extreme points $x_1, x_2, \ldots, x_\nu$ and the masses $w_1, w_2, \ldots, w_\nu$ in Theorem 4.2 define the T_p-optimal discriminating design. This follows from Part (iii) of the theorem that is closely related to condition (3.7) in Corollary 3.4. Indeed, assume that (iii) in the theorem is satisfied and consider a design ξ^* with weights $w_1, w_2, \ldots, w_\nu$ at the points $x_1, x_2, \ldots, x_\nu$. It follows for all $v \in V$ that $\|\varepsilon^*\| \, \ell(v) = \int_{\mathcal{A}} \langle f(x) - u^*(x), v(x) \rangle \, d\xi^*(x) = 0$, and by inserting the elements $v_1, v_2, \ldots, v_n$ of the basis of V we obtain precisely condition (3.7). Consequently, there exists a T_p-optimal discriminating design with at most $n+1$ support points. As we will see in Lemma 5.3, functions satisfying only some of the properties in Theorem 4.2(iii) will also play an important role.

4.2. *The number of support points – the generic case.* By the characterization theorem there exists an optimal design with at most $n+1$ support points. If the number of points in the set $\mathcal{A}$ equals $n+1$, then the masses $w_1, w_2, \ldots, w_{n+1}$ of an optimal design can be calculated by the n equations (3.7) together with the normalization $\int_{\mathcal{A}} d\xi(x) = 1$. In most real-life problems, however, the number of support points is substantially smaller than $n+1$, and we obtain from (3.7) more equations than unknown masses. In this case the problem is ill-conditioned and the numerical computation of the masses will be more sophisticated. The following example illustrates the statement on the support.

Example 4.3 We reconsider Example 2.2 for the polynomial regression models. The weights $p_{2,1}$ and $p_{3,2}$ are chosen as positive numbers. Since all functions are polynomials, we may assume $\mathcal{X} = [-1, +1]$ without loss of generality. A quadratic polynomial f_1 is approximated by linear polynomials in the first component, and a cubic polynomial f_2 is approximated by quadratic polynomials in the second component. Therefore, $V = \mathcal{P}_1 \times \mathcal{P}_2$, where $\mathcal{P}_k$ denotes the set of polynomials of degree $\leq k$.

We note that the character of the approximation problem does not change if we subtract a linear polynomial from f_1 and a quadratic polynomial from f_2. Therefore we can assume that $f(x) = (\rho_2 x^2, \rho_3 x^3)^T$. Symmetry arguments show that the best approximating functions will be polynomials with the same symmetry, and we obtain the reduced approximation problem

$$\min_{\theta_1, \theta_2 \in \mathbb{R}} \sup_{x \in [-1,1]} (p_{2,1} |\rho_2 x^2 - \theta_1|^2 + p_{3,2} |\rho_3 x^3 - \theta_2 x|^2).$$

We now fix the given parameters as $\rho_2 = \rho_3 = 1$ and the weights in the T_p-optimality criterion as $p_{2,1} = p_{3,2} = 1/2$. The best approximation is given by $u^*(x) = (1/2, \, x)^T$, i.e., the first component is the best approximation of the univariate function f_1, and the second component interpolates

f_2 at the extreme points of $f_1 - u_1^*$. The function $\psi(x) = |f(x) - u^*(x)|^2 = (x^6 - x^4 + 1/4)/2$ is depicted in the left part of Figure 1. The support of the T_p-optimal discriminating design ξ^* is a subset of the set of extreme points $\mathcal{A} = \{-1, 0, +1\}$ of the function $|f - u^*|^2$. The linear functional ℓ in Theorem 4.2 is easily determined as $\ell(g) = \sqrt{2}[\frac{1}{4}g(-1) - \frac{1}{2}g(0) + \frac{1}{4}g(1)]$. The Characterization Theorem 4.2 yields the associated T_p-optimal discriminating design

$$\xi^* = \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}, \tag{4.11}$$

where the first line provides the support and the second one the associated masses. The degeneracy is now obvious. The dimension of the set $V \subset \mathcal{F}_2$ is $n = 5$, but the solution of the corresponding approximation problem has only 3 extreme points. This degeneracy is counter intuitive. When univariate functions are approximated by polynomials in $\mathcal{P}_2$, then by Chebyshev's theorem there are at least 4 extreme points. Although our approximation problem with 2-variate functions contains more functions and more parameters, the number of extreme points is smaller.

Note also that the second component is determined by interpolation and not by a direct optimization. The same designs are obtained, whenever $p_{2,1}\rho_2^2 \geq p_{3.2}\rho_3^2$. If this condition does not hold, we may have 4 extreme points, as shown in the right part of Figure 1 for the choice $\rho_2 = 1$; $\rho_3 = 4$. The solution is also degenerate. Here, the location of the support points depends on the value of ρ_3. In the mentioned case we obtain (subject to rounding) the T_p-optimal discriminating design with masses 0.18, 0.32, 0.32, 0.18 at the points -1, -0.48, 0.48 and 1.
The previous example shows that the cardinality of the support depends on the given parameters $\rho_{(1)}, \rho_{(2)}, \ldots, \rho_{(p)}$. The following definition helps to understand which cardinality is found in most cases.

Definition 4.4 *Let ξ^* be a T_p-optimal discriminating design for the given data $\rho_{(1)}, \rho_{(2)}, \ldots, \rho_{(p)}$ with $\nu \leq n+1$ support points. The design ξ^* is called a* generic point *if for all parameters in some neighborhood of $\rho_{(1)}, \rho_{(2)}, \ldots, \rho_{(p)}$ the corresponding T_p-optimal discriminating designs have the same number ν of support points.*

Our numerical experience leads to the following

Conjecture 4.5 *If a T_p-optimal discriminating design is a generic point, then its support consists of*

$$\max_{(i,j)\in\mathcal{I}} \dim \Theta^{(j)} + 1$$

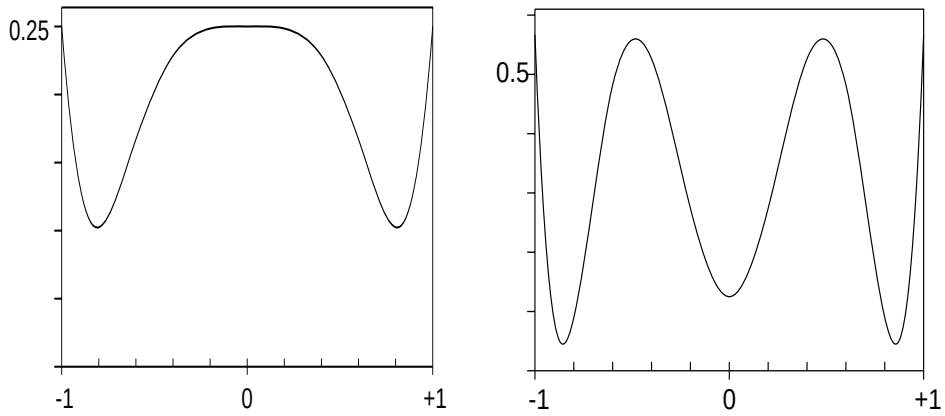

Figure 1. *Error functions $\psi(x) = |f(x) - u^*(x)|^2$ in the equivalence theorem for Example 4.3. Left panel: $\rho_2 = \rho_3 = 1$; right panel: $\rho_2 = 1, \rho_3 = 4$.*

points.

It has been observed in the literature that the number of points in the support can be smaller than $n + 1$ [see e.g. Dette and Titoff (2009)], but computations for $d \leq 2$ do not give the correct impression how large the reduction can be.

5. Linearization and duality. The Equivalence Theorem 3.1 and Theorem 3.2 show that the maximization of $T(\xi)$ is related to a minimization problem. This duality is also reflected in the Characterization Theorem 4.2. We will now consider Newton's iteration for the computation of best approximations.

In each step of the iteration an approximating function u in the family V is improved simultaneously with a *reference set* $\mathcal{S} = \{x_1, x_2, \ldots, x_\nu\}$ that is considered as an approximation of the set $\mathcal{A}$ of extreme points which contains the support of T_p-optimal discriminating designs. Thus we focus on the minimization problem, but we will obtain the associated weights $\{w_1, w_2, \ldots, w_\nu\}$ by duality considerations. Note that in this section we regard duality in connection with the linearized problems and the involved linear programs.

Given a guess u for the approximating function and a finite reference set $\mathcal{S}$, the quadratic term of the correction v in the Binomial formula is temporarily ignored. As usual, let $\varepsilon := f - u$. We replace the optimization problem

(5.1)

$$\max_{x_i \in \mathcal{S}} \big| f(x_i) - u(x_i) - v(x_i) \big|^2 = \max_{x_i \in \mathcal{S}} \big\{ |\varepsilon(x_i)|^2 - 2 \langle \varepsilon(x_i), v(x_i) \rangle + |v(x_i)|^2 \big\} \to \min_{v \in V}!$$

by the linear program

$$\max_{x_i \in \mathcal{S}} \left\{ |\varepsilon(x_i)|^2 - 2 \langle \varepsilon(x_i), v(x_i) \rangle \right\} \to \min_{v \in V}! \tag{5.2}$$

While the left-hand side of (5.1) is obviously bounded from below, this is not always true for the optimization problem (5.2). The boundedness, however, is essential for the algorithm.

Definition 5.1 A function $u \in V$ is called *dual feasible* for the reference set $\mathcal{S}$, if the left-hand side of (5.2) is bounded from below.

The notation of dual feasibility will be clear from the dual linear program (5.4) and Lemma 5.2 below. We will also see in Lemma 5.3 that only the dual feasible functions are associated to a design ξ in the sense of (2.2).

The minimization of a linearized functional on a finite set $\mathcal{S} = \{x_i\}_{i=1}^{\nu}$ with $\nu \geq n+1$ as in (5.2) will be the basis of our algorithm. For a given error function $\varepsilon = f - u$ and a reference set with ν points $x_1, x_2, \ldots, x_\nu$ we may use the representation (4.5) and rewrite the primal problem (5.2) as a linear program for the $n+1$ variables $E, \alpha_1, \alpha_2, \ldots, \alpha_n$:

$$\begin{aligned} E &\to \min! \\ 2 \sum_{m=1}^{n} \alpha_m \langle \varepsilon(x_i), v_m(x_i) \rangle + E &\geq |\varepsilon(x_i)|^2, \quad i = 1, 2, \ldots, \nu. \end{aligned} \tag{5.3}$$

Obviously, there exists a feasible point for this linear program, since the inequalities are satisfied by $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$ and $E = \|\varepsilon\|^2$.

The dual program to (5.3) contains the equations for the ν weights w_1, $w_2, \ldots, w_\nu$ with the adjoint matrix, where we can drop the factor 2 for the sake of simplicity

$$\begin{aligned} \sum_{i=0}^{\nu} w_i |\varepsilon(x_i)|^2 &\to \max! \\ \sum_{i=1}^{\nu} w_i \langle \varepsilon(x_i), v(x_i) \rangle &= 0, \quad \forall\, v \in V, \\ \sum_{i=1}^{\nu} w_i &= 1, \quad w_i \geq 0, \quad i = 1, 2, \ldots, \nu. \end{aligned} \tag{5.4}$$

The following result of duality theory will play an important role [for a proof see Papadimitriou and Steiglitz (1998)].

Lemma 5.2 *The linear program* (5.4) *has a feasible point and a solution if and only if the objective function in the linear program* (5.3) *is bounded from below, i.e.,*

$$\min_{v \in V} \max_{0 \leq i \leq n} \langle \varepsilon(x_i), v(x_i) \rangle \geq 0.$$

If the linear program (5.4) has a feasible point, there is a solution with at most $n+1$ positive weights. We obtain a linear functional ℓ of the form (4.9) with these parameters where $\|\ell\| = 1$ and $V \subset \ker(\ell)$. We have $\ell(\varepsilon) < \|\varepsilon\|$, whenever u is not a best approximation. Since the values of the primal program (5.3) and the dual program (5.4) coincide, we also have

$$E = \sum_{i=1}^{\nu} w_i |\varepsilon(x_i)|^2.$$

The final result of this section shows that the evaluation of the functional T defined in (2.4) for a given design ξ is strongly related to dual feasibility.

Lemma 5.3 *Let $u \in V$ and $\mathcal{S} = \{x_i\}_{i=1}^{\nu}$. The following statements are equivalent.*

(i) The function u is dual feasible for the reference set $\mathcal{S}$.
(ii) There exist nonnegative weights w_i, $i = 1, 2, \ldots, \nu$, such that

$$\sum_{i=1}^{\nu} w_i \langle \varepsilon(x_i), v(x_i) \rangle = 0$$

holds for all $v \in V$.
(iii) There exists a design ξ supported on $\mathcal{S}$ such that

$$u = \operatorname*{argmin}_{v \in V} \int_{\mathcal{X}} |f(x) - v(x)|^2 \, d\xi(x).$$

Proof. The equivalence of (i) and (ii) is a direct consequence of Lemma 5.2. Note that for $t \in \mathbb{R}$ and $v \in V$,

(5.5)
$$\sum_{i=1}^{\nu} w_i |(f - u - tv)(x_i)|^2 = \sum_{i=1}^{\nu} w_i \Big(|\varepsilon(x_i)|^2 - 2t \langle \varepsilon(x_i), v(x_i) \rangle + t^2 |v(x_i)|^2 \Big).$$

If (ii) holds with the weights w_i, then the expression (5.5) attains its minimum at $v = 0$. Hence, u is the solution of the minimization of $\int_{\mathcal{X}} |f(x) - u(x)|^2 \, d\xi(x)$ for the design ξ with the support $\mathcal{S}$ and the masses $w_1, \ldots, w_\nu$ from condition (ii). If (ii) does not hold, then the minimum of (5.5) is not obtained at $t = 0$ for one $v \in V$. Therefore the minimum is not attained at u. □

6. The algorithm. Each step of our iterative procedure consists of two parts. The first part deals with the improvement of the approximating function and the reference set. It focuses on the approximation problem. The

second part is concerned with the computation of the associated masses. The dual linear program is embedded in a saddle point problem. Thus computations for the primal problem and the dual problem may alternate during the iteration. The small number of support points of T_p-optimal discriminating designs (as described in Conjecture 4.5) has impact on both parts.

The iteration starts with a set of parameters $\theta_{(i,j)}$ and a reference set of about $n+1$ points which divide the interval $\mathcal{X}$ into subdomains of equal size. Of course, any prior information may be used for getting a better initial guess.

6.1. *Newton's method and its adaptation.* The improvement of the approximation on a given reference set will be done iteratively by Newton's method. In order to avoid the introduction of an additional symbol, we focus on one step of the iteration for the given input u_0, the corresponding error function $\varepsilon_0 = f - u_0$, and the reference set $\mathcal{S}_0$. The simplest Newton step

> *Given u_0 and $\mathcal{S}_0$, find a solution of the linear program* (5.3) *for $u = u_0$, set $v = \sum_m \alpha_m v_m$.*
> *Take $u_1 = u_0 + v$ as the result of the Newton step*

looks natural, however, it can be only the basis of our algorithm. We take three actions. For convenience, we use the notation $\|g\|_{\mathcal{S}} := \sup_{x\in\mathcal{S}} |g(x)|$.

(1) *Newton steps on subspaces.* Referring to the notation in Section 2 we write the space of approximating functions as a sum of d subspaces

$$V = \oplus_{(i,j)\in\mathcal{I}} V_{(i,j)}\,, \tag{6.1}$$

where $V_{(i,j)}$ contains those functions in V that correspond to $\{\eta_j(\cdot,\theta_{(i,j)}) \mid \theta_{(i,j)} \in \Theta^{(j)}\}$. The linear program that is obtained from (5.3) by the restriction of the functions $\sum_m \alpha_m v_m$ to the subspace $V_{(i,j)}$will be denoted as $(5.3)_{(i,j)}$.
The improvement of the approximation on the reference set will be done iteratively by Newton's method. The linearization (5.2), however, will be considered for the subspaces $V_{(i,j)}$ and not for V. In other words, the d linear programs $(5.3)_{(i,j)}$ are performed separately. It follows from Conjecture 4.5 that we have dual feasibility only on lower dimensional spaces. Indeed, the splitting (6.1) creates dual feasible problems, or the defect is one-dimensional; and the regularization described in item (3) below is the correct remedy. Moreover, another improvement without the splitting will be provided in combination with the evaluation of the masses in Part 2 of the iteration step. (Note that we have the same splitting in the evaluation of θ^* according to (2.3).)

(2) *The damped Newton method.* The Newton correction v will be multiplied by a damping factor t. By definition of the Newton method we have $\max_{x_i \in \mathcal{S}_0}\{|\varepsilon_0(x_i)|^2 - 2\langle \varepsilon_0(x_i), v(x_i)\rangle\} < \|\varepsilon_0\|^2_{\mathcal{S}_0}$ if we have not yet obtained the solution of the actual minimum problem. Since

$$|(f - u_0 - tv)(x_i)|^2 = |\varepsilon_0(x_i)|^2 - 2t\,\langle \varepsilon_0(x_i), v(x_i)\rangle + O(t^2),$$

it follows that $\|f - u_0 - tv)\|^2_{\mathcal{S}_0} < \|\varepsilon_0\|^2_{\mathcal{S}_0}$ for sufficiently small positive factors t; and thus an improvement is generated. Let

$$T := \{1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, \ldots, 2^{-7}, 0\}$$

and determine

$$t = \operatorname*{argmin}_{t \in T} \|f - u_0 - tv\|_{\mathcal{S}_0}\,. \tag{6.2}$$

The standard set of damping factors $1, 2^{-1}, 2^{-2}, \ldots$ has been augmented by the element 0, and therefore the new approximation is at least as good as the old one.

(3) *Regularization by adding a bound.* By definition the objective function E is not bounded from below in the linear program $(5.3)_{(i,j)}$ if u_0 is not feasible with respect to $V_{(i,j)}$. Therefore, we add the restriction $E \geq 0$ to the linear programs.

At the end of this part of the iteration step we have an improved approximation u_1. Extreme points of $f - u_1$ that are not yet obtained in $\mathcal{S}_0$ are added to this set. A decision on the augmentation of the reference set is easy when the error curve is shown on the monitory of the computer. Furthermore, we mark the points in $\mathcal{S}_0$ to which a positive mass was given by the dual linear program associated to $(5.3)_{(i,j)}$ for one pair $(i,j) \in \mathcal{I}$. The points in the reference set are relabel such that $x_1, x_2, \ldots, x_\mu$ are the marked ones.

6.2. *Computation of best designs.* The adapted Newton step in the first part of the iteration step has provided an improved error curve $\varepsilon_1 = f - u_1$ and simultaneously a set of marked points, say $\{x_1, x_2, \ldots, x_\mu\}$. Let ξ be a design with this support and masses $\{w_1, w_2, \ldots, w_\mu\}$ that are not yet known. We look for a correction v with the representation (4.5) such that $u_1 + v$ is associated to $T(\xi)$ in the spirit of (2.2), i.e., we have to minimize

$$\sum_{i=1}^{\mu} w_i \Big|\varepsilon_1(x_i) - \sum_{k=1}^{n} \alpha_k v_k(x_i)\Big|^2 \;=\; \alpha^T A \alpha - 2 w^T R \alpha + b^T w, \tag{6.3}$$

where the elements of the matrices $A = (A_{jk})_{j,k=1,\dots,n}$, $R = (R_{ik})_{i=1,\dots,\mu}^{j=1,\dots,n}$ and the vector $b = (b_i, \dots, b_n)^T$ are defined by

$$\begin{aligned} A_{jk} &:= \sum_i w_i \langle v_j(x_i), v_k(x_i) \rangle , \\ R_{ik} &:= \langle \varepsilon_1(x_i), v_k(x_i) \rangle , \\ b_i &:= |\varepsilon(x_i)|^2. \end{aligned} \tag{6.4}$$

The optimal design among all designs supported at $\{x_1, \dots, x_\mu\}$ is determined by the solution of the saddle point problem

$$\max_w \min_\alpha \{\alpha' A \alpha - 2w' R\alpha + b'w\}, \tag{6.5}$$

where we will ignore the dependence of the matrix A on w for a moment. Reasonable weights in (6.4) will be specified below. The inner optimization problem in (6.5) is solved by

$$A\alpha = R^T w \tag{6.6}$$

and we arrive at the quadratic program

$$\max_w \{-w^T R A^{-1} R^T w + b^T w \mid e^T w = 1,\ w_i \geq 0\}, \tag{6.7}$$

where $e := (1, 1, \dots, 1)^T$ is a μ-vector. In order to check whether all masses are positive, we compute an approximate solution $\tilde{w}$ by solving the linear program

$$\begin{aligned} \sum_{i=1}^{\nu} |(RA^{-1}R^T \tilde{w})_i| &\to \text{min!} \\ \sum_{i=1}^{\nu} \tilde{w}_i &= 1, \quad \tilde{w}_i \geq 0, \quad i = 1, 2, \dots, \mu. \end{aligned} \tag{6.8}$$

We observed in our numerical calculations that all masses are positive, whenever at least 2 points have been marked in Part 1 of the procedure. After removing points with zero mass $\tilde{w}_i$, if necessary, we can ignore the restrictions $w_i \geq 0$, and the problem (6.7) is solved by the linear saddle point equation

$$\begin{pmatrix} A & -R^T & \\ -R & & e \\ & e^T & \end{pmatrix} \begin{pmatrix} \alpha \\ w \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} b \\ 1 \end{pmatrix}. \tag{6.9}$$

Now we are in a position to specify which masses are inserted in (6.4) when the matrix A is calculated. We start with equal masses $w_i = 1/\mu$ for $i = 1, 2, \ldots, \mu$ when we build the matrix for the linear program (6.8). The masses $\tilde{w}_i$ from the linear program are then used in the definition of the matrix A for the saddle point equation (6.9). The solution of (6.9) yields the masses for the improved design ξ. By definition, these masses are used when the criterion $T(\xi)$ is evaluated.

The evaluation of $T(\xi)$ according to (2.4) provides also corrections of the parameters. Let $u_2 = u_1 + v$ be the associated function in V. By definition the sum of weighted squares $\sum_i |\varepsilon(x_i)|^2$ is smaller for u_2 than for u_1. If the errors are nearly equilibrated, it follows that $\max_i |\varepsilon(x_i)|$ will also be smaller for u_2 than for u_1. Therefore, we look for a damping factor t such that the norm of the error $\|f - (u_1 + tv)\|$ is as small as possible. The details of the damping procedure are the same as in the damped Newton method described in Section 6.1.

The value of $T(\xi)$ is a lower bound for the degree of approximation and provides a lower bound of the T_p-efficiency

$$\mathrm{Eff}_{T_p}(\xi) := \frac{T(\xi)}{\sup_\eta T(\eta)} \geq \frac{T(\xi)}{\|\varepsilon\|^2}\,. \tag{6.10}$$

In particular, we have a stopping criterion for the algorithm. The iteration will be stopped if the guaranteed T_p-efficiency is sufficiently close to 1.

6.3. *Adaptation to nonlinear models.* When the models $\eta_1, \eta_2, \ldots, \eta_k$ depend nonlinearly on the parameters, the approximating function $u(x, \theta)$ depends in a (possibly) nonlinear way on the parameter θ. The gradient space defined by

$$\Big\{\eta(\cdot, \tilde{\theta}) + \sum_{k=1}^{n} \alpha_k \frac{\partial}{\partial \theta_k} \eta(\cdot, \theta)|_{\theta=\tilde{\theta}},\ \alpha \in \mathbb{R}^n\Big\} \tag{6.11}$$

is a linear subspace and all the procedures described for linear spaces can be applied to this gradient space. Only the computation of $T(\xi)$ for given ξ requires more effort. The minimization in its definition of $T(\xi)$ can be done by Newton's method. The linearization uses those formulas that are related to the minimization in the gradient space. Thus the algorithm can also deal with nonlinear models.

7. Numerical results. We confirm the efficiency of the new algorithm by numerical results for three examples with linear and nonlinear regression functions. A fourth example can be found in Section C of the online

Table 2
The results of the new algorithm for Example 7.1.

j	$\|\varepsilon_{j,1}\|^2$ Part 1	$\|\varepsilon_{j,2}\|^2$	$T(\xi_j)$ Part 2	$\frac{T(\xi_j)}{\|\varepsilon_{j,2}\|^2}$	support	reference set
0	12.5					$\mathcal{S} = \{-1, -0.5, -0.1, 0, 0.1, 0.5, 1\}$
1	2.3513					
2	0.6092					
3	0.3391	0.2434	0.0146	0.0600	$\{-1, 0.3, 1\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{0.3\}$
4	0.2012	0.1556	0.1144	0.7350	$\{-1, 0.2, 1\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{0.2\}$
5	0.1401	0.1287	0.1029	0.8002	$\{-1, -0.3, 1\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{-0.3\}$
6	0.1268	0.1265	0.1225	0.9685	$\{-1, 0.1, 1\}$	
7	0.1261	0.1260	0.1244	0.9872	$\{-1, -0.05, 1\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{-0.05\}$
8	0.1259	0.1258	0.1246	0.9906	$\{-1, 0, 04\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{0.04\}$

appendix. We also provide a comparison with the algorithm proposed by Atkinson and Fedorov (1975a). Each iteration step is performed in the examples in less than 1 or 2 seconds on a five years old personal computer. The quotient $T(\xi_j)/\|\varepsilon_j\|^2$ in the tables shows the lower bound for the efficiency defined in (6.10). When we distinguish between Part 1 and Part 2 of the iteration step, an index is added to the iteration count. In particular, we distiguish the error functions $\varepsilon_{j,1}$ and $\varepsilon_{j,2}$ obtained in Part 1 and Part 2 of the iteration. The ratio $T(\xi_j)/\|\varepsilon_{j,2}\|^2$ in the tables shows the lower bound for the efficiency defined in (6.10).

Example 7.1 We consider once more Example 2.2, fix $p_{2,1} = p_{3,2} = \frac{1}{2}$, set

$$f(x) = (\eta_2(x, \rho_{(2)}), \eta_3(x, \rho_{(3)}))^T = (1 + x + x^2, 1 + x + x^2 + x^3)^T,$$

and start the algorithm with $u_0 = (0, 0)^T$, i.e., $\theta_{(2,1)} = (0,\, 0)$, $\theta_{(3,2)} = (0,\, 0,\, 0)$. The initial guess u_0 implies that the functions obtained during the iteration do not have the symmetry properties discussed in Example 4.3.

The results of the new algorithm are displayed in Table 2. After 8 iteration steps we obtain a discriminating design with at least 99% efficiency. In the first part of the iteration the lower bound is very small and of no use, but it is increasing rapidly during the iteration. In Figure 2 we display the shape of the error function in the first 3 iterations. We observe that the location of the extreme points changes substantially in the first steps of the algorithm. A comparison with Figure 1 shows that afterwards there are no substantial changes of the shape. The resulting discriminating design puts

Table 3
The results of the algorithm proposed by Atkinson and Fedorov (1975a) in Example 7.1 (left part) and Example 7.2 (right part).

j	$\|\|\psi\|\|$	$T(\xi_j)$	$\frac{T(\xi_j)}{\|\|\psi\|\|}$	$\|\|\psi\|\|$	$T(\xi_j)$	$\frac{T(\xi_j)}{\|\|\psi\|\|}$
1	0.2172	0.1041	0.4791	0.0104	0.0033	0.3150
2	0.3995	0.0743	0.1860	0.0133	0.0034	0.2560
3	0.3189	0.0778	0.2440	0.0241	0.0045	0.1880
4	0.1539	0.1216	0.7903	0.0099	0.0055	0.5583
5	0.1974	0.1195	0.6055	0.0131	0.0055	0.4199
6	0.2337	0.1137	0.4868	0.0094	0.0063	0.6682
7	0.2045	0.1143	0.5592	0.0093	0.0060	0.6471
8	0.1347	0.1240	0.9206	0.0121	0.0062	0.5161
9	0.1732	0.1217	0.7029	0.0079	0.0065	0.8228
10	0.2055	0.1186	0.5773	0.0104	0.0064	0.6153
11	0.1791	0.1199	0.6694	0.0099	0.0064	0.6502
12	0.1356	0.1243	0.9165	0.0091	0.0065	0.7166
13	0.1651	0.1200	0.7267	0.0081	0.0066	0.8223
14	0.1640	0.1229	0.7493	0.0088	0.0065	0.7371
15	0.1714	0.1213	0.7078	0.0097	0.0065	0.6686
16	0.1362	0.1243	0.9130	0.0070	0.0067	0.9550

the masses 0.241, 0.501 and 0.258 at the points -1, 0.04 and 1, respectively. The parameters may be compared with the exact optimal ones $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ at the points given in (4.11). The parameters corresponding to the solution of the nonlinear approximation problem defined by the right hand side of (3.2) are given by $\overline{\theta}_{(2,1)} = (1.501, 1.002)$, $\overline{\theta}_{(3,2)} = (0.996, 1.976, 0.958)$.

For the sake of comparison we also present in the left part of Table 3 the corresponding results for the first 16 iterations of the algorithm proposed by Atkinson and Fedorov (1975a). This method starts with an initial guess, say ξ_0, and computes successively new designs $\xi_1, \xi_2, \ldots$ as follows:

(1) At stage s a point $x_{s+1} \in \mathcal{X}$ is determined such that $\psi(x_{s+1}, \xi_s) =$

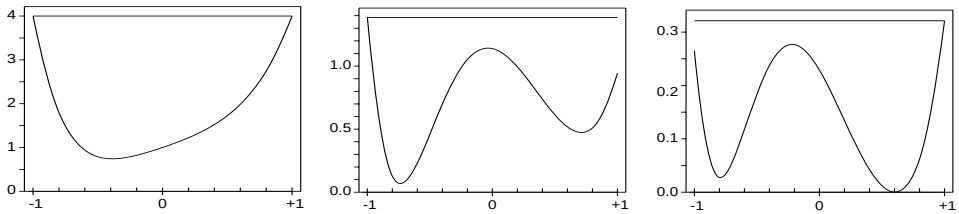

Figure 2. *Error curve $|f - u|^2$ in the first three iteration steps for Example 7.1.*

$\sup_{x\in\mathcal{X}} \psi(x,\xi_s)$, where the function ψ is defined in (3.1).

(2) The updated design ξ_{s+1} is defined by $\xi_{s+1} = (1-\alpha_s)\xi_s + \alpha_s\delta_{x_{s+1}}$, where δ_x is the Dirac measure at point x and $(\alpha_s)_{s\in\mathbb{N}_0}$ is any sequence of positive numbers satisfying $\alpha_s \to 0$; $\sum_{s=0}^{\infty}\alpha_s = \infty$; $\sum_{s=0}^{\infty}\alpha_s^2 < \infty$.

This procedure provides the design

$$\xi = \begin{pmatrix} -1 & -0.2 & -0.1 & 0 & 0.1 & 1 \\ 0.23 & 0.18 & 0.12 & 0.1 & 0.17 & 0.20 \end{pmatrix}$$

in 12 iteration steps, and its efficiency is at least 92%. The final design contains an unnecessarily large support although several design points with low weight have been removed during the computations. Note that neither the sup-norm of the function ψ is decreasing, nor the lower bound $T(\xi)/\|\psi\|$ is increasing. In particular if the iteration is continued, the lower bound for the efficiency of the calculated design is decreasing again. This effect is at first compensated after the 16th iteration, where the bound for the efficiency is 91% (but not 92% as after the 12th iteration). This "oscillating behavior" was also observed in other examples and seems to be typical for the frequently used algorithm proposed by Atkinson and Fedorov (1975a).

Table 4
The results of the new algorithm for Example 7.2.

j	$\|\varepsilon_{j,1}\|^2$	$\|\varepsilon_{j,2}\|^2$	$T(\xi_j)$	$\frac{T(\xi_j)}{\|\varepsilon_{j,2}\|^2}$	support	reference set
	Part 1		Part 2			
0	1.25301					$\mathcal{S} \leftarrow \{1,2,4,$ $6,8,10\}$
1	0.040044					
2	0.012839	0.008738	0.005404	0.6184	$\{0.7, 4.5, 10\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{0.7, 4.5\}$
3	0.006957	0.006827	0.006757	0.9897	$\{0.5, 3.6, 10\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{0.5, 3.6\}$
4	0.006805	0.006797	0.006786	0.9996	$\{0.5, 3.4, 10\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{3.4\}$
5	0.006793	0.006789	0.006786	0.9999		
6	0.006788	0.006787	0.006786	0.9999		

Example 7.2 In order to demonstrate that the algorithm can be used when dealing with nonlinear regression models, we consider two rival models $\eta_1(x,\theta) = \frac{\theta_{11}x}{x+\theta_{12}}$, $\eta_2(x,\theta) = \theta_{21}(1-e^{-\theta_{22}x})$, where $\rho_{(1)} = (2.0,\ 1.0)$ and $\rho_{(2)} = (2.5,\ 0.5)$. The weights in the criterion (2.4) are $p_{1,2} = p_{2,1} = 1/2$. The Newton method is started with $\theta_{(1,2)} = (1,1)$, $\theta_{(2,1)} = (2, 0.5)$, and $\mathcal{S} = \{1,2,4,6,8,10\}$. The degree of approximation is close to the optimum

already after 6 iteration steps, and the guaranteed efficiency is 99.9%. The resulting design has masses 0.311, 0.415 and 0.274 at the points 0.5, 3.4 and 10.0, respectively, while the parameters of the solution of the approximation problem on the right hand side of (3.2) are given (subject to rounding) by the parameters $\bar{\theta}_{(1,2)} = (3.008, 1.809)$, and $\bar{\theta}_{(2,1)} = (1.721, 0.865)$. The determination of the parameter θ^* that minimizes $T(\xi)$ as defined in (2.3) is done by Newton's method. It yields the best θ in a neighborhood of the computed solution. Therefore, we have also performed an extensive global search for the minimum and found a minimum that equals the result of Newton's method up to rounding errors. Now, the plot of the corresponding function ψ in the equivalence Theorem 3.1 is shown in Figure 3. We see that the design is in fact T_p-optimal discriminating. Note that the support of the resulting design consists of 3 points in accordance with Conjecture 4.5. The corresponding results for the algorithm proposed by Atkinson and Fedorov (1975a) are displayed in the right part of Table 3. The algorithm needs 16 iterations in order to find a design with masses 0.32, 0.03, 0.21, 0.12, 0.06, 0.27 at the (unnecessarily large set of) points 0.5, 3.0, 3.3, 3.4, 3.8, 10. Here the lower bound of the efficiency is only 95.5% if we take the best information from the previous steps. The new algorithm is obviously much faster.

Example 7.3 We consider T_p-optimal discriminating designs for the four competing dose-response models listed in Table 1 in the introduction and the design space $\mathcal{X} = [0, 500]$. Here, $d = 6$ comparisons and $n = 15$ parameters are involved. Moreover, the model η_4 is nonlinear. We use the weights $p_{i,j} = 1/6$ if $i > j$ and $p_{i,j} = 0$ otherwise in the criterion (2.4).

The corresponding results are displayed in Table 5, which shows that only 9 iteration steps are required in order to obtain a design with at least 99.9% efficiency. The resulting T_p-optimal discriminating design puts masses 0.255, 0.212, 0.358, 0.175 at the points 0, 78, 245 and 500, respectively. We finally

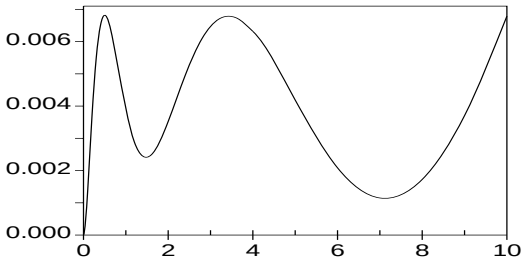

Figure 3. *The function ψ in the equivalence Theorem 3.1 for Example 7.2.*

Table 5
The results of the new algorithm for Example 7.3.

j	$\|\varepsilon_{j,1}\|^2$ Part 1	$\|\varepsilon_{j,2}\|^2$	$T(\xi_j)$ Part 2	$\frac{T(\xi_j)}{\|\varepsilon_{j,2}\|^2}$	support	reference set
0	16661					$\mathcal{S} \leftarrow \{0, 30, 60, 90 \ldots, 450, 500\}$
1	12646					
2	9727	8923	275	0.0309	$\{0, 50, 290, 450\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{50, 290\}$
3	8246	6901	764	0.1108	$\{0, 60, 290, 450\}$	
4	5835	5081	2462	0.4846	$\{0, 70, 260, 500\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{70, 260\}$
5	4543	4170	3016	0.7233	$\{0, 80, 250, 500\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{80, 250\}$
6	4048	3619	3168	0.8754	$\{0, 80, 240, 500\}$	$\mathcal{S} \leftarrow \{0, 70, 80, 240, 250, 500\}$
7	3446	3270	3194	0.9989		
8	3201	3199	3195	0.9980	$\{0, 78, 240, 500\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{78\}$
9	3197	3196	3195	0.9998		

note that we were not able to find a design with a guaranteed efficiency of 80% using the algorithm proposed by Atkinson and Fedorov (1975a).

8. Concluding remarks. Our main theoretical result relates T_p-optimal discriminating designs to an approximation problem for vector-valued functions (Theorem 3.2). By duality theory we show that there exist T_p-optimal designs with at most $n+1$ support points, where n is the number of parameters in the approximation problem (which coincides with the total number of parameters of all regression functions used in the comparisons). These results are sufficient if we are interested only one or two comparisons among the rival models. In this case the computations can be done by an exchange-type algorithm that was already proposed by Atkinson and Fedorov (1975a). This procedure is still the common tool for dealing with design problems whenever $d = 1$ or $d = 2$.

The situation is different and the construction of T_p-optimal discriminating designs becomes extremely difficult and challenging if three or more comparisons are involved. The number of support points can now be much smaller than $n+1$, where n is the total number of parameters of the models involved in the T_p-optimality criterion. Although a reduction of this number was already observed in the case $d = 2$, the amount of the reduction and its impact become clear only when optimal discriminating design problems with $d \geq 3$ pairwise comparisons are studied. For example, we have $n = 15$ parameters in the dose-finding problems listed in Table 1, but the support

of the T_p-optimal discriminating design consists of only 4 points.

Therefore, there are substantial differences between our new algorithm and the generalization of the method by Atkinson and Fedorov (1975a) beyond the case $d = 1$. Our algorithm is based on the related approximation problem (Theorem 3.2), and additionally we also add combinatorial aspects (addition (iii) in Section 6.1), which accelerate the speed of convergence. Dual linear programs associated to small subproblems determine the support of the resulting design and prevent the algorithm from providing designs with too many support points. The masses are simultaneously computed by a stabilized version of the equations in Corollary 3.4, while the commonly used algorithms in each iteration step involve an update of the mass at only one point and a renormalization.

Our numerical examples in Section 7 and in the online appendix show that the new algorithm is able to to solve T_p-optimal discriminating design problems of higher dimensions in situations where the classical methods fail.

Acknowledgments. We are very grateful to the referees and the associate editor for their constructive comments on an earlier version of this manuscript. In particular, one referee encouraged us to include examples with a larger number of pairwise comparisons. By these investigations we gained more insight in the optimization problem, which led to a further improvement of the proposed algorithm. We also want to thank Stefan Skowronek for providing a code for the numerical calculations and Martina Stein, who typed parts of this manuscript with considerable technical expertise. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C2) of the German Research Foundation (DFG).

References.

Atkinson, A. C. (2008a). *DT*-optimum designs for model discrimination and parameter estimation. *Journal of Statistical Planning and Inference* **138** 56-64.

Atkinson, A. C. (2008b). Examples of the use of an equivalence theorem in constructing optimum experimental designs for random-effects nonlinear regression models. *Journal of Statistical Planning and Inference* **138** 2595-2606.

Atkinson, A. C. and Fedorov, V. V. (1975a). The designs of experiments for discriminating between two rival models. *Biometrika* **62** 57-70.

Atkinson, A. C. and Fedorov, V. V. (1975b). Optimal design: Experiments for discriminating between several models. *Biometrika* **62** 289-303.

Braess, D. (1986). *Nonlinear Approximation Theory.* Springer, Berlin.

Bretz, F., Pinheiro, J. C. and Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **61** 738-748.

Brosowski, B. (1968). *Nicht-lineare Tschebyscheff-Approximation.* Bibliographisches Institut.

Cheney, E. W. (1966). *Introduction to Approximation Theory.* MacGraw-Hill, New York.

Dette, H. and Haller, G. (1998). Optimal designs for the identification of the order of a Fourier regression. *Annals of Statistics* **26** 1496-1521.

Dette, H., Melas, V. B. and Shpilev, P. (2012). T-optimal designs for discrimination between two polynomial models. *The Annals of Statistics* **40** 188-205.

Dette, H. and Titoff, S. (2009). Optimal discrimination designs. *Annals of Statistics* **37** 2056-2082.

Fedorov, V. V. (1981). Design of model testing experiments. In *Symposia Mathematica* 171-180.

Fedorov, V. and Khabarov, V. (1986). Duality of optimal designs for model discrimination and parameter estimation. *Biometrika* **73** 183-190.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics* **2** 849-879.

López-Fidalgo, J., Tommasi, C. and Trandafir, P. C. (2007). An optimal experimental design criterion for discriminating between non-normal models. *Journal of the Royal Statistical Society, Series B* **69** 231-242.

Meinardus, G. (1967). *Approximation of Functions: Theory and Numerical Methods.* Springer, Berlin.

Papadimitriou, C. H. and Steiglitz, K. (1998). *Combinatorial Optimization: Algorithms and Complexity.* Dover (2nd edition).

Pinheiro, J., Bretz, F. and Branson, M. (2006). Analysis of dose-response studies: Modeling approaches. In *Dose Finding in Drug Development* (N. Ting, ed.) 146-171. Springer-Verlag, New York.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression.* John Wiley and Sons Inc., New York.

Song, D. and Wong, W. K. (1999). On the construction of G_{rm}-optimal designs. *Statistica Sinica* **9** 263-272.

Stigler, S. (1971). Optimal experimental design for polynomial regression. *Journal of the American Statistical Association* **66** 311-318.

Tommasi, C. (2009). Optimal designs for both model discrimination and parameter estimation. *Journal of Statistical Planning and Inference* **139** 4123-4132.

Tommasi, C. and López-Fidalgo, J. (2010). Bayesian optimum designs for discriminating between models with any distribution. *Computational Statistics & Data Analysis* **54** 143-150.

Ucinski, D. and Bogacka, B. (2005). T-optimum designs for discrimination between two multiresponse dynamic models. *Journal of the Royal Statistical Society, Ser. B* **67** 3-18.

Wiens, D. P. (2009). Robust discrimination designs, with Matlab code. *Journal of the Royal Statistical Society, Ser. B* **71** 805-829.

ONLINE APPENDIX: "OPTIMAL DISCRIMINATING DESIGNS FOR SEVERAL COMPETING REGRESSION MODELS"

By Dietrich Braess and Holger Dette

Ruhr-Universität Bochum

APPENDIX A: SOME REMARKS ON THE CONVERGENCE

A.1. Proof of convergence. In this section we briefly address the problem of convergence of the proposed algorithm. When T_p-optimal discriminating designs for linear models have to be determined, the approximation problem is a convex optimization problem, and the iteration converges to a best approximation. In the general case of nonlinear models, we only expect convergence to a local best approximation or to a *critical point* in a rigorous setting; cf. Braess (1986). We assume that the functions $\theta \to \eta(\cdot, \theta)$ are differentiable functions on open sets in $\mathbb{R}^n$.

Definition A.1 *A function $\eta(\cdot, \theta)$ is a* critical point *for the T_p-optimal discriminating design problem if $\eta(\cdot, \theta)$ is a best approximation to f in the gradient space defined in* (6.11).

The condition in turn may be expressed in terms of Gateau derivatives: $\eta(\cdot, \theta)$ is a critical point if

$$\langle f - \eta(\cdot, \theta), v \rangle = 0 \quad \forall\, v = \sum_{k=1}^{n} \alpha_k \frac{\partial}{\partial \theta_k} \eta(\cdot, \theta),\ \alpha \in \mathbb{R}^n.$$

Theorem A.2 *Let $u_0 = \eta(\cdot, \theta_0)$ be the initial guess for the iteration and $c_0 = \|f - u_0\|$. If the level set*

$$\{\eta(\cdot, \theta);\ \|f - \eta(\cdot, \theta)\| \le c_0\} \tag{A.1}$$

is compact, the iteration converges to a critical point of the T_p-optimal discriminating design problem.

Moreover, if additionally all competing regression models are linear, convergence to a best approximation holds.

Proof. For convenience, we restrict ourselves to linear models. The extension to the weaker results for nonlinear models is straight-forward but more technical.

Let $(u_s)_{s\in\mathbb{N}}$ be the sequence generated by the iteration and $u_{s,1}$ be the result of the first part of the iteration step $s+1$. The step length control guarantees that

$$\|f-u_{s+1}\| \le \|f-u_{s,1}\| \le \|f-u_s\|, \ s\in\mathbb{N}, \tag{A.2}$$

and the sequence $(u_{s,1})_{s\in\mathbb{N}}$ is contained in a compact subset of $V \subset \mathcal{F}_d$. Consequently a subsequence of $(u_{s,1})_{s\in\mathbb{N}}$ converges to a function $\hat{u} = \eta(\cdot,\hat{\theta})$.

Suppose that

$$\|f-\hat{u}\|^2 > \Phi := \inf_{v\in V} \|f-v\|^2. \tag{A.3}$$

Let $\mathcal{A} = \{x_1, x_2, \ldots, x_\mu\}$ be the set of extreme points of $\hat{\varepsilon} = f - \hat{u}$. We consider the correction determined by (6.9) if we proceed as in Section 6.1 for $u_1 = \hat{u}$:

$$\begin{array}{lll} A\alpha - R^T w & = & 0, \\ -R\alpha \qquad +\lambda e & = & -\frac{1}{2}b. \end{array}$$

We multiply the second line by w^T and make use of $w^T R\alpha = \alpha^T R w = \alpha^T A\alpha$ to obtain

$$-\alpha^T A\alpha + \lambda w^T e = -\frac{1}{2} w^T b.$$

Since A is positive definite and $w^T e = 1$, it follows that

$$-\lambda \le \frac{1}{2} w^T b - \alpha^T A\alpha \tag{A.4}$$

and

$$b - 2R\alpha \le -\lambda e \le (w^T b)e \le (\max_i b_i)e \tag{A.5}$$

with equality only if $\alpha = 0$ and $b_1 = b_2 = \ldots = b_\mu$.

If $\alpha = 0$, it follows that $R^T w = 0$, since A is positive definite, and $\hat{\theta}$ is the best parameter for the design ξ with masses $w_1, w_2, \ldots, w_\mu$. The relations $b_1 = b_2 = \ldots = b_\mu$ imply that $\|f-\hat{u}\|^2 = T(\xi) \le \Phi$, and we have a contradiction to (A.3).

We turn to the case that we have strict inequality in one of the relations in (A.5) and set $v = \sum_k \alpha_k v_k$. We rewrite (A.5) in terms of function values and obtain

$$|\hat{\varepsilon}(x)|^2 - 2\,\langle \hat{\varepsilon}(x), v(x)\rangle \le \|\hat{\varepsilon}\|^2 - \delta_1 \quad \text{for } x = x_1, x_2, \ldots, x_\mu,$$

with some $\delta_1 > 0$. Now we proceed as in the proof of the classical Kolmogorov theorem. Continuity implies that the inequality above applies to a

neighborhood U of $\{x_1, x_2, \ldots, x_\mu\}$ if we replace δ_1 by $\frac{1}{2}\delta_1$ and we have for $x \in U$:

$$\begin{aligned} |(f - \hat{u} - tv)(x)|^2 &\leq \|f - \hat{u}\|^2 - \tfrac{1}{2}t\delta_1 + t^2\|v\|^2 \\ &\leq \|f - \hat{u}\|^2 - \tfrac{1}{4}t\delta_1 \quad \text{for } t \leq \delta_1/(4\|v\|^2). \end{aligned}$$

The design space $\mathcal{X}$ is assumed to be compact, and $|\hat{\varepsilon}(x)|^2 \leq \|\hat{\varepsilon}(x)\|^2 - \delta_2$ holds in the compact set $\mathcal{X}\backslash U$ with some $\delta_2 > 0$. Hence,

$$\|f - \hat{u} - tv\|^2 \leq \|f - \hat{u}\|^2 - \delta_3$$

holds for some $t > 0$ and $\delta_3 > 0$. Continuity of the algorithm implies

$$\|f - u_{s+1}\|^2 \leq \|f - u_{s,1}\|^2 - \frac{1}{2}\delta_3$$

for infinitely many s in the convergent subsequence. This contradicts (A.3). □

Note that the compactness condition in the previous theorem is obviously satisfied by linear models. We also note that convergence proofs, which are based on monotonicity arguments, do not provide information on the speed of convergence.

A.2. Speed of convergence. The convergence proof in the previous section is based on a compactness argument. Therefore, it provides no information on the speed of convergence. We will present a heuristic argument why we expect the fast convergence (at least in the linear case) that we found in the numerical examples.

Let $u_1 = \eta(\cdot, \theta_1)$ be the actual approximating function and $\mathcal{S}_1 = \{x_1, x_2, \ldots, x_\mu\}$ be the actual reference set. Set $\varepsilon_1 = f - u_1$ and $e = (e_1, e_2, \ldots, e_\mu)^T = (1, 1, \ldots, 1)^T$, and let A, R and b be defined as in (6.4). The correction vector α for the parameters θ and the weights $w = (w_1, w_2, \ldots, w_\mu)^T$ are determined by these equations that are repeated here for the reader's convenience,

$$\begin{array}{rrrcl} A\alpha & -R^T w & & = & 0, \\ 2Rw & & +\lambda e & = & b, \\ & e^T w & & = & 1. \end{array} \tag{A.6}$$

Now we ignore that the matrix A is computed with the masses in the vector $\tilde{w}$ and not with the the masses in the vector w. The latter masses define a design ξ_2. The parameter $\theta_2 = \theta_1 + \alpha$ and $u_2 = \eta(\cdot, \theta_2) = u_1 + v$ are associated to $T(\xi_2)$ due to the hypothesis above, i.e.,

$$T(\xi_2) = \sum_{i=1}^{\mu} w_i |f(x_i) - u_2(x_i)|^2.$$

We recall that the components of the vector $b-2R\alpha$ represent the linearization of $|f-u_1-v|^2$, i.e.,

$$|(f-u_1-v)(x_i)|^2 = (b-2R\alpha)_i + |v(x_i)|^2. \tag{A.7}$$

The equations in the second line of (A.6) show that all components of $b-2R\alpha$ are equal, and

$$\begin{aligned}(b-2R\alpha)_i &= \sum_j w_j (b-2R\alpha)_j \le \sum_j w_j |(f-u_2)(x_j)|^2 \\ &= T(\xi_2) \le \min_{\theta\in\Theta} \|f-\eta(\cdot,\theta)\|^2.\end{aligned}$$

By using (A.7) once more, we have

$$|(f-u_2)(x_i)|^2 \le \min_{\theta\in\Theta} \|f-\eta(\cdot,\theta)\|^2 + |v(x_i)|^2. \tag{A.8}$$

If an iteration step yields large corrections, we expect a reduction due to the term $\alpha^T A\alpha$ in (A.4). If the correction is small, then it follows from (A.8) that we get close to the minimum.

APPENDIX B: DETAILS OF THE ALGORITHM

In this section we present one iteration step with its details. As usual, an instruction $A \leftarrow B$ in the algorithm means that the value of A is replaced by B. Short loops over finite sets of indices, however, are written as in mathematical formulas and not as computer code. Lines that start with a # contain comments.

There will be a loop over all pairs in $\mathcal{I}$ as announced in Section 6.1. The pairs will be specified by the symbol κ instead of (i,j) since the letter i is used for other purposes. The index sets $\mathcal{J}_\kappa$ denotes those components of the vector $\theta=(\theta_1,\theta_2,\dots,\theta_n)^T$ that are associated to the subset V_κ.

As it was described in Section 6.1 the iteration may start with a function for the parameters $\theta_{(i,j)} = \rho_{(j)}$ and a reference set of about $n+1$ points which divide the interval $\mathcal{X}$ into subdomains of equal size.

Algorithm B.1 *(Newton steps in Part 1 of the iteration step)*

Input
an approximation $u_0 = \eta(x,\theta_0)$, that need not be dual feasible,
a reference set $\mathcal{S}_0 = \{x_i\}_{i=1}^{\nu}$,
Unmark all points $x_i \in \mathcal{S}_0$, (all old marks are removed)
for $\kappa \in \mathcal{I}$
{

Primal linear program
Coefficients of the linear program

$$\left.\begin{array}{rl} \varepsilon_0(x_i) & \leftarrow f(x_i) - u_0(x_i), \\ r_{im} & \leftarrow \left\langle \varepsilon_0(x_i), \frac{\partial}{\partial \theta_m}\eta(x_i,\theta_0)\right\rangle,\ m \in \mathcal{J}_\kappa \\ b_i & \leftarrow |\varepsilon_0(x_i)|^2, \end{array}\right\} i = 1,2,\ldots,\nu,$$

solve the linear program (with respect to $E, \alpha_1, \ldots, \alpha_n$)

$$\begin{array}{rcll} E & \to & \text{min!} & \\ 2\sum\limits_{m\in\mathcal{J}_\kappa} r_{im}\alpha_m + E & \geq & b_i, & i = 1,2,\ldots,\nu, \\ E & \geq & 0. & \end{array}$$

$\alpha_m \leftarrow 0$ for all $m \notin \mathcal{J}_\kappa$,

Step length control

Determine the step length $t \in \{1, 2^{-1}, \ldots, 2^{-7}, 0\}$
for which $\|f - \eta(\cdot, \theta_0 + t\alpha)\|$ is minimal.
$\theta_0 \leftarrow \theta_0 + t\alpha$.

Mark candidates for points in the support
(in the case of dual feasible functions)

if $E > 0$
{
Solve the dual linear program

$$\begin{array}{rcl} \sum_{i=1}^{\nu} b_i w_i & \to & \text{max!} \\ \sum_{i=1}^{\nu} r_{im} w_i & = & 0,\ m \in \mathcal{J}_\kappa, \\ \sum_{i=1}^{\nu} w_i & = & 1,\ w_i \geq 0,\ i = 1,2,\ldots,\nu. \end{array}$$

mark the points x_i for which $w_i > 0$
}
else
{
print "subproblem κ is not dual feasible",
}
}
# end of κ-loop
$\theta_1 \leftarrow \theta_0$,
$u_1 \leftarrow \eta(\cdot, \theta_1)$,
Decision on the update of $\mathcal{S}_0$ and the support
show the function $|\varepsilon_1(\cdot)|$ on the monitory of the computer,
$\mu \leftarrow$ number of marked points,
relabel points of $\mathcal{S}_0$ such that $x_1, x_2, \ldots x_\mu$ are the marked points,
replace x_i by the local maximum of $|\varepsilon_1|$ next to x_i for $i = 1,2,\ldots,\mu$,

add local maxima of $|\varepsilon_1|$ to $\mathcal{S}_0$ that are not close to points in $\mathcal{S}_0$,

\# Output

$\mathcal{S}_1 \leftarrow \mathcal{S}_0$,

supp $\leftarrow \{x_1, x_2, \ldots, x_\mu\}$,

θ_1. □

As a consequence of Conjecture 4.5 the reference set should contain at least $\max_{(i,j)\in\mathcal{I}} \dim \Theta^{(j)} + 1$ points, but the number may be larger. There are usually large changes of the error function in the first iteration steps, and it is reasonable to have surplus points in the reference set. If the algorithm asks to replace a point in the reference set by a new one, one may add the new one and keep the old one. The computing times are so small that extra points have no measurable effect on the computing time, but they may be useful in controlling large changes. Later, when there are only small changes, it is reasonable to delete points that are far away from local maxima. The decision is left to the user and does not cause a problem, since the user may decide to keep the points in uncertain cases.

Algorithm B.2 *(Linearized saddle point problems in Part 2 of the iteration step)*

\# Input

an approximation $u_1 = \eta(\cdot, \theta_1)$,

a reference set $\mathcal{S}_1 = \{x_i\}_{i=1}^{\nu}$ which is the approximation of the support

a subset $\{x_i\}_{i=1}^{\mu} \subset \mathcal{S}_1$,

\# First approximation of ξ

\# Coefficients of the linear program

$$\left.\begin{array}{rll} \bar{w}_i & \leftarrow 1/\mu, & \\ \varepsilon_1(x_i) & \leftarrow f(x_i) - u_1(x_i), & \\ b_i & \leftarrow |\varepsilon_1(x_i)|^2, & \\ \partial u_{im} & \leftarrow \frac{\partial}{\partial \theta_m}\eta(x_i, \theta_1), & m = 1, 2, \ldots, n \\ r_{im} & \leftarrow \langle \varepsilon_0(x_i), \partial u_{im} \rangle\,, & m = 1, 2, \ldots, n, \end{array}\right\} i = 1, 2, \ldots, \mu,$$

$a_{km} \leftarrow \sum_{i=1}^{\mu} \bar{w}_i \langle \partial u_{ik}, \partial u_{im} \rangle$ for $k, m = 1, 2, \ldots n$,

\# Here and below: $R := (r_{im})$, $A := (a_{km})$, $Q := (q_{ij})$, ∂u is one symbol

$Q \leftarrow RA^{-1}R^T$,

solve the linear program (with respect to $\delta, \tilde{w}_1, \ldots, \tilde{w}_\mu$)

$$\begin{array}{rcll} \delta & \rightarrow & \min! & \\ \sum_{j=1}^{\mu} q_{ij}\tilde{w}_j - \delta & \leq & 0, & i = 1, 2, \ldots, \mu, \\ \sum_{j=1}^{\mu} q_{ij}\tilde{w}_j + \delta & \leq & 0, & i = 1, 2, \ldots, \mu, \\ \sum_{j=1}^{\mu} \tilde{w}_j & = & 1, & \tilde{w}_i \geq 0,\ i = 1, 2, \ldots, \mu, \end{array}$$

eliminate points x_i with $\tilde{w}_i = 0$, if necessary,
We assume that $\tilde{w}_i > 0$ holds for $1 \le i \le \mu$
and drop the description of the elimination procedure
with the associated relabeling
Linearized saddle point problem
$a_{km} \leftarrow \sum_{i=1}^{\mu} \tilde{w}_i \langle \partial u_{ik}, \partial u_{im} \rangle$ for $k, m = 1, 2, \dots n$,
$e \leftarrow (1, 1, \dots, 1)^T$,
solve the system of linear equations

$$\begin{pmatrix} A & -R^T & \\ -R & & e \\ & e^T & \end{pmatrix} \begin{pmatrix} \alpha \\ w \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} b \\ 1 \end{pmatrix}.$$

ξ is given by $\{x_i\}_1^\mu$ and $\{w_i\}_1^\mu$
The minimization for determining $T(\xi)$ is done by 3 Newton steps
$\theta_2 \leftarrow \theta_1$,
for $\sigma = 1, 2, 3$
{

$$\left.\begin{aligned} u_2 &\leftarrow \eta(\cdot, \theta_2), \\ \varepsilon_2(x_i) &\leftarrow f(x_i) - u_2(x_i), \\ b_i &\leftarrow |\varepsilon_2(x_i)|^2, \\ \partial u_{im} &\leftarrow \tfrac{\partial}{\partial \theta_m} \eta(x_i, \theta_2), \qquad m = 1, 2, \dots, n, \\ r_{im} &\leftarrow \langle \varepsilon_2(x_i), \partial u_{im} \rangle, \quad m = 1, 2, \dots, n, \end{aligned}\right\} i = 1, 2, \dots, \mu,$$

matrix A and vector c for linear equation
$a_{km} \leftarrow \sum_{i=1}^{\mu} \tilde{w}_i \langle \partial u_{ik}, \partial u_{im} \rangle$ for $k, m = 1, 2, \dots n$,
$c \leftarrow R^T w$,
solve the linear equation $A\gamma = d$,
Step length control
determine $t \in \{1, 2^{-1}, \dots, 2^{-7}, 0\}$
for which $\sum_{i=1}^{\mu} w_i |f(x_i) - \eta(x_i, \theta_0 + t\gamma)|^2$ is minimal.
$\theta_2 \leftarrow \theta_0 + t\gamma$,
}
# end of Newton loop
$T(\xi) \leftarrow \sum_{i=1}^{\mu} w_i |f(x_i) - \eta(x_i, \theta_2)|^2$,
Improve the approximating function by recalling the saddle point equation
determine $t \in \{1, 2^{-1}, \dots, 2^{-7}, 0\}$ for which $\|f - \eta(\cdot, \theta_1 + t\alpha)\|$ is minimal,
Output
$\theta_2 \leftarrow \theta_1 + t\alpha$,
$u_2 \leftarrow \eta(\cdot, \theta_2)$,
efficiency $\leftarrow T(\xi)/\|f - u_2\|^2$. □

Table 6
The results of the new algorithm for Example C.1.

j	$\|\varepsilon_{j,1}\|^2$	$\|\varepsilon_{j,2}\|^2$	$T(\xi_j)$	$\frac{T(\xi_j)}{\|\varepsilon_{j,2}\|^2}$	support	reference set
	Part 1		Part 2			
0	14082					$\mathcal{S} \leftarrow \{0, 30, 60, 90, \ldots, 450, 500\}$
1	9018					
2	8666	7408	1368	0.1818	$\{0, 50, 250, 500\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{50, 250\}$
3	5668	4518	2058	0.3798	$\{0, 80, 250, 500\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{80\}$
4	4722	4582	3267	0.7131		
5	4513	4399	3433	0.7804		
6	4226	4123	3556	0.8625		
7	4026	3894	3591	0.9223	$\{0, 70, 250, 500\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{70\}$
8	3782	3715	3618	0.9739	$\{0, 70, 240, 500\}$	
9	3681	3648	3618	0.9918		
10	3622	3622	3621	0.9998	$\{0, 75, 225, 500\}$	$\mathcal{S} \leftarrow \mathcal{S} \cup \{75, 235\}$

APPENDIX C: MORE NUMERICAL RESULTS

Example C.1 In this example we provide a further illustration of the performance of the new algorithm if $d \geq 3$ pairwise comparisons are addressed in the T_p-optimal discriminating design criterion. The competing regression models including the specified parameters are given in Table 1 in the introduction. Here the weights are given by $p_{2,1} = p_{3,1} = p_{3,2} = p_{4,1} = p_{4,2} = 1/5$ and 0 otherwise. The size of the problem is characterized by $d = 5$ and $n = 12$ for this choice.

We start with the true parameters and the reference set

$$\mathcal{S} = \{0, 30, 60, \ldots, 450, 500\},$$

and the results of the iteration are shown in Table 6. Although we have 12 free parameters in the nonlinear approximation problem, the T_p-optimal discriminating design has only 4 support points. The T_p-optimal discriminating design calculated in 10 iteration steps is given by

$$\xi^* = \begin{pmatrix} 0 & 75 & 235 & 500 \\ 0.26 & 0.18 & 0.38 & 0.18 \end{pmatrix} \tag{C.1}$$

and has an efficiency of at least 99%.

We also tried to calculate the T_p-optimal discriminating design by an application of the procedure proposed by Atkinson and Fedorov (1975a), but this algorithm failed in the calculation of the T_p-optimal discriminating

design. A heuristical explanation of this observation is that the objective function is extremely flat. For example, the design

$$\xi = \begin{pmatrix} 0 & 70 & 240 & 500 \\ 0.33 & 0.28 & 0.28 & 0.11 \end{pmatrix},$$

which is not close to ξ^*, still has an efficiency of more than 93 %. If a support point is split into several ones, then the objective function becomes even more flat.

Fakultät für Mathematik
44780 Bochum, Germany
E-mail: dietrich.braess@rub.de
holger.dette@rub.de