Optimal discriminating designs for several competing regression models

Dietrich Braess
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
e-mail: dietrich.braess@rub.de

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum, Germany
e-mail: holger.dette@rub.de

November 23, 2011

Abstract

The problem of constructing optimal 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) and demonstrate that optimal designs with respect to this type of criteria can be obtained by solving (nonlinear) vector-valued approximation problems. We provide a characterization of the best approximations in this context and use these results to develop an efficient algorithm for the determination of the optimal discriminating designs. The results are illustrated by fnumerical examples.

Keyword and Phrases: Optimal design; model discrimination; vector-valued approximation

AMS Subject Classification: 62K05, 41A30, 41A50

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

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

Often there exist several plausible models which may be appropriate for a fit to the data. A typical example are dose-finding studies, where various models have been developed for describing the dose-response relation [see Dette et al. (2008)]. In these and similar cases the first step of the data analysis consists in the identification of an appropriate model from a class of competing regression models. The optimal design problem for a situation of this type has a long history. Early work on discriminating designs is found in Stigler (1971) and Studden (1982), who determined designs for discriminating between two nested univariate polynomials by minimizing the volume of the confidence ellipsoid for the parameters corresponding to the extension of the smaller model. Several authors have continued to work on this approach in various other models [Spruill (1990), Dette (1994), Dette and Haller (1998), Song and Wong (1999), Zen and Tsai (2002) or Zen and Tsai (2004) among others].

In a pioneering paper Atkinson and Fedorov (1975a) proposed an alternative criterion, called T-optimality criterion, which provides a design such that the sum of squares for a lack of fit test is large. This optimality criterion has found considerable attention in the statistical literature [see e.g. Fedorov (1981), Denisov et al. (1981), Fedorov and Khabarov (1986) for early and Ucinski and Bogacka (2005), López-Fidalgo et al. (2007), Atkinson (2008b), Atkinson (2008a), Tommasi (2009) or Wiens (2009) for some more recent references]. Atkinson and Fedorov (1975b) extended their approach later to designs for discriminating between a class of given regression models, say $\mathcal{M} = \{\eta_1, \eta_2, \ldots, \eta_k\}$, $k \geq 2$. In general, the problem of finding T-optimal designs, either analytically or numerically, is a very hard and challenging one. Recently Dette and Titoff (2009) considered the case $k = 2$. They explored the relation between the two concepts of discriminating designs by relating the T-optimal design problem to a problem in approximation theory, which indicate the difficulties.

In the present paper we construct optimal discriminating designs for $k \geq 3$ competing regression models where none of the models is selected in advance to be tested against all other models. We consider a weighted T-optimality criterion which is a slight modification of the criterion introduced by Atkinson and Fedorov (1975b). It is demonstrated that these design problems are closely related to vector-valued approximation problems. In particular, we show that 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. Vector-valued approximation theory has not been studied intensively in the literature, and we are only aware of the investigations of Brosowski (1968) who considered some special nonlinear families. Therefore we study this approximation problem in Section 4 and provide a characterization of the best vector-valued approximation that generalizes the classical Kolmogorov criterion [Kolmogorov (1948) or Meinardus (1967)]. In Sections 5 and 6 we use these results to develop an efficient algorithm for the calculation of best approximations which provide the T-optimal discriminating designs. Finally, in Section 7 we illustrate our approach by several numerical examples. The algorithm solves the corresponding approximation problem in less than 20 iterations and determines simultaneously the optimal design. Details of the main step of the algorithm are given in Section 8.

2 Preliminaries

Following Kiefer (1974) we consider approximate designs that are defined as probability measures with finite support on a compact design space $\mathcal{X}$. The support points of an (approximate) design ξ give the locations where observations are to be taken, while the weights give the corresponding relative proportions of observations at these points. Let the design ξ have positive masses $w_1, \ldots, w_\nu$ at the distinct points $x_1, \ldots, x_\nu$, respectively, and assume that N observations can be made by the experimenter. In this case the quantities $w_i N$ are rounded to integers, say N_i, satisfying $\sum_{i=1}^{\nu} N_i = N$. The experimenter takes N_i observations at the location x_i ($i = 1, \ldots, \nu$). 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, which varies in the set $\Theta^{(j)}$ ($j = 1, \ldots, k$). Atkinson and Fedorov (1975b) proposed to fix one model in $\mathcal{M}$, say η_1 with vector of parameters $\rho_{(1)}$, and to determine a discriminating design by maximizing

$$\text{(2.1)} \qquad \min_{j=2}^{k} \Delta_{1,j}(\xi) \quad \text{where} \quad \Delta_{1,j}(\xi) = \inf_{\theta_{(j)} \in \Theta_{(j)}} \int_{\mathcal{X}} [\eta_1(x, \rho_{(1)}) - \eta_j(x, \theta_{(j)})]^2 d\xi(x).$$

If the competing regression models $\eta_1, \ldots, \eta_k$ are not nested, it is not clear which model is to be fixed in this approach, and it is useful to have more "symmetry" in the concept. For illustration consider the case of two competing models, say $\eta_i(x, \theta_{(1)}), \eta_j(x, \theta_{(2)})$, 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 $\Delta_{1,2}$ and $\Delta_{2,1}$, corresponding to the specification of the model η_1 or η_2, respectively, where

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

($i \neq j$). 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 corresponding to the minimum is denoted as

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

We assume its existence, and the dependence of this parameter on the value $\rho_{(i)}$ is not reflected in our notation. 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 $\sum_{i \neq j} p_{i,j} = 1$. Following Atkinson and Fedorov (1975b) a design ξ^* is called T_p-*optimal discriminating* for the class of models $\mathcal{M} = \{\eta_1, \ldots, \eta_k\}$ if it maximizes the functional

$$\text{(2.4)} \qquad \Delta(\xi) = \sum_{1 \leq i \neq j \leq k} p_{i,j} \Delta_{i,j}(\xi).$$

Throughout this paper, we will denote the parameter $\theta^*_{(i,j)}(\xi^*)$ defined in (2.3) by $\theta^*_{(i,j)}$ whenever there is no danger of confusion. For the choice

$$p_{i,j} > 0 \quad (j = 2, \dots, k), \qquad p_{i,j} = 0 \quad (i = 2, \dots, k,\ j = 2, \dots, k;\ i \neq j), \tag{2.5}$$

the criterion (2.4) reduces to a similar optimality criterion as considered by Atkinson and Fedorov (1975b) in the special case (2.1). The criterion (2.4) provides a 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.

Remark 2.1 In the case $k = 2$ Tommasi and López-Fidalgo (2010) proposed to maximize a weighted mean of efficiencies that yields in the situation considered here the criterion

$$\sum_{1 \le i \neq j \le k} \widetilde{p}_{i,j} \frac{\Delta_{i,j}(\xi)}{\Delta_{i,j}(\xi^*_{i,j})}, \tag{2.6}$$

where $\xi^*_{i,j}$ denotes the design maximizing the criterion $\Delta_{i,j}$ defined in (2.2). Both criteria are equivalent if the weights are chosen as

$$p_{i,j} = \frac{\widetilde{p}_{i,j}}{\Delta_{i,j}(\xi^*_{i,j})} \bar{p}^{-1}$$

and $\tilde{p} = \sum_{1 \le i \neq j \le k} \widetilde{p}_{i,j}$. For sake of a simple notation we consider the criterion (2.4) throughout this paper, but in applications standardization should be taken into account [see Dette (1997)].

We will relate the optimal discriminating design problem to a nonlinear vector-valued approximation problem. To be precise, let the weights $p_{i,j}$ for the criterion (2.4) be given, denote the set of indices corresponding to the positive weights

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

We assume without loss of generality that $\mathcal{I}$ can be decomposed in $p \le k$ subsets $\mathcal{I}_1, \dots, \mathcal{I}_p$ of the form $\mathcal{I}_i := \{(i,j) \in \mathcal{I} \mid 1 \le j \le k\}$. This means that for each model η_i, $(i = 1, 2, \dots, p)$, a parameter is fixed and it is to be discriminated from the other ones in the set $\mathcal{I}_i$. Define

$$\lambda_i = \#\mathcal{I}_i, \quad d = \sum_{i=1}^{p} \lambda_i \tag{2.7}$$

as the cardinality of $\mathcal{I}_i$ and $\mathcal{I}$, respectively, and consider the space of vector-valued functions defined on $\mathcal{X}$, i.e., $\mathcal{F}_d = \{g : \mathcal{X} \to \mathbb{R}^d\}$. Given a function

$$g = (g_{ij})_{(ij) \in \mathcal{I}} \in \mathcal{F}_d\,,$$

we define a maximum norm by

$$\|g\| = \sup_{x \in \mathcal{X}} |g(x)| \tag{2.8}$$

where $|g(x)|^2 = \sum_{i=1}^{p} \sum_{j \in \mathcal{I}_i} p_{i,j}\, g_{ij}^2(x)$ denotes a *weighted Euclidean norm* on $\mathbb{R}^d$. In this framework, given two functions $f, g \in \mathcal{F}_d$, their distance is $\|f - g\|$. Next, define the d-dimensional vector-valued function

$$f(x) = \Big(\underbrace{\eta_1(x, \rho_{(1)}), \ldots, \eta_1(x, \rho_{(1)})}_{\lambda_1}, \ldots, \underbrace{\eta_p(x, \rho_{(p)}), \ldots, \eta_p(x, \rho_{(p)})}_{\lambda_p} \Big)^T, \tag{2.9}$$

where each function $\eta_j(x, \rho_{(j)})$ appears λ_j times in the vector $\eta(x)$, and consider the vector-valued approximating functions

$$\eta(x, \theta) = \Big(\underbrace{(\eta_j(x, \theta_{(1,j)}))_{j \in \mathcal{I}_1}}_{\lambda_1}, \ldots, \underbrace{(\eta_j(x, \theta_{(p,j)}))_{j \in \mathcal{I}_p}}_{\lambda_p} \Big) \in \mathcal{F}_d\,. \tag{2.10}$$

The corresponding parameters are collected in the vector

$$\theta = \Big((\theta_{(1,j)})_{j \in \mathcal{I}_1}, \ldots, (\theta_{(p,j)})_{j \in \mathcal{I}_p} \Big) \in \Theta = \otimes_{i=1}^{p} \otimes_{j \in \mathcal{I}_i} \Theta^{(j)}. \tag{2.11}$$

Hence,

$$\dim \Theta = n := \sum_{i=1}^{p} \sum_{j \in \mathcal{I}_i} \dim \Theta^{(j)}.$$

The following examples illustrate this general notation.

Example 2.2 Consider the case $k = 3$ and assume that all weights in the criterion (2.4) are positive. Here no model is preferred as basis model. In this case we address for 6 possible pairwise comparisons, and we have $p = k = 3$,

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

which gives $\lambda_1 = \lambda_2 = \lambda_3 = 2,\ d = 6$. We obtain for the vectors in (2.9) and (2.10)

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

with

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

Example 2.3 Consider the discrimination between $k = 3$ nested polynomial models, e.g.,

$$\begin{aligned} \eta_1(x, \theta_{(1)}) &= \theta_{10} + \theta_{11}x, \\ \eta_2(x, \theta_{(2)}) &= \theta_{20} + \theta_{21}x + \theta_{22}x^2, \\ \eta_3(x, \theta_{(3)}) &= \theta_{30} + \theta_{31}x + \theta_{32}x^2 + \theta_{33}x^3. \end{aligned} \tag{2.12}$$

It is appropriate to choose only the weights $p_{2,1}$ and $p_{3,2}$ in the criterion (2.4) as positive numbers in order to obtain a design for identifying the degree of the polynomial, which yields

$$\mathcal{I} = \{(2,1), (3,2)\}, \quad \mathcal{I}_1 = \{(2,1)\}, \quad \mathcal{I}_2 = \{(3,2)\}.$$

Thus we have $p = 2$, $\lambda_1 = 1$, $\lambda_2 = 1$, $d = 2$. The functions f and $\eta(\cdot, \theta)$ are given by

$$\begin{aligned} f(x) &= \Big(\eta_2(x, \rho_{(2)}), \eta_3(x, \rho_{(3)})\Big)^T, \\ \eta(x, \theta) &= \Big(\eta_1(x, \theta_{(2,1)}), \eta_2(x, \theta_{(3,2)})\Big)^T = (\theta_{10} + \theta_{11}x, \theta_{20} + \theta_{21}x + \theta_{22}x^2)^T, \end{aligned}$$

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

3 Characterization of optimal designs

The T_p-optimality of a given design ξ can be checked by an equivalence theorem that can be proved by the same arguments as used by Atkinson and Fedorov (1975b) in the situation (2.1). As usual, the following properties are silently assumed to hold (note that the assumptions are always satisfied by linear models).

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

A2. Let ξ^* be a T_p-optimal discriminating design. The parameter

$$\theta^* = (\theta^*_{(i,j)})_{(i,j)\in\mathcal{I}}$$

defined by (2.3) exists, is unique and an interior point of Θ.

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 $x \in \mathcal{X}$*

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

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

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 (2.8).

Theorem 3.2 *For the criterion* (2.4)*, we have*

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

Moreover, if ξ^* *maximizes the criterion* (2.4)*, then we have for the vector* $\theta^* = (\theta^*_{(i,j)})_{(i,j)\in\mathcal{I}}$ *defined in* (2.3)

$$\|\eta(x) - \eta(x,\theta^*)\| = \inf_{\theta\in\Theta} \|\eta(x) - \eta(x,\theta)\| = \Delta(\xi^*). \tag{3.3}$$

Remark 3.3 Condition (3.3) means that the parameter θ^* defined in (2.3) corresponds to the best approximation of the function η in (2.9) by functions of the form (2.10) with respect to the norm (2.8). Moreover, the support of the T_p-optimal discriminating design ξ^* for the class $\mathcal{M}$ satisfies

$$\text{supp}(\xi^*) \subset \Big\{ x \in \mathcal{X} \;\Big|\; |\eta(x) - \eta(x,\theta^*)| = \|\eta - \eta(\cdot,\theta^*)\| \Big\} =: \mathcal{A}. \tag{3.4}$$

Proof of Theorem 3.2 and Remark 3.3. Let $\tilde\theta \in \Theta$. We obtain from (2.2), (2.4), and $\int_{\mathcal{X}} d\xi = 1$:

$$\begin{aligned}
\Delta(\xi) &= \sum_{(i,j)\in\mathcal{I}} p_{i,j} \inf_{\theta_{(i,j)}\in\Theta^{(j)}} \int_{\mathcal{X}} [\eta_i(x,\rho_{(i)}) - \eta_j(x,\theta_{(i,j)})]^2 d\xi(x) \\
&\le \sum_{(i,j)\in\mathcal{I}} p_{i,j} \int_{\mathcal{X}} [\eta_i(x,\rho_{(i)}) - \eta_j(x,\tilde\theta_{(i,j)})]^2 d\xi(x) \\
&= \int_{\mathcal{X}} \big|\eta(x) - \eta(x,\tilde\theta)\big|^2 d\xi(x) \le \|\eta - \eta(\cdot,\tilde\theta)\|^2.
\end{aligned}$$

Since $\tilde\theta$ is an arbitrary parameter in Θ, if follows that $\Delta(\xi) \le \inf_{\theta\in\Theta} \|\eta - \eta(\cdot,\theta)\|^2$, and

$$\sup_\xi \Delta(\xi) \le \inf_{\theta\in\Theta} \|\eta - \eta(\cdot,\theta)\|^2. \tag{3.5}$$

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 yields for the T_p-optimal discriminating design

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

which proves Theorem 3.2. The statement on the support points of ξ^* in Remark 3.3 follows directly from these considerations. □

Example 3.4 Consider the situation in Example 2.2, where we investigated 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 design problem can be found by solving the nonlinear vector-valued approximation problem

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

Example 3.5 Consider the situation in Example 2.3 where we are interested in the problem of discriminating between linear and quadratic and between a quadratic and cubic model. In this case we have $p_{2,1} > 0$ and $p_{3,2} > 0$, and the corresponding approximation problem is given by

$$\begin{aligned}(3.6)\qquad \inf\Big\{\sup_{x\in\mathcal{X}}\Big(&p_{2,1}|\rho_{20} + \rho_{21}x + \rho_{22}x^2 - \theta_{10} - \theta_{11}x|^2 \\ &+ p_{3,2}|\rho_{30} + \rho_{31}x + \rho_{32}x^2 + \rho_{33}x^3 - \theta_{20} - \theta_{21}x - \theta_{22}x^2|^2\Big)\Big|\theta_{10},\theta_{11},\theta_{20},\theta_{21},\theta_{22} \in \mathbb{R}\Big\},\end{aligned}$$

where $\rho_{(2)} = (\rho_{20}, \rho_{21}, \rho_{22})$ and $\rho_{(3)} = (\rho_{30}, \rho_{31}, \rho_{32}, \rho_{33})$ denote the fixed parameters for the models η_2 and η_3, respectively.

Now we turn to the situation that the nonlinear approximation problem has been solved and that we know the parameter $\bar\theta = \Big((\bar\theta_{(i,j)})_{j\in\mathcal{I}_1}, \ldots, (\bar\theta_{(p,j)})_{j\in\mathcal{I}_p}\Big)$ corresponding to the best approximation, i.e.,

$$(3.7)\qquad \|\eta - \eta(\cdot,\bar\theta)\|^2 = \min_{\theta\in\Theta} \|\eta - \eta(\cdot,\theta)\|^2.$$

By Theorem 3.2 and Remark 3.3 the support of the T_p-optimal discriminating design is contained in the set $\mathcal{A}$ defined in (3.4). The associated design (more precisely the weights at the support points) has still to be determined.

Corollary 3.6 *Assume that a parameter $\bar\theta$ defined in* (3.7) *exists and is an interior point of* Θ. *Moreover assume that the n derivatives*

$$\nabla_{\theta_{(i,j)}}\eta_j(x,\theta_{(i,j)}),\quad (i,j)\in\mathcal{I},$$

span an n-dimensional subspace of $\mathcal{F}_d$.
(a) If a design ξ is a T_p-optimal discriminating design for the class $\mathcal{M}$, then

$$(3.8)\qquad \int_{\mathcal{A}} \Big(\eta_i(x,\rho_{(i)}) - \eta_j(x,\bar\theta_{(i,j)})\Big)\nabla_{\theta_{(i,j)}}\eta_j(x,\theta_{(i,j)})\Big|_{\theta_{(i,j)}=\bar\theta_{(i,j)}} d\xi(x) = 0$$

holds for all $(i,j)\in\mathcal{I}$, where $\nabla_{\theta_{(i,j)}}$ denotes the gradient with respect $\theta_{(i,j)}$.

(b) Conversely, if ξ satisfies (3.8), $\mathrm{supp}(\xi) \subset \mathcal{A}$*, and the function*

$$\theta \longrightarrow \int_{\mathcal{A}} \sum_{i,j \in \mathcal{I}} p_{i,j} [\eta_i(x, \rho_{(i)}) - \eta_j(x, \theta_{(i,j)})]^2 d\xi(x) \tag{3.9}$$

has a unique minimum, then ξ is a T_p-optimal discriminating design for the class $\mathcal{M}$. Futhermore, the uniqueness assumptions can be dropped if all the competing models are linear.

Sketch of a proof. If condition (3.8) is not satisfied, there is a direction such that the expression on the right-hand side of (3.9) decreases. Thus (3.8) 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 here also sufficient. □

Remark 3.7 If there are at least $n+1$ extreme points in the set $\mathcal{A}$, an optimal design can be calculated by the n equations (3.8) together with the normalization $\int_{\mathcal{A}} d\xi(x) = 1$. This theoretical argument, however, is not applicable for the numerical solution of optimal designs for real-life problems. In most cases the number of extreme points is smaller than $n+1$, which complicates the determination of the set $\mathcal{A}$ and the corresponding design substantially.

The numerical examples in Section 7 and the consideration of Example 4.5 will show that practical problems lead to T_p-optimal designs with less than $n+1$ points in most cases. This is true even if the competing models are linear. We need more information on the approximation problem in order to deal with the mentioned degeneracy when T_p-optimal designs are determined by numerical methods.

4 Chebyshev Approximation of d-Variate Functions

By Theorem 3.2 a T_p-optimal discriminating design can be determined by solving an approximation problem in the space of d-variate functions on the compact design space $\mathcal{X}$. In this section we will investigate these problems in more detail for the case of linear models in order to be prepared for the computation and evaluation of the efficiency of (nearly) best approximations. We will see that there always exists a T_p-optimal discriminating design with at most $n+1$ support points and demonstrate that the number of support points is often less than $n+1$. This degeneracy confirms the nonlinearity of the approximation problem. Numerical algorithms for the computation of best approximations cannot proceed like the Remez algorithm, which is the common tool to solve the approximation problem in the case $d=1$ [see Cheney (1966)].

We will avoid double indices throughout this section because the main purpose here is to gain more insight in the approximation problem corresponding to the optimal design problem. To be precise, assume that the function $f = (f_1, \dots, f_d)$ defined by (2.9) is an element of the set

$$\mathcal{F}_d = C(\mathcal{X})^d$$

of d-variate continuous functions on a compact set $\mathcal{X}$.

In the case of linear models, equation (2.10) defines an n-dimensional linear subspace

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

where $v_1, v_2, \dots, v_n \in \mathcal{F}_d$ denotes a basis of 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.,

$$\|f - u^*\| = \min_{v \in V} \|f - v\|. \tag{4.2}$$

As stated in (2.8), the norm $\|\cdot\|$ refers to the maximum-norm on $\mathcal{X}$, $\|g\| := \sup_{x\in\mathcal{X}} |g(x)|$, where the weighted Euclidean norm $|\cdot|$ and the corresponding inner product in $\mathbb{R}^d$

$$|r|^2 := \sum_{l=1}^{d} p_l\, |r_l|^2, \qquad \langle \tilde{r}, r\rangle := \sum_{l=1}^{d} p_l\, \tilde{r}_l r_l, \quad r, \tilde{r} \in \mathbb{R}^d. \tag{4.3}$$

are now written with single indices. Specifically, the function values $f(x)$ and $v(x)$ for $v \in V$ are d-dimensional vectors. The family V defined in (4.1) is a linear space, and the classical Kolmogorov criterion [see Meinardus (1967)] can be generalized for d-variate functions. This generalization will be denoted as Kolmogorov criterion again. Note that the nonlinear character of the procedures for determining best approximations does not matter here.

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

$$\mathcal{A} := \{x \in \mathcal{X} \mid |\varepsilon(x)| = \|\varepsilon\|\}. \tag{4.4}$$

be the set of extreme points of the error function

$$\varepsilon := f - u\,. \tag{4.5}$$

The d-variate function u is a best approximation to f in V if and only if

$$\min_{x\in\mathcal{A}} \langle \varepsilon(x), v(x)\rangle \le 0 \quad \textit{for all } v \in V. \tag{4.6}$$

Proof. Let $u \in V$ and assume that (4.6) holds. Given $v \in V$, there exists a point $x_0 \in \mathcal{A}$ such that $\langle \varepsilon(x_0), v(x_0) \rangle \le 0$. Hence,

$$|(f-u-v)(x_0)|^2 = |\varepsilon(x_0)|^2 - 2\,\langle \varepsilon(x_0), v(x_0) \rangle + |v(x_0)|^2 \ge |\varepsilon(x_0)|^2,$$

and u is a best approximation.

In order to prove the converse, assume that u is a best approximation and that there exists $v_0 \in V$ such that

$$\langle \varepsilon(x), v_0(x) \rangle > 0 \quad \forall\, x \in \mathcal{A}. \tag{4.7}$$

Since the set $\mathcal{A}$ is compact, we have $\delta := 2\inf_{x\in\mathcal{A}} \langle \varepsilon(x), v_0(x) \rangle > 0$, which yields $\langle \varepsilon(x), v_0(x) \rangle > \delta$ for all x in some open neighborhood $\tilde{\mathcal{A}}$ of $\mathcal{A}$. If $t \le \delta/\|v\|^2$, it follows from this relation that

$$\begin{aligned} |f(x) - u(x) - tv_0(x)|^2 &= |f(x) - u(x)|^2 - 2t\,\langle \varepsilon(x), v_0(x) \rangle + t^2|v_0(x)|^2 \\ &\le |f(x) - u(x)|^2 - 2t\delta + t\delta \le \|f-u\|^2 - t\delta, \qquad x \in \tilde{\mathcal{A}}. \end{aligned}$$

If t is sufficiently small, the error is smaller on the compact set $\mathcal{X}\backslash\tilde{\mathcal{A}}$ as well, and u is not a best approximation. Therefore, the proof is complete. □

Now we turn to the consideration of inconsistent inequalities and specifically to the case that the system (4.7) is not solvable. Let $v_1, v_2, \ldots, v_n$ be a basis of V. Assume that u is a best approximation of the function f. Lemma 3.1 and the representation

$$v(x) = \sum_{m=1}^{n} \alpha_m v_m(x) \tag{4.8}$$

lead to the unsolvable system for $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T \in \mathbb{R}^n$:

$$\sum_{m=1}^{n} \alpha_m r_m(x) > 0 \quad \forall\, x \in \mathcal{A}\,, \tag{4.9}$$

where we use the notation $r_m(x) := \langle \varepsilon(x), v_m(x) \rangle$. These numbers are considered as the components of a vector $r(x)$, and we make use of the following lemma. An elementary proof is provided on p. 19 in the book by Cheney (1966).

Lemma 4.2 *(Lemma on linear inequalities)*
Let $M \subset \mathbb{R}^n$ be a compact set. Then the following statements are equivalent:
(i) The system of inequalities

$$\langle r, y \rangle > 0 \quad \textit{for all} \;\; r \in M$$

has no solution $y \in \mathbb{R}^n$.
(ii) The convex hull of M contains the origin. □

It follows from Lemma 4.2 that (4.9) is not solvable if the origin in $\mathbb{R}^n$ is an element of the convex hull of the vectors

$$\{r(x) = (r_1(x), \dots, r_n(x))^T,\ x \in \mathcal{A}\}.$$

By Carathéodory's theorem there are $\nu \le n+1$ points $x_1, \dots, x_\nu \in \mathcal{A}$ and numbers $w_1, \dots, w_\nu \ge 0$ such that $\sum_{i=1}^{\nu} w_i = 1$ and

$$\sum_{i=1}^{\nu} w_i r(x_i) = \sum_{i=1}^{\nu} w_i \langle \varepsilon(x_i), v(x_i) \rangle = 0 \quad \forall\, v \in V. \tag{4.10}$$

Theorem 4.3 *(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 $\nu \le n+1$ points $x_1, x_2, \dots, x_\nu \in \mathcal{A}$ such that

$$\min_{1 \le i \le \nu} \langle \varepsilon(x_i), v(x_i) \rangle \le 0 \quad \forall v \in V. \tag{4.11}$$

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

$$\ell(g) := \frac{1}{\|\varepsilon\|} \sum_{i=1}^{\nu} w_i \langle \varepsilon(x_i), g(x_i) \rangle \tag{4.12}$$

satisfies

$$\ell(\varepsilon) = \|\varepsilon\|, \quad \|\ell\| = 1, \quad \textit{and} \quad V \subset \ker(\ell). \tag{4.13}$$

Furthermore u remains a best approximation if the domain of the approximation problem is reduced to the finite set of points $\{x_i\}_{i=1}^{\nu}$.

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 (4.12) with the parameters x_i and w_i from (4.10). By the Cauchy–Schwarz inequality we obtain $\langle \varepsilon^*(x_i), g(x_i) \rangle \le |\varepsilon^*(x_i)|\, |g(x_i)| \le \|\varepsilon^*\|\, \|g\|$ with equality if $g = \varepsilon^*$. Since $\sum_i w_i = 1$, it follows that $\ell(g) \le \|g\|$, again with equality if $g = \varepsilon^*$, and the properties in (4.13) are verified.

Finally, assume that $u \in V$ and a functional with the properties (4.13) exists. We have for any $v \in V$

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

and u is a best approximation. □

Remark 4.4 Note that part (iii) of Theorem 4.3 is in the spirit of Theorem 1.1 in Singer (1970). This characterization is closely related to condition (3.8) in Corollary 3.6. To be precise assume that (iii) in Theorem 4.3 is satisfied and consider a design ξ^* with weights $w_1, \dots, w_\nu$ at the points $x_1, \dots, x_\nu$. 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 $v_1, v_2, \dots, v_n$ of the basis of V we obtain precisely condition (3.8). As we will see in the following discussion, functions satisfying only some of the properties in Theorem 4.3(iii) will also play an important role. Moreover, because of the linearity assumption it is easy to see that the solution of the optimization problem

$$\inf_{u \in V} \int |f(x) - u(x)|^2 d\xi^*(x)$$

is unique. Therefore Corollary 3.6 shows that ξ^* with masses w_i at the points x_i $(i = 1, \dots, \nu)$ is a T_p-optimal discriminating design, and this design has at most $n + 1$ support points.

Following the terminology in optimization theory we call the case $\nu = n+1$ the generic case; this means that this case is usually encountered. The next example, however, shows that degeneracy (i.e., $\nu < n + 1$) occurs already in simple cases.

Example 4.5 We reconsider Example 3.5 for the polynomial regression models (2.12). 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. The approximation problem is specified in (3.6). A quadratic polynomial f_1 is to be approximated by linear polynomials in the first component, and a cubic polynomial f_2 is to be 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 may investigate the reduced problem

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

Only 2 parameters are active, and by the Characterization Theorem there are optimal designs with at most 3 extreme points.

We now fix the given parameters $\rho_2 = \rho_3 = 1$ and the weights in the T_p-optimality criterion as $p_{2,1} = p_{3,2} = 1/2$. The best approximation is

$$u^*(x) = (1/2,\, x)^T,$$

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 $\ell(g)$ in Theorem 4.3 is given by

$$\ell(g) = \sqrt{2}\,\big(\frac{1}{4}g(-1) - \frac{1}{2}g(0) + \frac{1}{4}g(1)\big).$$

By the Characterization Theorem the associated T_p-optimal discriminating design is

$$\xi^* = \begin{Bmatrix} -1 & 0 & 1 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{Bmatrix}, \tag{4.14}$$

where the first line provides the support and the second one the associated masses. The degeneracy is now obvious. We have $n = 5$, but 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 both 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

$$p_{2,1}\rho_2^2 \geq p_{3.2}\rho_3^2. \tag{4.15}$$

If condition (4.15) 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 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

$$\xi^* = \begin{Bmatrix} -1 & -0.48 & +0.48 & +1 \\ 0.18 & 0.32 & 0.32 & 0.18 \end{Bmatrix}.$$

5 Concept of an algorithm – linearization

The characterization of best approximations in the previous section conceals that we have a nonlinear problem in the case of $d \geq 2$ from the viewpoint of their numerical construction. We

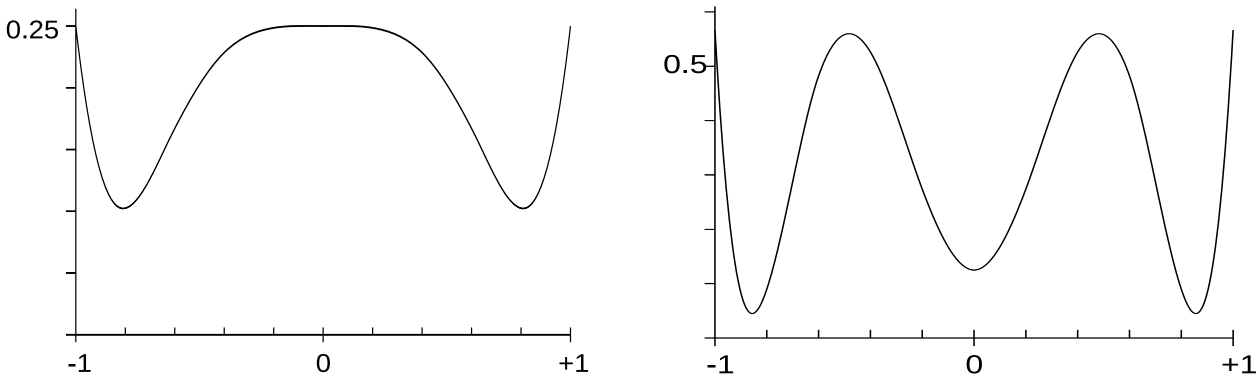

Figure 1: *Error functions* $\psi(x) = |f(x) - u^*(x)|^2$ *in the equivalence theorem for Example 4.5. Left panel:* $\rho_2 = \rho_3 = 1$*; right panel:* $\rho_2 = 1, \rho_3 = 4$.

will develop an algorithm that provides a T_p-optimal discriminating design in terms of the masses w_i and of the support points x_i. We restrict ourselves to linear models not only for the sake of simplicity. We want to emphasize in this way that the nonlinearities and the possible degeneracies are already encountered with linear models. The extension of the algorithm to nonlinear models is straightforward and will be given in Remark 6.2 below.

We propose a descent algorithm for the computation of a best approximation. During the iteration there will not only be a sequence of functions in the family V computed, but also a set of weights $\{w_1, \ldots, w_\nu\}$ and points $\{x_1, \ldots, x_\nu\}$. The collection of those points in an iteration step is denoted as *reference set*. The set $\{x_1, \ldots, x_\nu\}$ converges to a set which contains the set $\mathcal{A}$ of extreme points and by Remark 3.3 the support of the optimal design. The weight w_i corresponding to a point x_i converges to 0, whenever x_i belongs to a sequence that converges to a point in $\mathcal{X}\backslash\mathcal{A}$. The iteration contains two kinds of updates.

(A) For a given reference set $\mathcal{S} = \{x_1, \ldots, x_\nu\}$ $(\nu \geq n+1)$ the approximating function is updated while the reference set is kept. In particular, the differences between the errors $|\varepsilon(x_i)|$ at the ν points of the reference set is reduced. This procedure is denoted as an *approximate equilibration* or *equilibration*, for short. The result of these calculations is an "improved" approximating function, say u.

Additionally we obtain the weights for a design ξ with support S by solving a dual linear program. The computation of the quantity $\Delta(\xi)$ (cf, Remark 5.3 below) yields a lower bound of the T_p-efficiency

$$\mathrm{Eff}_{T_p}(\xi) = \frac{\Delta(\xi)}{\sup_\eta \Delta(\eta)} \geq \frac{\Delta(\xi)}{\|\varepsilon\|^2} \tag{5.1}$$

where $\varepsilon = f - u$. In particular, this provides a stopping criterion for the algorithm. The iteration will be stopped if the lower bound is close to 1.

(B) The reference set is updated while the approximating function is kept. Candidates for the new set are points in the domain for which the error is larger than the error on the actual reference set. Points at which the error is small can be dropped in the reference set.

Looking at the Characterization Theorem one expects that reference sets with $n+1$ points are a good choice. However, situations as shown in Example 4.5 are typical. Frequently we find degeneracies, and therefore we have to care for robustness of the procedures. Robustness is more easily achieved if reference sets $\mathcal{S}$ with more than $n+1$ points are admitted. This means that there may exist points which do not contribute to the sum in (4.10). These points can be dropped in Step (B) when new points are inserted.

The error curve $|f-u|$ may be shown on the screen of the computer after each update of u. Usually it is not difficult to decide whether and how the reference set is to be improved. Therefore we focus on Step (A) and the equilibration, since this is the more involved part of the iteration. Moreover, usually several steps of type (A) are performed between updates of the reference set with steps of type (B); cf. the tables in Section 7.

We proceed in the spirit of Newton's method and ignore temporarily the quadratic term of the correction v in the Binomial formula. Specifically, given a finite reference set $\mathcal{S}$ and a guess u for the approximating function, we replace the optimization problem

$$\max_{x_i\in\mathcal{S}} \big|f(x_i)-u(x_i)-v(x_i)\big|^2 = \max_{x_i\in\mathcal{S}}\big\{|\varepsilon(x_i)|^2 - 2\,\langle\varepsilon(x_i),v(x_i)\rangle + |v(x_i)|^2\big\} \to \min_{v\in V}! \tag{5.2}$$

by the linear program

$$\max_{x_i\in\mathcal{S}}\big\{|\varepsilon(x_i)|^2 - 2\,\langle\varepsilon(x_i),v(x_i)\rangle\big\} \to \min_{v\in V}! \tag{5.3}$$

While the left-hand side of (5.2) is obviously bounded from below, this is not always true for the optimization problem (5.3). The boundedness, however, is essential for the algorithm.

Definition 5.1 A function $u\in V$ is called *dual feasible* for the set $\mathcal{S}$, if the left-hand side of (5.3) is bounded from below.

The notation of dual feasibility will be clear from the dual linear program (5.5) and Lemma 5.2. We will also see in Remark 5.3 that dual feasible functions are associated to a design in the spirit of (2.2).

The minimization of a linearized functional on a finite set $\mathcal{S}=\{x_i\}_{i=1}^{\nu}$ with $\nu\ge n+1$ as in (5.3) will be the basis of our algorithm. We will get more insight from the dual programs. In particular, we obtain the masses w_i of the optimal discriminating designs and (equivalently) of the functional (4.12) defined in Theorem 4.3.

For a given actual error function ε and a reference set with ν points $x_1, \ldots, x_\nu$ we may use the representation (4.8) and rewrite the (primal) linear program (5.3) as a linear program for the $n+1$ variables $E, \alpha_1, \alpha_2, \ldots, \alpha_n$

$$
\begin{aligned}
E &\rightarrow \min! \\
2 \sum_{m=1}^{n} \alpha_m \langle \varepsilon(x_i), v_m(x_i) \rangle + E &\geq |\varepsilon(x_i)|^2, \quad i = 1, 2, \ldots, \nu.
\end{aligned}
\tag{5.4}
$$

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 (5.4) contains the equations for the ν weights $w_1, w_2, \ldots, w_\nu$ with the adjoint matrix, where we can drop the factor of 2

$$
\begin{aligned}
\sum_{i=0}^{\nu} w_i |\varepsilon(x_i)|^2 &\rightarrow \max! \\
\sum_{i=0}^{\nu} w_i \langle \varepsilon(x_i), v(x_i) \rangle &= 0, \quad \forall\, v \in V, \\
\sum_{i=0}^{\nu} w_i &= 1, \quad w_i \geq 0, \quad i = 1, 2, \ldots, \nu.
\end{aligned}
\tag{5.5}
$$

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* (5.5) *has a feasible point and a solution if and only if the target functional in the linear program* (5.4) *is bounded from below, i.e.,*

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

Since V is a linear space, this condition is equivalent to

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

Remark 5.3 If the linear program (5.5) has a feasible point, there is a solution for which at most $n+1$ of the w_i are positive. We obtain a linear functional ℓ of the form (4.12) with these parameters that satisfies

$$\|\ell\| = 1 \quad \text{and} \quad V \subset \ker(\ell).$$

We have $\ell(\varepsilon) < \|\varepsilon\|$ as long as we have not reached a best approximation. Since the values of the primal program (5.4) and the dual program (5.5) coincide, we have also $E = \sum_{i=0}^{m} w_i |\varepsilon(x_i)|^2$.

Moreover, we obtain a lower bound for the degree of approximation. To be precise, let ξ be the measure with masses w_i at the points x_i. From Theorem 3.2 and (5.5) it follows that

$$
\begin{aligned}
\inf_{v\in V} \|f-v\|^2 \geq \Delta(\xi) &= \sum_i w_i \inf_{v\in V} |(f-u-v)(x_i)|^2 \\
&= \inf_{v\in V} \sum_i w_i \Big(|\varepsilon(x_i)|^2 - 2\,\langle \varepsilon(x_i), v(x_i)\rangle + |v(x_i)|^2 \Big) \\
&= \inf_{v\in V} \sum_i w_i \Big(|\varepsilon(x_i)|^2 + |v(x_i)|^2 \Big) \;=\; \sum_i w_i |\varepsilon(x_i)|^2. \qquad (5.6)
\end{aligned}
$$

We emphasize that more information is provided by (5.6) since the infimum in the last line of (5.6) is attained at $v = 0$. Given a dual feasible function $u = \eta(\cdot, \theta)$, we have $\theta = \theta^*(\xi)$ for the design ξ specified above. *We have found a design ξ to which u is associated in the spirit of the characterization* (2.2).

Example 5.4 For the sake of transparency, we illustrate the existence of dual infeasible functions by an example with univariate functions on the interval $[-1, 1]$. Let V be the set of linear polynomials and $f(x) := 1 + x + x^3$. The function $u = 0$ is not dual feasible for the set $S = \{-1, 0, +1\}$. Indeed, we have a relation as in (5.5) with a negative weight,

$$
\frac{3}{8}\varepsilon(-1)v(-1) + \frac{6}{8}\varepsilon(0)v(0) - \frac{1}{8}\varepsilon(+1)v(+1) = 0 \quad \forall v \in V,
$$

and there does not exist such a relation with positive weights. By the Kolmogorov criterion, this is in accordance with the fact that the function $v_0 := 1 + 2x$ satisfies

$$
\varepsilon(x_i)v_0(x_i) > 0 \quad \text{for } i = 0, 1, 2,
$$

and consequently the function $u = 0$ is not dual feasible.

6 Newton's Method and its Adaptation

The improvement of the approximation on a given reference set $\mathcal{S}$ will be done iteratively by Newton's method. In order to avoid the introduction of one more symbol for the specific iteration, we focus on one step of the iteration for the given input u_0 and corresponding error function ε_0. The following algorithm looks natural, it is however only the basis of our algorithm:

Given u_0 and $\mathcal{S}$, find a solution of the linear program (5.4) *for $u = u_0$, set $v = \sum_m \alpha_m v_m$, and $u_1 = u_0 + v$ is the result of the Newton step.*

In order to achieve a robust procedure, we have to modify it into three directions, which will be explained in the following discussion. The damping as in item (2) is a standard tool in modern algorithms, but the other modifications are specific for the optimal design problem. From the discussion in the previous section we already know that we have to admit a function u_0 in the input of an iteration step which is not necessarily dual feasible. The degeneracy of the support of the optimal design makes it also difficult to find a dual feasible approximation.

For convenience, we use the notation $\|g\|_{\mathcal{S}} := \sup_{x\in\mathcal{S}} |g(x)|$.

(1) Choose a bound $\bar{\alpha}$ and add the restriction

$$\sum_{m=1}^{n} |\alpha_m| \le \bar{\alpha}$$

to the linear program (5.4). Now the domain of the feasible vectors α is bounded, and there always exists a solution even if the original input corresponds to a dual infeasible problem. – The parameter $\bar{\alpha}$ will also be updated; cf. the next item.

(2) The Newton correction v will be multiplied by a damping factor t. Obviously we have only to care for the case that u_0 is not yet a best approximation. By definition of the Newton method,

$$\max_{x\in\mathcal{S}}\{|\varepsilon_0(x)|^2 - 2\,\langle\varepsilon_0(x_i), v(x_i)\rangle < \|\varepsilon_0\|_{\mathcal{S}}^2.$$

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 $\|f-u_0-tv)\|_{\mathcal{S}}^2 < \|\varepsilon_0\|_{\mathcal{S}}^2$ for sufficiently small positive factors t and for this choice an improvement is generated. Let $T := \{2, 1, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, \ldots, 0\}$ and determine

$$t = \operatorname*{argmin}_{t\in T} \|f-u_0-tv\|_{\mathcal{S}}\,. \tag{6.1}$$

The standard set of damping factors: $1, 2^{-1}, 2^{-2}, \ldots$ has been augmented. The number $0 \in T$ guarantees that the new approximation is at least as good as the old one. If the minimum is attained at $t = 2$, this is a hint that the bound $\bar{\alpha}$ is too small. In any case, $\bar{\alpha}$ should be replaced in the next iteration step by a number between $\bar{\alpha}$ and $t\bar{\alpha}$.

Since the functions $t \to |(f-u_0-tv)(x_i)|^2$ are quadratic polynomials, the execution of (6.1) is very cheap.

(3) As illustrated in Example 4.5, the best approximations have less than $n+1$, say ν, extreme points in many cases. Therefore it is natural to consider optimization problems also on ν-dimensional subspaces. There is one more motivation. Given a design ξ, the minimization of $\Delta(\xi)$ can be split into the d subproblems of minimizing $\Delta_{(i,j)}(\xi)$ for $(i,j) \in \mathcal{I}$. – Of

course, the pairs (i, j) refer to the notation in Section 2.

Now we write the space of approximating functions as a sum of d subspaces

$$V = \oplus_{(i,j)\in\mathcal{I}} V_{(i,j)}\,, \tag{6.2}$$

where $V_{(i,j)}$ contains those functions in V that correspond to $\{\eta_j(\cdot, \theta_{(i,j)}) \mid \theta_{(i,j)} \in \Theta_{(i,j)}\}$. In addition to the optimization (5.4), we solve the d optimization problems with V replaced by the subspaces $V_{(i,j)}$, $(i, j) \in \mathcal{I}$.

Remark 6.1 As was just noted, the support of the optimal design consists of less than $n+1$ points in many cases of actual interest. This has another consequence. We cannot decide by a simple inspection of the error curve during the iteration whether we are already close to the optimum or not. We may do it by comparing the upper bound for the degree of approximation $\|\varepsilon\| = \|f - u\|$ with the lower bound $\Delta(\xi)$ for an appropriate ξ. This gives the lower bound (5.1), which can be used as a stopping criterion of the algorithm. If u is dual feasible, we get an associated ξ from the dual program (5.5) and (5.6). Otherwise we will relax the dual program and solve (8.2) below. The points with negative weights are dropped, and the posivive weights are renormalized to have $\sum w_i = 1$. We obtain a reasonable ξ, at least in the neighborhood of an optimum. The negative weights will be small there, and the procedure above implies only small changes. Thus designs and their corresponding efficiencies are computed simultaneously by solving the approximation problem iteratively.

A detailed code is postponed to Section 8.

Remark 6.2 *(Remark on the adaptation to nonlinear models)*
If the models $\eta_1, \eta_2, \ldots, \eta_k$ depend nonlinear on the parameters, the approximating function $u(x, \theta)$ depends in a (possibly) nonlinear way on θ, and the linear program (5.4) has to be replaced by the linear program

$$\begin{aligned} E &\to \min! \\ 2\sum_{m=1}^{n} \alpha_m \left\langle \varepsilon(x_i), \frac{\partial}{\partial\theta_m} u(x_i) \right\rangle + E &\geq |\varepsilon(x_i)|^2, \quad i = 1, 2, \ldots, \nu. \end{aligned} \tag{6.3}$$

The solution α yields the update of θ. The adaptation of the Newton method described at the beginning of this section is performed also here in an obvious way, and we illustrate the application of the algorithm in the context of nonlinear regression models in Example 7.2.

Table 1: *The results of the algorithm for Example 7.1. The fourth column shows the lower bound for the efficiency defined in* (5.1)*.*

j	$\|\varepsilon_j\|^2$	$\Delta(\xi_j)$	$\frac{\Delta_{lin}(\xi_j)}{\|\varepsilon_j\|^2}$	dual feasible	t	extras
0	12.5	0.0168	0.0013	no		$\bar{\alpha} \leftarrow 1$
1	4	0.0444	0.011	no	1	
2	0.6944	0.0204	0.0294	no	1	
3	0.3209	0.0063	0.0196	no	0.5	
4	0.2161	0.0356	0.1649	no	0.25	$\bar{\alpha} \leftarrow 0.5$
5	0.1901	0.0689	0.3623	no	1	
6	0.1684	0.0746	0.4432	no	0.5	$\mathcal{S} \leftarrow \mathcal{S} \cup \{-0.2\}$
7	0.1501	0.0963	0.6417	no	1	
8	0.1342	0.1138	0.848	yes	1	
9	0.1318	0.1189	0.902	yes	0.25	$\bar{\alpha} \leftarrow 0.1$
10	0.1283	0.1200	0.9348	no	1	
11	0.1254	0.1223	0.9754	no	0.25	$\bar{\alpha} \leftarrow 0.05$
12	0.1252	0.1229	0.9816	no	0.25	

7 Numerical results

In this section we illustrate the algorithm described in the previous section in two examples with linear and nonlinear regression functions. The first example considers the linear case.

Example 7.1 We consider once more the situation as in Example 3.5, 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 Newton iteration 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 iterated functions do not have the symmetry properties discussed in Example 4.5. The reference set at the start is $S := \{-1, -0.5, -0.1, 0, 0.1, 0.5, 1\}$, and the point -0.2 was added in step 6 of the iteration.

The results of the algorithm are displayed in Table 1. After 12 iteration steps we obtain an approximation such that the degree of approximation does not exceed the lower bound by more than 2%, and the efficiency of the computed discriminating design is larger than 98%. In the first part of the iteration the lower bound is very small and is of no use. Note also that the bound is not monotonously increasing in the iterations. 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 iteration steps. The final result in Figure 1 shows that afterwards there

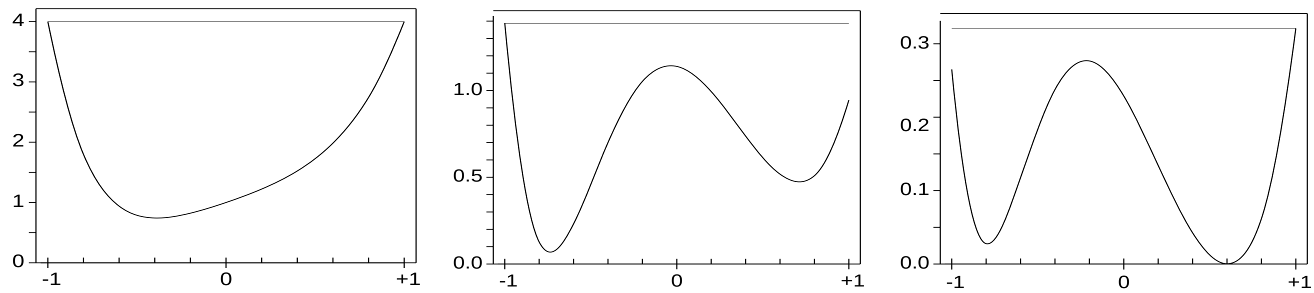

Figure 2: *Error curve $|f-u|^2$ in Example 7.1 in the first three iteration steps.*

are no great changes of the shape. The resulting discriminating design

$$\xi^* = \begin{Bmatrix} -1 & 0 & 0.1 & 1 \\ 0.2516 & 0.3392 & 0.1562 & 0.2530 \end{Bmatrix},$$

[where $\theta_{(2,1)} = (1.50005,\ 0.99991)$, $\theta_{(3,2)} = (0.99998,\ 1.99041,\ 1.0057)$] may be compared with the exact optimal one given in (4.14). Note that the maximum in Figure 1.a is very flat. This and the degeneracy yields two points 0 and 0.1 instead of one extreme point 0. If we would continue with the iteration, the two points would eventually coalesce.

Table 2: *The results of the algorithm for Example 7.2.*

j	$\|\varepsilon_j\|^2$	$\Delta_{lin}(\xi_j)$	$\frac{\Delta_{lin}(\xi_j)}{\|\varepsilon_j\|^2}$	dual feasible	t	extras
0	1.25301	0.0062	0.0049	no		$\bar{\alpha} \leftarrow 1$
1	0.07348	0.00401	0.0545	no	1	
2	0.01632	0.00521	0.3189	no	1	$\mathcal{S} \leftarrow \mathcal{S} \cup \{0.6\}$
3	0.00722	0.00642	0.8898	no	0.25	$\bar{\alpha} \leftarrow 0.25$
4	0.00707	0.00672	0.9499	yes	0.0625	$\mathcal{S} \leftarrow \mathcal{S} \cup \{0.5, 3.3\}$
5	0.00681	0.00671	0.9854	no	0.0625	$\bar{\alpha} \leftarrow 0.05$
6	0.00680	0.00678	0.9965	no	0.125	$\mathcal{S} \leftarrow \mathcal{S} \cup \{3.4, 3.5\}$
7	0.00679	0.00678	0.9992	no	0.25	

Example 7.2 In order to demonstrate that the algorithm can be used for the calculation of T_p-optimal discriminating designs in case of nonlinear regression models we consider two rival models

$$\eta_1(x,\theta) = \frac{\theta_{11}x}{x+\theta_{12}}, \quad \eta_2(x,\theta) = \theta_{21}(1-e^{-\theta_{22}x})$$

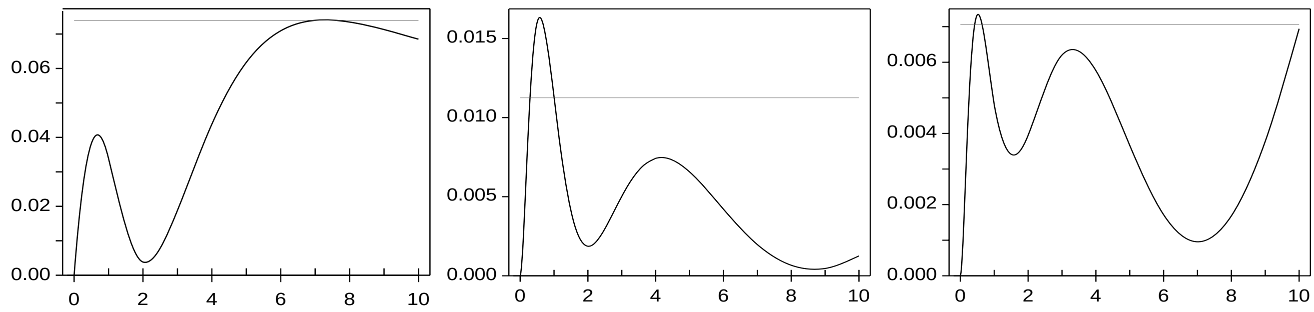

Figure 3: *Error curve $|f-u|^2$ in Example 7.2 in the first three iteration steps.*

where $\rho_{(1)} = (2.0,\, 1.0)$ and $\rho_{(2)} = (2.5,\, 0.5)$. The weights in the criterion (2.6) are defined by $p_{1,2} = p_{2,1} = 1/2$. 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 error curves in Figure 3 show that the referencee set has to be updated after the second iteration step. The degree of approximation is close to the optimum already after 4 iteration steps. Further iteration steps improve the lower bound for the efficiency defined in (5.1). The resulting design is

$$\xi^* = \begin{Bmatrix} 0.5 & 3.4 & 3.5 & 10.0 \\ 0.304 & 0.143 & 0.278 & 0.275 \end{Bmatrix},$$

and the parameters corresponding to the best uniform approximation are given by (subject to rounding) $\bar\theta_{(1,2)} = (3.006,\, 1.804)$, and $\bar\theta_{(2,1)} = (1.721,\, 0.868)$. The efficiency of 99% listed in Table 2 refers to the linearization and therefore to the consideration in the neighborhood of the solution. Nevertheless we can use the equivalence Theorem 3.1 to check if this design is in fact T_p-optimal. We have performed an extensive search for the parameter θ^* defined in (2.3) and found that it equals $\bar\theta$ subject to rounding, i.e. $\bar\theta^* = \bar\theta$. The corresponding plot of the function ψ^* in the equivalence Theorem 3.1 is shown in Figure 4. We observe that the design ξ^* is in fact a T_p-optimal discriminating design.

Note that the support of the resulting design is much smaller than the reference sets during the iteration. The degeneracy here has the effect that the second extreme point is split into the two points 3.4 and 3.5. We observe that more damping of the Newton steps is required in this example. It is known that approximation problems with exponential functions are ill conditioned compared to polynomial problems.

8 Details of the algorithm

In this section we present one Newton step for the equilibration with its details. As usual, an instruction $A \leftarrow B$ in the algorithm means that the value of A is to be replaced by B. The loops over finite sets of indices, however, are written as in mathematical formulas and not as usually

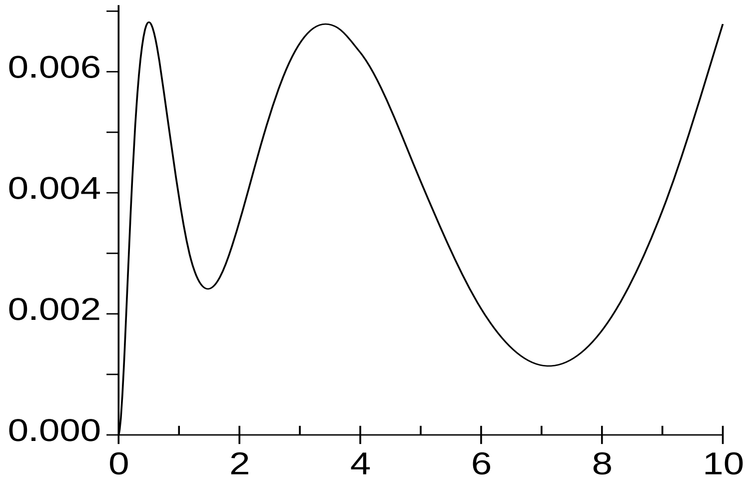

Figure 4: *The function ψ in the equivalence Theorem 3.1 for Example 7.2.*

in computer codes.

Algorithm 8.1 *(One Newton step for the equilibration)*

Input

an approximation $u_0 = \sum_{m=1}^{n} \gamma_m v_m$, that need not be dual feasible,
a reference set $\mathcal{S} = \{x_i\}_{i=1}^{\nu}$,
a bound $\bar{\alpha}$ for the correction of the vector $\gamma = (\gamma_1, \ldots, \gamma_n)^T$.

Coefficients of the linear program

$$\left.\begin{array}{rl} \varepsilon_0(x_i) & \leftarrow f(x_i) - u_0(x_i), \\ r_{im} & \leftarrow \langle \varepsilon_0(x_i), v_m(x_i) \rangle\,,\ m = 1, 2, \ldots, n, \\ b_i & \leftarrow |\varepsilon_0(x_i)|^2, \end{array}\right\} i = 1, 2, \ldots, \nu,$$

Computation of the weights

Solve the linear program (with respect to $w_1, \ldots, w_\nu$)

$$\begin{array}{rcl} \sum_{i=1}^{\nu} b_i w_i & \to & \max! \\ \sum_{i=1}^{\nu} r_{im} w_i & = & 0, \quad m = 1, 2, \ldots, n, \\ \sum_{i=1}^{\nu} w_i & = & 1, \quad w_i \geq 0,\ i = 1, 2, \ldots, \nu. \end{array} \tag{8.1}$$

print 'reference set' $\mathcal{S}$, 'actual error' b, 'weights' w.

If the linear program (8.1) has a solution, then

print 'lower bound of the degree of approximation', $\sum_i w_i |\varepsilon_0(x_i)|^2$

else

{

How far are we from dual feasibility?
Solve the linear program [by setting $w_i = w_i^+ - w_i^-$, $w_i^+ \geq 0$, $w_i^- \geq 0$]

$$\begin{array}{rcl} \sum_{i=1}^{\nu} b_i \max(w_i, 0) & \rightarrow & \max! \\ \sum_{i=1}^{\nu} r_{im} w_i & = & 0,\ m = 1, 2, \ldots, n, \\ \sum_{i=1}^{\nu} |w_i| & = & 1. \end{array}$$

print 'relaxed weights' w.
$w_i \leftarrow \max\{w_i, 0\},\ i = 1, 2, \ldots, \nu$,
Normalize weights to obtain $\sum_i w_i = 1$.
Define ξ by x_i, w_i # and compute the lower bound $\Delta(\xi)$.

$$\left.\begin{array}{l} B_m \leftarrow \sum_{i=1}^{\nu} w_i r_{im}, \\ A_{Mm} \leftarrow \sum_{i=1}^{\nu} w_i \langle v_M(x_i), v_m(x_i) \rangle,\ M = 1, 2, \ldots, n, \end{array}\right\} m = 1, 2, \ldots, n.$$

Solve $A\alpha = B$,
$\Delta(\xi) \leftarrow \sum_{i=1}^{\nu} w_i |\varepsilon_0(x_i)|^2 - \sum_{m=1}^{n} B_m \alpha_m$.
print 'lower bound of the degree of approximation', $\Delta(\xi)$.
}

Computation of the basic Newton correction
Solve the linear program (with respect to $E, \alpha_1, \ldots, \alpha_n$)

$$\begin{array}{rcl} E & \rightarrow & \min! \\ 2\sum_{m=1}^{n} \alpha_m r_{im} + E & \geq & |\varepsilon_0(x_i)|^2, \quad i = 1, 2, \ldots, \nu, \\ \sum_{m=1}^{n} |\alpha_m| & \leq & \bar{\alpha}. \end{array}$$

Newton correction δu (without damping).
$\delta u \leftarrow \sum_{m=1}^{n} \alpha_m v_m$,
The degree of approximation is determined for several damping factors t:
The errors are quadratic polynomials in t: $|(f - u_0 - t\delta u)(x_i)|^2 = p_i(t)$.
Define $p_i(t) = (1 - t)b_i + e_i t + q_i t^2$.

$$\left.\begin{array}{l} e_i \leftarrow b_i - 2\sum_{m=1}^{n} r_{im}\alpha_m, \\ q_i \leftarrow |\delta u(x_i)|^2 \end{array}\right\} i = 1, 2, \ldots, \nu,$$

print 'errors after the Newton step with quadratic terms ignored', e.
$T \leftarrow \{2, 1, 1/2, 1/4, 1/8, 1/16, 0\}$.

$$\left.\begin{array}{l} \tilde{b}_t \leftarrow \max_{0 \leq i \leq \nu}\{p_i(t)\}, \\ \text{print 'new error with damping', } t,\ \tilde{b}_t, \end{array}\right\} t \in T.$$

Determine best damping factor.
$t \leftarrow \operatorname{argmin}\{\tilde{b}_t\}$,

Final update.
$u_0 \leftarrow u_1 = u_0 + t\delta u,$
$\gamma_m \leftarrow \gamma_m + t\alpha_m, \; m = 1, 2, \ldots, n,$
If $(t > 0)$ then $\bar{\alpha} \leftarrow t\bar{\alpha}$.
print 'coefficients of new approximation', $\gamma_m, \; m = 1, \ldots, n.$ □

Acknowledgements. The authors want to thank Stefan Skowronek for providing a code for the numerical calculations. – 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(9):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.

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

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

Denisov, V. I., Fedorov, V. V., and Khabarov, V. I. (1981). Tchebysheff approximation in problems of constructing asymptotic locally optimal designs for discriminating regression experiments. In Fedorov, V. and Nalimov, V., editors, *Problems in Cibernetics (Voprosy Kibernetiki in Russian), Linear and Non-Linear Parametrization in Optimal Design*, pages 3–10. Akademiya Nauk RSFSR, Nauchnyj Sovet po Kompleksnoj Probleme Kibernetika, Moskva.

Dette, H. (1994). Discrimination designs for polynomial regression on a compact interval. *Annals of Statistics*, 22(2):890–904.

Dette, H. (1997). Designing experiments with respect to "standardized" optimality criteria. *Journal of the Royal Statistical Society, Ser. B*, 59:97–110.

Dette, H., Bretz, F., Pepelyshev, A., and Pinheiro, J. (2008). Optimal designs for dose-finding studies. *Journal of the American Statistical Association*, 104(483):1225–1237.

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. and Titoff, S. (2009). Optimal discrimination designs. *Annals of Statistics*, 37(4):2056–2082.

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

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

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2(5):849–879.

Kolmogorov, A. N. (1948). A remark concerning the polynomials of P. L. Chebyshev which deviate the least from a given function. [in Russian]. *Uspehi Mat. Nauk*, 3:216–221.

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

Singer, I. (1970). *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces.* Springer, Berlin.

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

Spruill, M. C. (1990). Good designs for testing the degree of a polynomial mean. *Sankhya, Ser. B*, 52(1):67–74.

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

Studden, W. J. (1982). Some robust-type d-optimal designs in polynomial regression. *Journal of the American Statistical Association*, 77(380):916–921.

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(1):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.

Zen, M.-M. and Tsai, M.-H. (2002). Some criterion-robust optimal designs for the dual problem of model discrimination and parameter estimation. *Sankhya: The Indian Journal of Statistics*, 64:322–338.

Zen, M.-M. and Tsai, M.-H. (2004). Criterion-robust optimal designs for model discrimination and parameter estimation in Fourier regression models. *Journal of Statistical Planning and Inference*, 124:475–487.