Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2012-59

On an integrated compound criterion

Elena Bukina[a] and Milan Stehlík

March 2012

[a]Université de Nice-Sophia Antipolis

Abstract

In this paper we introduce a class of integrated compound criteria. We will show that for a specific choice the criterion can mimic well a behavior of an integrated mean square prediction error (IMSPE). Such a method can be algorithmized and thus gave an insight into the relation of D_α and IMSPE criterion. Theoretical and numerical aspects are discussed and examples provided.

keywords: Compound Designs, Computer experiments, Correlated errors, Experimental design; Equidistant design; Parameterized covariance functions, Fredholm equation, reproducing kernel, regularization.

1 Introduction and Setup

Probably the main reason why optimum design principles are not frequently used in spatial data analysis is that the observations are correlated. Consequently, the corresponding optimal design questions must cope with the existence and detection of an error correlation structure, problems largely unaccounted for by traditional optimal design theory. In all of these situations there arise a number of issues, which require special techniques - for a recent discussion see (Müller and Stehlík (2009)). The statistical model we consider in the paper is the so called random field, given by

$$Y\left(x\right)=\eta(x,\beta)+\varepsilon_{\gamma}\left(x\right) \tag{1}$$

with design points (coordinates of monitoring sites) $\xi_n=\{x_1,...,x_n\}$ taken from a compact design space $\mathcal{X}=X^n, X=[a,b], -\infty<a<b<\infty$. The parameters β are unknown and the variance-covariance structure of the errors depends on parameters γ. However, one can, if one is willing to make distributional assumptions, employ the ML-estimators. For the full parameter set $\{\beta,\gamma\}$ the information matrix then exhibits the block diagonal form

$$\left(\begin{array}{cc} M_\beta(\xi) & 0 \\ 0 & M_\gamma(\xi) \end{array}\right).$$

Outputs from various environmental measurements are often approximated as realizations of correlated random fields. Two approaches are considered to design experiments for a correlated random field when the objective is to obtain precise predictions over the whole experimental domain. The first one corresponds to a compound D_α-optimality criterion for both the trend and covariance parameters introduced by (Müller and Stehlík (2010)). The second one relies on an approximation of the mean squared prediction error already proposed in the literature. In (Müller and Pronzato (2009)) was conjectured, and shown on an example, that for some particular settings both approaches yield similar optimal designs, thereby revealing a sort of equivalence theorem for random fields. For estimations of spatial fields a classical criterion is the Empirical Kriging prediction error. Here we have to minimize the so-called kriging variance $Var[\widehat{Y}(x|\xi)]=E[(\widehat{Y}(x|\xi)-Y(x))^2]$ (Mean Squared Prediction Error - MSPE), where $\widehat{Y}(x|\xi)$ denotes the best linear unbiased predictor of $Y(x)$ based on the design points in ξ. The EK-optimal design minimizes the criterion function $\psi(\xi) = \max_{x \in X} \widehat{Var[\widehat{Y}(x|\xi)]}$. However, this criterion is difficult to compute. Much easier criterion is so-called D_α-optimality criterion $\phi(\alpha, \xi)$ introduced by (Müller and Stehlík (2010)), which is defined in terms of the Fisher Information matrix. The D_α-optimal design with a weighting factor α maximizes the criterion function $\phi(\alpha, \xi) = |M_\beta(\xi)|^\alpha |M_\gamma(\xi)|^{1-\alpha}, \quad 0 \leq \alpha \leq 1$, there $M_\beta(\xi)$ and $M_\gamma(\xi)$ are information matrices for the parameters β and γ. See (Kiseľák and Stehlík (2008)) for the details on the structure of $M_\beta(\xi)$ and $M_\gamma(\xi)$ in the case of Ornstein-Uhlenbeck process.

Here we concentrate on IMSPE, instead of MSPE. The main aim of the kriging technique consists in predicting output of the simulator on the experimental region, and for any untried location x the estimation procedure is focused on the BLUP $\hat{Y}(x)$. Thus natural criteria will minimize suitable functionals of the MSPE. Since often the prediction accuracy is related to the entire prediction region X^n, a very practical design criterion is Integrated MSPE given by $\int_{X^n} \sigma^{-2} MSPE(\hat{Y}(\xi)) d\xi$. IMSPE was used in several papers (see (Sacks, Schiller and Welch (1989)) or (Crary (2002))) for construction of optimal designs.

In this paper we introduce the integral variant of the compound criterion by operator

$$(L\,h)(\xi) = \int_0^1 \phi(\alpha, \xi) h(\alpha)\, d\alpha, \tag{2}$$

where $h \in D$, D being a space of functions. We study some of the properties of this criterion. We show that having equation $-IMSPE = (L\,h)(\xi)$ we can find such an $h \in L^2(0,1)$ (at least numerically).

Definition (2) is given by the direct integration of D_α criterion given by (Müller and Stehlík (2010)). There are several possibilities of its modification. One possible modification is related to region for α. There is no particular reason for restriction $\alpha \in [0,1]$. However, we assume in this paper that α belongs to a compact interval and thus 'for the sake of simplicity' we normed this interval to be $[0,1]$. Also an important question is the dimensionality of α. For designs with large number of points, multivariate $h(\alpha_1, ..., \alpha_m)$ may be more appropriate. Then $Lh = \int_{[0,1]^m} |M_\beta|^{\alpha_1} .. |M_\beta|^{\alpha_m} |M_r|^{1-\alpha_1} .. |M_r|^{1-\alpha_m} h(\alpha_1, ..., \alpha_m) d\alpha_1 ... d\alpha_m$

In (2) we integrate directly compound criterion $\phi(\alpha, \xi)$ given by (Müller and Stehlík (2010)). Until now we do not have justification whether this choice is optimal, i.e. other form of compounding, generally $F_\alpha(\det M_\beta, \det M_r)$ for an appropriate function F may give a better properties of integrated criterion $\int_0^1 F_\alpha(\det M_\beta, \det M_r) h(\alpha)\, d\alpha$. As an alternative we may introduce $\phi_2(\alpha, \xi) = |M_\beta(\xi)|^{-\alpha/p} |M_\gamma(\xi)|^{-(1-\alpha)/q}, \quad 0 \leq \alpha \leq 1, p = \dim(\beta), q = \dim(\gamma)$ and study Fredholm equation $\psi = L_2 h$, where operator L_2 is given by kernel ϕ_2. From numerical reasons (ill conditioned matrix) we use also criterion $F_\alpha(\det M_\beta, \det M_r) = \alpha \log |M_\beta(\xi)| + (1-\alpha) \log |M_\gamma(\xi)|$. Notice, that in the latter case we obtain the criterion of the form $A \log |M_\beta(\xi)| + (1-A) \log |M_\gamma(\xi)|$ where $A = \int_0^1 \alpha h(\alpha) d\alpha$ (from practical reason we later regularize by $h > 0, ||h|| = 1$). Moreover, this approach relates to a general compound criterion treated in literature (see e.g. (McGree et al. (1988))).

The paper is organized as follows. In the next section we study some properties of class of integrated compound criteria (2). In section 2 we provide (by means of Fredholm integral operators) the theoretical backgrounds for existence of specific criterions, related to IMSPE for a large class of stochastic processes. In sections 3 and 4 we illustrate by numerical experiments and theoretical arguments the need for regularization of Fredholm equation. Discussion concludes the paper.

2 On some properties of a class of compound criteria

In this section we reveal some properties of integrated compound criteria (2).

2.1 Integrated compound criteria mimics nugget effect

As was observed (see e.g. (Kiseľák and Stehlík (2008))) two point D-optimal design $\{x_1, x_2\}$ for a constant trend parameter η (and taking γ as parameter of interest) is "collapsing", e.g. maximum of determinant of Fisher information is attained for $x_1 = x_2$. As can be easily checked, this "collapsing" employs also in the case of compound criterion $\phi(\alpha, \xi)$ for all $\alpha \in (0, 1)$. However a non-constant h may regularize this "collapsing" effect (see following Example 1), e.g. integrating of $\phi(\alpha, \xi)$ with non-constant h works similarly as "nugget effect".

Example 1 *Let us consider OU process,2 point design, estimation also for a correlation parameter*

$$M_\beta = \frac{2}{1+e^{-\gamma d}} \text{ and } M_\gamma = \frac{d^2 e^{-2\gamma d}(1+e^{-2\gamma d})}{(1-e^{-2\gamma d})^2}. \tag{3}$$

$$F(d,\alpha) = \alpha M'_\beta M_\gamma + (1-\alpha) M_\beta M'_\gamma. \tag{4}$$

For the case $\gamma = 1$,

$$\lim_{d\to\infty} F(d,\alpha) = \lim_{d\to\infty} \frac{\partial}{\partial\alpha}\phi(d,\alpha) = -\frac{1}{2} + \frac{3\alpha}{4}. \tag{5}$$

For $0 < \alpha < 1$ $\max \phi(d, \alpha)$ *is attained at* $d = 0$ *This means, that* D_α *optimal design is attained for* $d = 0$. *We call this phenomenon "collapsing effect", it was reported recently (see e.g. (Crary (2002)) and for recent discussion (Müller and Stehlík (2009))).*

Let us consider

$$(Lh)(d) = \int_0^1 \phi(d,\alpha)h(\alpha)\,d\alpha,$$

then for $h(\alpha) = 1$ *the maximum* $\max(Lh)(d)$ *is at* $d = 0$ *(i.e. not improvement in collapsing effect)..*

However, for $h(\alpha) = \alpha, \alpha^2, e^\alpha, \sin\alpha$

$$\arg\max(Lh)(d) > 0.$$

This means improvement, i.e. no more collapsing! That means, we have obtained the regularization of collapsing by integral compound criteria. The other way of regularizing of collapsing is to employ a nugget (see (Müller and Stehlík (2009))).

If we know theoretically, that $\psi = Lh$ is solvable in some class H of function, then, instead of optimizing of ψ we may optimize Lh for $h \in H$. Notice, that optimizing of Lh may be different than optimizing of its kernel ϕ (see Example 1, where Lh has a different optimal designs for $h = 1$ and $h(\alpha) = \alpha$).

2.2 Integrated compound criteria mimics IMSPE

Next example illustrates the case of Ornstein Uhlenbeck process.

Example 2 ***Ornstein-Uhlenbeck model with constant trend***

In this example we consider Ornstein-Uhlenbeck model with constant trend given by $\psi(\xi) = IMSPE$. *Let us have Gaussian random field* $Y(x) = \mu + \epsilon_\gamma(x)$, $corr(\epsilon_\gamma(x), \epsilon_\gamma(y)) = \exp(-\gamma|x-y|)$. *Then (see (Baldi-Antognini and Zagoraiou (2010)))* $\psi(\xi) = IMSPE = 1 - \frac{n-1}{\gamma} + 2A(\xi) + \frac{B(\xi)}{C(\xi)}$, *where* $A(\xi) = \sum_{i=1}^{n-1} a(d_i)$, $B(\xi) = \sum_{i=1}^{n-1} b(d_i)$, $C(\xi) = \sum_{i=1}^{n-1} c(d_i)$, *and* $a(d) = \frac{d}{\exp(2\gamma d)-1}$, $b(d) = d + \frac{\exp(\gamma d)-1}{\exp(\gamma d)+1}$, $c(d) = d + \frac{3(1-\exp(2\gamma d))+2\gamma d \exp(\gamma d)}{\gamma(\exp(\gamma d)+1)^2}$

And kernel has the form $\phi(\alpha, \xi) = (1 + \sum_{i=1}^{n-1} \frac{e^{\gamma d_i}-1}{e^{\gamma d_i}+1})^\alpha (\sum_{i=1}^{n-1} \frac{d_i^2(e^{2\gamma d_i}+1)}{(e^{2\gamma d_i}-1)^2})^{1-\alpha}$

Numerical experiments for this setup are given in Section 4.

Example 3 ***Inverse Problem***

Computing h from Lh = y is indirect or inverse problem. Indirect problems arise in numerous applications. The main difference to "classical" inference is that h cannot be computed directly. Such problems have been investigated intensively (most of the work focused on the construction of estimators of h and determination of convergence properties). In many areas the data are sampled using an uniform design. (Biedermann et al. (2011)) constructs an optimal design minimizing IMSE of indirect regression estimator, which is constructed by estimating the coefficients in the singular value decomposition of the corresponding operator. They discuss two regularization schemes (Tihhonoff and spectral cut-off regularization).

3 Fredholm operators and KWET

Under the suitable assumptions the integral operator L is a compact operator. The dimension of the null space of operator acting on h equals to the co–dimension of its range. The range is characterized as those ψ that are orthogonal to the null space of the transpose of the operator, namely with the kernel $K'(\alpha, x) = K(x, \alpha)$. This was proven by Fredholm for such operators, see e.g. (Lax (2002)).

Here we discuss the applicability of this approach for KWET and solvability of the appropriate Fredholm equations. We may have a look on a $\psi = Lh$ by considering various spaces of h.

We can employ Fredholm approach by using an approximative approaches from numerical analysis. In such a case the solution of Fredholm integral equations of the first kind is considered in terms of a linear combination of eigenfunctions of the kernel. Practical and theoretical difficulties appear when any corresponding eigenvalue is very small, and practical solutions are obtained which exclude the small eigensolutions and which are exact for a slightly perturbed integral equation. The problem is essentially ill posed, in the sense that there are many solutions which satisfy exactly an integral equation slightly perturbed from the original, and we might therefore seek a "smooth" solution rather than an exact solution. For more details on analytical properties see e.g. (Baker et al. (1964)). The practical approach in numerics is well done by (Hansen (1992)). We have found practically very important to choose a big weights for a large values of criterion $\psi(\xi)$.

Now let us formulate theoretical conditions for solvability of Fredholm equation in our setup. Let us consider $h \in H_R$ where H_R is either $L^2[0,1]$ or reproducing kernel Hilbert space with reproducing kernel $R(s,t), s,t \in [0,1]$. We assume that the kernel given by

$$Q(\xi_1,\xi_2) = \begin{cases} \int_0^1 \phi(\xi_1,\alpha)\phi(\xi_2,\alpha)d\alpha, & \text{if } H_R \text{ is } L^2[0,1], \\ \int_0^1\int_0^1 \phi(\xi_1,u)\phi(\xi_2,v)R(u,v)dudv & \text{otherwise,} \end{cases} \tag{6}$$

is continuous on $[0,1]^2$. Now we can formulate some results on solvability.

Lemma 1 *Let us consider Fredholm equation $Lh = \psi$. Then $\psi \in L(H_R)$ if and only if $\psi(\xi) = \sum_v a_v q_v(\xi)$ and*

$$|||\psi|||^2 := \sum_s \frac{\psi_s^2}{\lambda_v^2} < \infty \tag{7}$$

where $\psi_s = \int_{X^n} \psi(\xi)q_s(\xi)d\xi$ and $Q(\xi,\nu) = \sum_s \lambda_v q_s(\xi)q_s(\nu)$ is Mercer-Hilbert-Schmidt expansion (see (Riesz and Nagy (1955))) of kernel Q.

The next lemma formulates sufficient conditions in terms of kernel ϕ (i.e. D_α-criterion) for existence of solution to $\psi = Lh$.

Lemma 2 *Let kernel ϕ has expansion $\phi(\xi,\alpha) = \sum_i \mu_i u_i(\xi)v_i(\alpha)$, $\psi \in$ span $\{u_i\}$ and $|||\psi|||_\phi^2 := \sum_i \frac{\psi_i^2}{\mu_i^2} < \infty$, where $\psi_i = \int_{X^n} \psi(\xi)u_i(\xi)d\xi$. Then there exists solution of $\psi = Lh$*

Proof of lemma 2 We integrate expansion equality $\phi(\xi,\alpha) = \sum_i \mu_i u_i(\xi)v_i(\alpha)$, (this is possible since kernel $\phi \in L^2$) by $\int_0^1 ..v_i(\alpha)d\alpha$ Thus we obtain $\sum_i \mu_i(v_i,h)u_i(\xi) = \sum_i(u_i,\psi)u_i(\xi)$ From that we have $(v_i,h) = (u_i,\psi)/\mu_i; i = 1,..$ And thus

$$h(\alpha) = \sum_i \frac{(u_i,\psi)}{\mu_i} v_i(\alpha),$$

From which Picard condition is $|||\psi|||_\phi^2 < \infty$. □

4 Numerical experiments and Regularization

The problem of solving numerically a Fredholm equation of the first kind is complicated by the fact that the inversion operator is not in general continuous so that numerical stability is major problem. When observation is not given analytically, any error in these values can completely invalidate the solution. This we have observed in our numerical experiments, getting many eigenvalues close to 0 and other one being relatively huge ($\approx 10^5$). A common approach is to find a regularized solution by minimizing the functional $||Lh - \psi||^2 + \lambda\Omega(h)$ over a suitable set of functions. Here, Ω, the stabilizing functional, is a non-negative functional chosen to be small for h having desirable properties (typically smoothness). λ, the regularization parameter, is set to give an appropriate balance (trade-off) between the value of $\Omega(h)$ and the error $||Lh - \psi||^2$ (see e.g. (Tikhonov and Arsenin (1977)) for a seminal paper). Regularization with

$$\Omega(h) = \int \sum_{i=0}^{m} a_i |f^{(i)}|^2 \tag{8}$$

(so called Tikhonov regularization) has been extensively studied. (Wahba (1977)) developed the method called generalized cross validation (GCV) to estimate λ directly from the data.

The following numerical experiments (see Section 4) illustrate that we need regularization in our setup. One may wonder, why a need for regularization in discretized setup? As (Hansen (1992)) states, even that for problems with finite rank, Pickard condition (in Lemma 1,2) is always satisfied (from a purely mathematical point of view), discrete problems always suffer from some combination of measurement errors, discretization errors, and rounding errors and the solution is extremely sensitive to these errors. Beside this argument, we have some reasons which solution h suits well for our purposes (i.e.$h > 0$, etc.).

4.1 Construction of the minimization problem

In this section we reformulate the problem of equivalence of the two criteria of optimality as a quadratic optimization problem. Then we present numerical methods used to solve it.

Recall the equivalence conjecture: there exists such a function $h^\star(\alpha)$ that the following holds $\psi(\xi) \simeq \int\limits_0^1 \Phi(\xi|\alpha)h^\star(\alpha)\,d\alpha$. Here $\psi(\xi)$ denotes a criterion related to prediction and $\Phi(\xi|\alpha)$ denotes D_α-optimality criterion.

The problem is how to find such a function $h(\alpha)$, so that the two criteria are equivalent. Given $\psi(\xi)$ and $\Phi(\xi|\alpha)$, we are interested in minimization of the following function

$$f_\xi = \left[\psi(\xi) - \int\limits_0^1 \Phi(\xi|\alpha)h(\alpha)\,d\alpha\right]^2, \tag{9}$$

with respect to $h(\alpha)$. For the sake of simplicity, consider that $h(\alpha)$ is of the form h$(\alpha) = \sum_{i=1}^d w_i\delta_{\alpha_i}$, for some $\underline{\alpha} = (\alpha_1, \ldots, \alpha_d)^T$, $0 \le \alpha_i \le 1$. Then our problem can be written as

$$f = \sum_{j=1}^{q} f_{\xi_j} \quad = \quad \sum_{j=1}^{q} \left[\psi(\xi_j) - \sum_{i=1}^{d} w_i \Phi(\xi_j|\alpha_i)\right]^2,$$

$$= \sum_{j=1}^{q} \left[\psi(\xi_j) - \Phi^T(\xi_j|\underline{\alpha})\underline{w}\right]^2,$$

$$= \sum_{j=1}^{q} \psi^2(\xi_j) + \underline{w}^T \left(\sum_{j=1}^{q} \Phi(\xi_j|\underline{\alpha})\Phi^T(\xi_j|\underline{\alpha}) \right) \underline{w} - 2\underline{w}^T \sum_{j=1}^{q} \psi(\xi_j)\Phi(\xi_j|\underline{\alpha}).$$

Thus the problem of minimization of the function f is equivalent to minimization of the following function

$$\tilde{f} = \frac{1}{2}\, \underline{w}^T A \underline{w} - b^T \underline{w} + c, \tag{10}$$

where

$$A = \sum_{j=1}^{q} \Phi(\xi_j|\underline{\alpha})\Phi^T(\xi_j|\underline{\alpha}), \quad b = \sum_{j=1}^{q} \psi(\xi_j)\Phi(\xi_j|\underline{\alpha}), \quad c = \frac{1}{2}\sum_{j=1}^{q} \psi^2(\xi_j).$$

Hence, our aim is to find the optimum of a quadratic function $\tilde{f}$. The Hessian A of the objective function $\tilde{f}$ is a symmetric positive semi-definite matrix.

4.1.1 Constrained case

In the previous section we considered that there is no constraints on the function $h(\alpha)$. However, it can be useful to require $h(\alpha)$ to have some certain properties. For instance, we can impose the following constraint on $h(\alpha)$

$$\int_0^1 h(\alpha) d\alpha = 1.$$

In this case our problem can be formulated as

$$minimize\ \tilde{f}, such\ that\ \int_0^1 h(\alpha) d\alpha = 1. \tag{11}$$

It corresponds to a constrained quadratic optimization problem. Since we assumed that $h(\alpha)$ is in the form $h(\alpha) = \sum_{i=1}^{d} w_i \delta_{\alpha_i}$, the constraint here can be simply rewritten as $\sum_{i=1}^{d} w_i = 1$.

The usual trick when solving a constrained quadratic optimization problem is penalty approach. It consists in introducing some penalty term σ to the original function $\tilde{f}$ and then solving a corresponding (equivalent) unconstrained problem.

In our case first we rewrite the constraint $\sum_{i=1}^{d} w_i = 1$ in the form $y^T w = 1$, where $y = (1, \ldots, 1)^T$, and then we define $\mathrm{f}_\sigma(x) = \tilde{f} + \sigma(y^T w - 1)^2 = \tilde{f} + \sigma(w^T y y^T w - 2w^T y + 1)$,
$= \frac{1}{2} w^T A w - w^T b + c + \sigma(w^T y y^T w - 2w^T y + 1)$
$= \frac{1}{2} w^T (A + 2\sigma y y^T) w - w^T(b + 2\sigma y) + \sigma + c$. Now we can solve this unconstrained problem with some large penalty term σ instead of the original constrained problem (11).

4.1.2 Weighted case

We also can be interested in minimization of the weighted function (or function with the scaling factor a)

$$f_\xi^a = \left[\psi(\xi) + a - \int_0^1 \Phi(\xi|\alpha)h(\alpha)\,d\alpha\right]^2, \tag{12}$$

where a is some weight. First, for each fixed weight a we shall minimize $\sum_{i=1}^q f_{\xi_i}^a$ with respect to $h(\alpha)$, such that $\int_0^1 h(\alpha)d\alpha = 1$. Then we minimize the resulting function with respect to the scaling factor a. Thus, we have

$$\min_a \min_{h(\alpha):\int_0^1 h(\alpha)d\alpha=1} \sum_{i=1}^q f_{\xi_i}^a = \tag{13}$$

$$\min_a \left\{ \min_{\int_0^1 h(\alpha)d\alpha=1} \sum_{i=1}^q \left[\psi(\xi_i) + a - \int_0^1 \Phi(\xi_i|\alpha)h(\alpha)\,d\alpha\right]^2 \right\}. \tag{14}$$

For each fixed weight a, we minimize the following function

$\mathrm{f}_a = \sum_{i=1}^q f_{\xi_i}^a = \sum_{j=1}^q \left[\psi(\xi_j) + a - \sum_{i=1}^d w_i \Phi(\xi_j|\alpha_i)\right]^2$

$= \sum_{j=1}^q (\psi(\xi_j) + a)^2 + \underline{w}^T \left(\sum_{j=1}^q \Phi(\xi_j|\underline{\alpha})\Phi^T(\xi_j|\underline{\alpha})\right) \underline{w}$

$- 2\underline{w}^T \sum_{j=1}^q (\psi(\xi_j) + a)\Phi(\xi_j|\underline{\alpha})$,

subject to $\int_0^1 h(\alpha)d\alpha = 1$.

The problem of minimization of the function f_a is equivalent to minimization of the following function

$$\tilde{f} = \frac{1}{2}\,\underline{w}^T A \underline{w} - b^T \underline{w} + c, \tag{15}$$

where

$$A = \sum_{j=1}^q \Phi(\xi_j|\underline{\alpha})\Phi^T(\xi_j|\underline{\alpha}), \quad b = \sum_{j=1}^q (\psi(\xi_j) + a)\Phi(\xi_j|\underline{\alpha}), \quad c = \frac{1}{2}\sum_{j=1}^q (\psi(\xi_j) + a)^2.$$

To minimize with respect to a we use the MATLAB function `fminbnd` within the interval $[-10, 10]$. To minimize f_a for each fixed a we shall use some recent gradient methods described in the next section.

4.2 Optimization methods

In this section we describe numerical methods we use to solve the optimization problems described in the previous sections.

Gradient algorithms

Consider a problem of minimizing a quadratic function in the following form

$$f(x) = \frac{1}{2}x^T A x - x^T b, \tag{16}$$

where $x \in \mathbb{R}^d$ is an unknown vector, A is a $d \times d$ symmetric positive-definite matrix such that

$$0 < m = \inf_{(z,z)=1} (Az, z) < M = \sup_{(z,z)=1} (Az, z) < \infty$$

and b is a given vector in $\mathbb{R}^d$. The gradient of the function (16) at point x_k is $\nabla f(x_k) = g_k = Ax_k - b$. Solution to the optimization problem (16) is $x^* = A^{-1}b$. However, usually it is a very difficult computational task to compute the inverse matrix A^{-1}, especially when the dimension of the problem d is large. In this case some numerical methods can be used to determine the solution x^*. A big class of numerical optimization methods is the class of gradient algorithms.

The general gradient method for solving problem (16) corresponds to the following iterative process

$$x_{k+1} = x_k - \gamma_k g_k\,, \;\; k = 0, 1, 2 \ldots \tag{17}$$

where $x_0 \in \mathbb{R}^d$ is a starting vector and $\gamma_k > 0$, the step-size at iteration k, is determined by some rule. The iterations (17) can also be rewritten in terms of residuals (gradients) g_k as

$$g_{k+1} = g_k - \gamma_k A g_k\,, \tag{18}$$

with $g_0 = Ax_0 - b \in \mathbb{R}^d$, the initial residual vector.

There exist numerous algorithm of gradient-type, the most famous of them being the Steepest Descent (SD) method (also known as Gradient Descent). This method was initially introduced by Cauchy in 1847 for solving systems of linear equations. The step length for the SD algorithm is determined by

$$\gamma_k = \arg\min_{\gamma} f(x_k - \gamma \nabla f(x_k)), \tag{19}$$

and in the quadratic case is given by

$$\gamma_k = \frac{(g_k, g_k)}{(Ag_k, g_k)}. \tag{20}$$

However, nowadays it is regarded as an algorithm with poor rate of convergence. Many other gradient methods were derived from SD and can be viewed as its adaptation, for example, the Barzilai and Borwein (BB) method introduced by (Barzilai and J.M. Borwein (1988)). The BB step-length is exactly the step length of the standard Steepest Descent but at the previous iteration

$$\gamma_k = \frac{(g_{k-1}, g_{k-1})}{(Ag_{k-1}, g_{k-1})}. \tag{21}$$

This method has much better convergence rate than that of SD, but convergence to the solution is completely non-monotonic.

The actual speed of convergence of any gradient method depends on the condition number $\rho = M/m$, *i.e.*, the ratio of the largest and the smallest eigenvalues of the matrix A. Larger the condition number is, worse the rate of convergence is.

An important characteristic of a numerical algorithm is its computational cost and possibility to parallelize the computations. The above methods require one matrix-vector multiplication per iteration and calculation of two inner products.

This inner products create so-called synchronization points: all computations have to wait till these calculations are done.

An efficient gradient algorithm with pre-defined step sizes

In (Bukina et al. (2009)) several methods to compute the step length for the gradient algorithm (17) were proposed. These methods consist in constructing a sequence of step sizes before the run of the main algorithm. Thus, the same sequence can be used for different problems.

More precisely, for the gradient algorithm

$$x_{k+1} = x_k - \gamma_k g_k \,,$$

the sequence of step lengths $\gamma_k = 1/\beta_k$ is constructed in advance using some pre-defined sequence of numbers. First we construct these numbers and then we rescale these points to the interval defined by the smallest and the largest eigenvalues of the matrix A. To construct a sequence $\{\beta_k\}$ first we compute a sequence $\{z_k\}$ having the arcsine density in $[0,1]$. Then we scale $\{z_k\}$ from $[0,1]$ to $[m,M]$ using the formula

$$\beta_k = m + (M-m)z_k, \quad k = 0,1,\dots$$

where m and M are the leading eigenvalues of A or their approximations.

Sequence of $\{z_k\}$ is constructed so that it has arcsine distribution with density

$$p(t) = \frac{1}{\pi\sqrt{t(1-t)}}, \;\; 0 \le t \le 1 \,, \tag{22}$$

on $[0,1]$. The sequence $\{u_k\}$ defined via

$$u_i = \frac{1}{\pi}\arccos(2z_i - 1), for\ all\ i = 1,\dots,k,$$

uniformly distributed in $[0,1]$ when $\{z_k\}$ has the density (22). Figure 1 shows frequencies of the sequence $\{z_k\}_0^{399}$ generated using a dynamical system described below and the corresponding arcsine density on $[0,1]$.

One of the most natural algorithms for computing points $\{z_k\}$ with density (22) would be to generate independent random points with this density. This, however, does not lead to reliable algorithms: the associated sequence of rates becomes too erratic. The use of dynamical systems for generating $\{z_k\}$ leads to more stable and efficient algorithms. A good example of a dynamical system used to generate points with the asymptotic density (22) is presented below.

Sequence 1: **Symmetric fractional parts of Golden Ratio.**

For all $k \ge 0$, define

$$u_k = \{\,\{\ \varphi(k+1)\}, \text{for } k = 2j1 - \{\varphi(k+1)\}, \text{for } k = 2j+1.$$

where

$$\varphi = (\sqrt{5}-1)/2 \simeq 0.61803\dots$$

is the Golden ratio and $\{a\}$ denotes the fractional part of a. The sequence $\{z_k\}$ is then defined via

$$z_k = \frac{1}{2} + \frac{1}{2}\cos(\pi u_k), \;\; k = 0,1,\dots \tag{23}$$

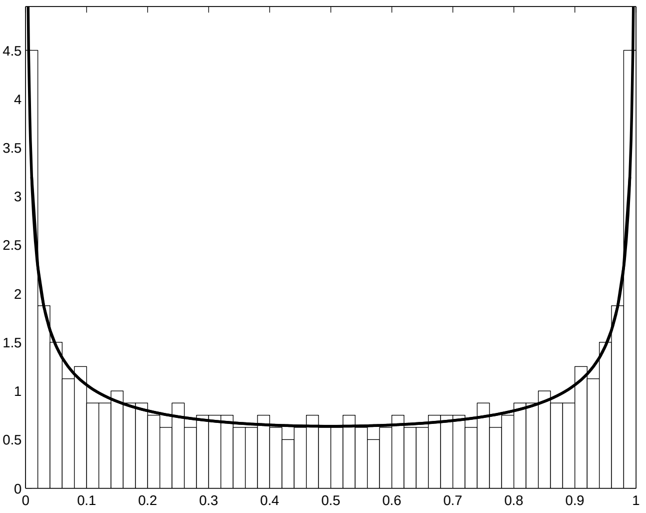

Figure 1: Frequencies of the sequence $\{z_k\}_0^{399}$ and the corresponding arcsine density on $[0, 1]$

Other sequences can be used to generate $\{z_k\}$: for instance, Chebyshev points (or Chebyshev nodes, the roots of Chebyshev polynomials of the fist kind), different logistic maps, etc. The advantage of the presented dynamical system is the fact we can construct as many step lengths as we want, we do not have to decide the number of iterations to perform beforehand.

Also ordering of the sequence $\{z_k\}$ is quite important: some sequences can lead to completely non-monotonic methods. Another advantage of this sequence is that the numbers come in symmetric pairs, so that we can use a largest z_k (and therefore the largest β_k) in each pair. For all $j \geq 0$, define $v_j = \{\varphi(j + 1)\}$, and $u_{2j} = \min\{v_j, 1 - v_j\}$, $u_{2j+1} = \max\{v_j, 1 - v_j\}$; the sequence $(z_k)_k$ is still defined using (23). We shall refer to the resulting sequence as Sequence 2. This special ordering of the step sizes results in more monotonic convergence of the associated gradient method.

The sequence $\{\beta_k\}$ is constructed from $\{z_k\}$ by scaling it to the interval $[m, M]$, where m and M are the extreme eigenvalues of A or their approximations. However, in most of the practical problems these eigenvalues are unknown, thus, they have to be estimated. In order to do so we shall define

$$\mu_k^\alpha = \frac{(A^\alpha g_k, g_k)}{(g_k, g_k)},$$

this value is called the moment of the order α. The first moment μ_k^1 turns out to be the Rayleigh quotient (also known as the Rayleigh-Ritz ratio), which is commonly used for the eigenvalue estimation. Also we have the following inequalities m $\leq \mu_1^{(k)} \leq \frac{\mu_2^{(k)}}{\mu_1^{(k)}} \leq \frac{\mu_3^{(k)}}{\mu_2^{(k)}} \leq \frac{\mu_4^{(k)}}{\mu_3^{(k)}} \leq \cdots \leq M$, see (Bukina et al. (2009)) for more details.

Moreover, we do not need to estimate the eigenvalues at every iteration: for each sequence there exist pre-specified set of iterations (called sequence of record moments) at which it is necessary to estimate. For any sequence $z_0, z_1, z_2, \ldots$, define the sequences of record moments $L_{\min} = \{L_{\min}(j)\}_{j=0}^\infty$ and $L_{\max} = \{L_{\max}(j)\}_{j=0}^\infty$ as follows: $L_{\min}(0) = L_{\max}(0) = 0$; $L_{\min}(j+1) = \min\{k > L_{\min}(j) : z_k < z_{L_{\min}(j)}\}$, $L_{\max}(j+1) = \min\{k > L_{\max}(j) : z_k > z_{L_{\max}(j)}\}$ for $j \geq 0$. The sequences of record moments related to sequence 2 are $L_{\min} = \{0, 1, 3, 5, 9, 15, \ldots\}$ and $L_{\max} = \{0, 2, 4, 8, 14, \ldots\}$ with $L_{\min}(j+1) = L_{\max}(j)+1$ for $j = 0, 1, \ldots$ See (Bukina et al. (2009)) for more information.

For our numerical tests we shall use Algorithm 3 presented in (Bukina et al. (2009)) and summarized in the scheme. The algorithm is initialized by using two iterations of Minimum Residues algorithm with $\gamma_k = (Ag_k, g_k)/(Ag_k, Ag_k)$, $k = 0, 1$, at which initial approximations to m and M, $\widehat{m}_{k+1}$ and $\widehat{M}_{k+1}$ respectively, are computed. To construct the sequence $\{z_k\}$ we use the symmetric fractional parts of Golden Ratio with largest z_k first in each symmetric pair (this sequence is denoted as Sequence 2 in (Bukina et al. (2009))).

Set $j = 0$ $k = 2, 3, \ldots$ 1. $\widehat{M}_k > \widehat{M}_{k-1}$ $\beta_k = \widehat{M}_k$ $\beta_k = \widehat{m}_k + (\widehat{M}_k - \widehat{m}_k)z_j$;

If $j \in L_{max}$, then add $k+1$ to I_{k+1}.

$j \leftarrow j+1$ 2. $x_{k+1} = x_k - \frac{1}{\beta_k} g_k$ 3. $g_{k+1} = Ax_{k+1} - b$

4. If $k \in I_k$ compute $\mu_1^{(k)} = \beta_k \left(1 - \frac{(g_k, g_{k+1})}{(g_k, g_k)}\right)$ and

$\frac{\mu_4^{(k-1)}}{\mu_3^{(k-1)}} = \beta_{k-1} + \beta_k \frac{(\beta_k(g_{k+1}-g_k)+\beta_{k-1}(g_{k-1}-g_k), g_{k+1}-g_k)}{(\beta_k(g_{k+1}-g_k)+\beta_{k-1}(g_{k-1}-g_k), g_{k-1}-g_k)}$,

update $\widehat{m}_{k+1} = \min\{\widehat{m}_k, \mu_1^{(k)}\}$ and $\widehat{M}_{k+1} = \max\{\widehat{M}_k, \mu_4^{(k-1)}/\mu_3^{(k-1)}\}$;

otherwise set $\widehat{m}_{k+1} = \widehat{m}_k$ and $\widehat{M}_{k+1} = \widehat{M}_k$.

This method is much faster than the standard gradient methods such as Steepest Descent and Minimal Residues and usually faster than the Barzilai-Borwein algorithm. It also exhibits monotonic behavior almost at every iteration. Moreover, this algorithm does not require computations with high accuracy and is advantageous in terms of computational cost: to perform k iterations, only $O(\log k)$ inner products need to be calculated versus $2k$ inner products for standard gradient methods. This algorithm can be easily parallelized: the synchronization is required only at $O(\log k)$ pre-specified iterations out of k iterations, as $k \to \infty$.

For more details on the construction of sequence of step lengths of this type and the behavior of associated gradient methods see (Bukina et al. (2009)).

5 Numerical experiments

In this chapter we present several numerical examples to illustrate the proposed approach.

5.1 Parameter setup

Using iterative gradient methods, we are trying to minimize the following objective function

$$\tilde{f} = \frac{1}{2}\, \underline{w}^T A \underline{w} - b^T \underline{w} + c, \tag{24}$$

where

$$A = \sum_{j=1}^{q} \Phi(\xi_j|\underline{\alpha})\Phi^T(\xi_j|\underline{\alpha}), \quad b = \sum_{j=1}^{q} \psi(\xi_j)\Phi(\xi_j|\underline{\alpha}), \quad c = \frac{1}{2}\sum_{j=1}^{q} \psi^2(\xi_j).$$

Recall that $\alpha = (\alpha_1, \ldots, \alpha_d)$, $0 \leq \alpha \leq 1$ and we are interested in a function $h(\alpha)$ which is assumed to be in the form $h(\alpha) = \sum_{i=1}^{d} w_i \delta_{\alpha_i}$. Here $\Phi(\xi|\alpha)$ denotes D_α-optimality criterion and $\psi(\xi)$ denotes a criterion related to prediction. For our numerical experiments we consider $\psi(\xi) = -IMSPE$.

When the criterion function for D_α-optimality criterion is $\Phi(\xi|\alpha) = |M_\beta(\xi,\gamma)|^\alpha |M_\gamma(\xi,\gamma)|^{1-\alpha}$, $0 \leq \alpha \leq 1$, the resulting matrix A is very ill-conditioned. This can lead to a slow convergence of gradient methods. The speed of convergence depends strongly on the spectrum of A, especially on the condition number - ratio of the smallest and the largest eigenvalues of the matrix. In this case the smallest eigenvalue m is very close to zero and the largest one M is usually of the form $c * 10^5$, where c is some positive constant. It is know that when the condition number is large, numerical methods converge slowly.

When using other criterion function $\Phi(\xi|\alpha) = \alpha \log |M_\beta(\xi,\gamma)| + (1-\alpha) \log |M_\gamma(\xi,\gamma)|$, $0 \leq \alpha \leq 1$, the condition number of A is still quite large, however, it is smaller than in the previous case. Thus, for our tests we have used log-based criterion function.

For the examples below we choose the dimension of the partition $\underline{\alpha}$, which is also the dimension of the problem, to be $d = 100$. Thus, matrix A is a symmetric $d \times d$ matrix with the largest eigenvalue M and the smallest eigenvalue m very close to zero. For the gradient algorithm the initial vector is $w_0 = (1, 1, \ldots, 1)$ and stopping criterion is $\|g\|_2 \leq tol$ with $tol = 10^{-4}$. For each example we plot the resulting function $h(\underline{\alpha})$, given by (4.1), and the two criteria as a function of $d_{12} = |\xi_2 - \xi_1|$.

We have used 2,3,4,5 and 6-point designs ξ in the following form $\xi_j = [0, \xi_2^j]$, $\xi_j = [0, \xi_2^j, 1]$, $\xi_j = [0, \frac{1}{2}\xi_2^j, \xi_2^j, 1]$, $\xi_j = [0, \frac{1}{3}\xi_2^j, \frac{2}{3}\xi_2^j, \xi_2^j, 1]$, $\xi_j = [0, \frac{1}{4}\xi_2^j, \frac{1}{2}\xi_2^j, \frac{3}{4}\xi_2^j, \xi_2^j, 0]$ with $1 \leq j \leq 200$, where ξ_2 is a random vector. In each example parameter of the covariance of the process r is equal to one.

5.2 Results: Unconstrained case

First, we do not impose any additional constraints on $h(\alpha)$, for the illustration see Figures 2 and 3. We plot the resulting values of w_i, $i = 1, \ldots, d$, against α, and the two criteria $\psi(\xi)$ and $D_\alpha(\xi)$. In the figures D_α denotes $\int_0^1 \Phi(\xi|\alpha) h(\alpha)\, d\alpha$, where as $h(\alpha)$ we use the optimal function that we have found. In this case, approximations are successful: in each example $D_\alpha(\xi)$ is very close to $\psi(\xi)$; the worst case corresponds to the 2-point design and is presented in Figure 2, however, even here the approximation is already quite good.

5.3 Results: Constrained case

Now we consider that the function $h(\alpha)$ is constrained so that

$$\int_0^1 h(\alpha) d\alpha = 1,$$

or simply $\sum_{i=1}^{d} w_i = 1$. In this case we apply the gradient method to minimize the function f_σ described in the previous chapter (see section 4.1.1 for more details), with the penalty term $\sigma = 10$. See Figure 4 for illustration.

In the constrained case, we were not able to construct reasonable approximations of $\psi(\xi)$. Perhaps, this happened because of the constraints imposed on $h(\alpha)$. For

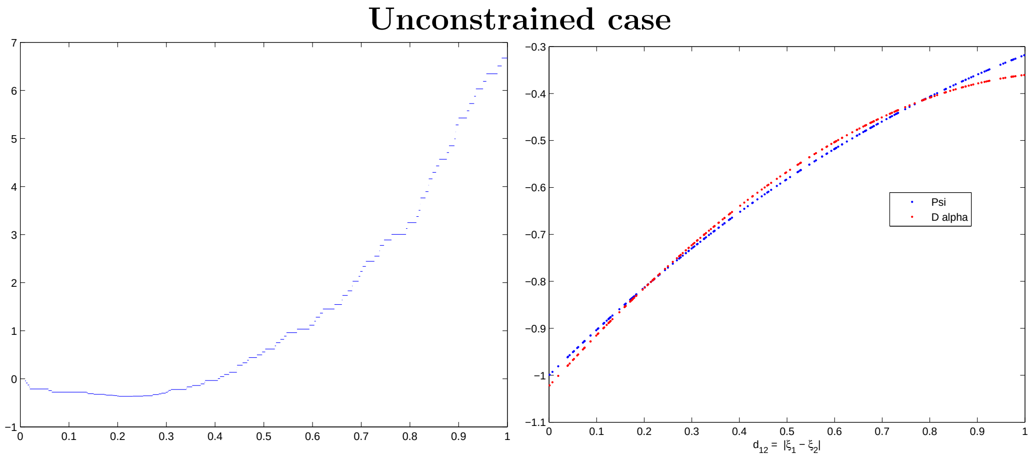

Figure 2: Values of w_i, $i = 1, \dots, d$, against $\alpha = (\alpha_1, \dots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for 2-point designs.

the optimal $h(\alpha)$ which satisfies the constraints and minimizes the function f_σ, the relative distance (the value of f_σ at the optimum $h(\alpha)$) is not small enough to provide a good fitting of the two criteria. This problem can be fixed by the introducing of a scaling factor a as described in section 4.1.2. This approach is implemented in the upcoming sections.

5.4 Results: Weighted and unconstrained case

In this section we present the results of the minimization of a function f_a with a scaling factor a, see section 4.1.2 for the explanation. First, we shall illustrate weighted and unconstrained case. It corresponds to the case described in the section 4.1.2 but when the constraints on $h(\alpha)$ are omitted. See Figures 5 and 6 for the results.

As in the unconstrained case we were able to obtain reasonable approximations of $\psi(\xi)$. Moreover, in this case minimization was also done with respect to the scaling factor a, thus we have slightly better results than in the simple case.

This is due to the more complex computations and an additional minimization with respect to the weight a. Of course, this minimization requires some extra calculations. For each fixed a the corresponding unconstrained problem is solved (from this point of view, our first example in the section 5.2 can be regarded as a unconstrained problem that corresponds to the scaling factor $a = 0$). Then amongst all tried weights, one weight a, for which f_a is taking the minimum value, is chosen. Thus, minimization with respect to a consists in sequential solution of unconstrained problems and then choosing the best of them. As a result, the computational time is increased but the precision of approximation is better.

It can be seen from the figures that for each optimal scaling factor (for the values of optimal a's see next section), the two criteria $\psi(\xi)$ and $D_\alpha(\xi)$ are very close to each other.

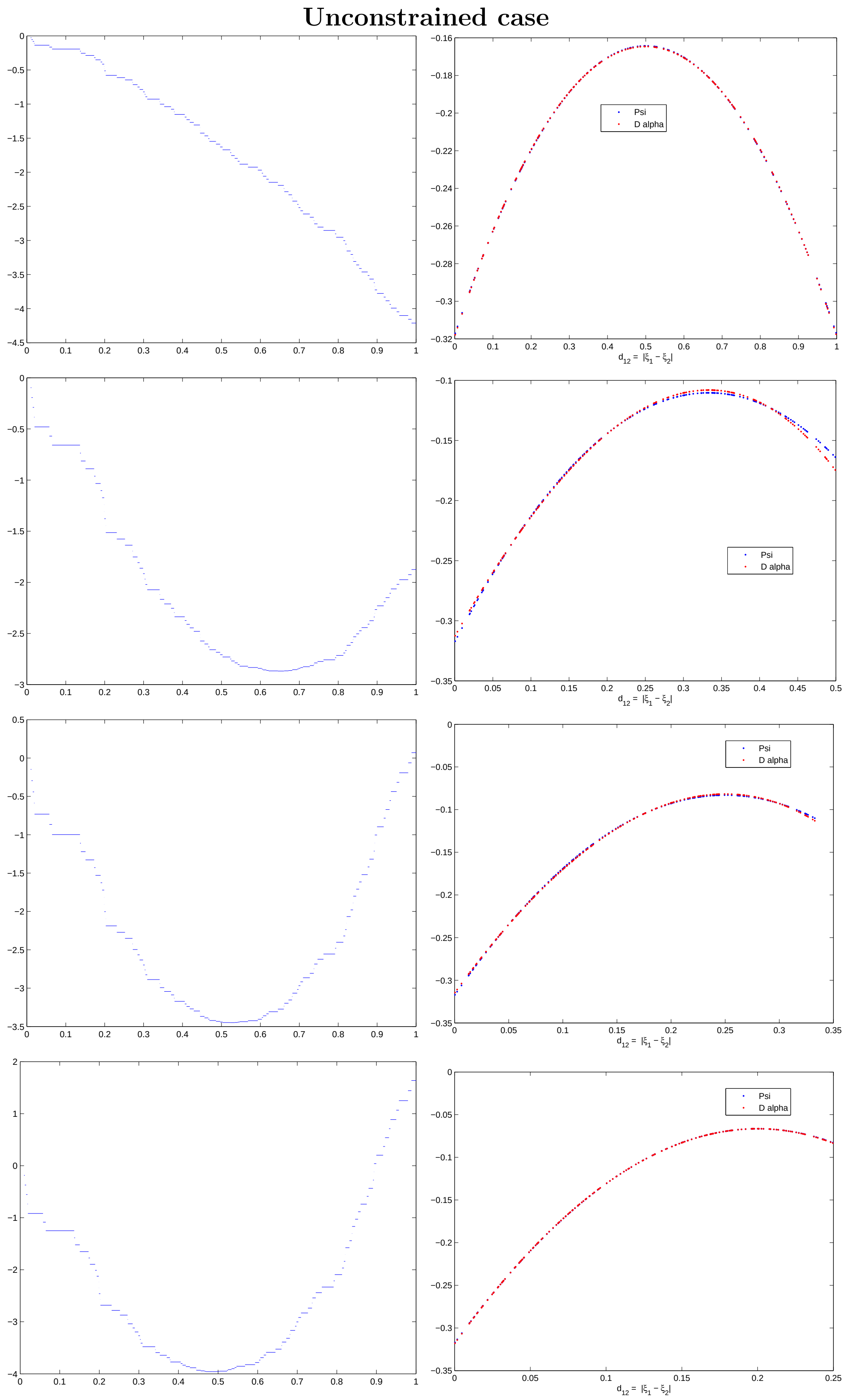

Figure 3: Values of w_i, $i = 1, \ldots, d$, against $\alpha = (\alpha_1, \ldots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for, from top to bottom, 3,4,5 and 6-point designs.

Constrained case

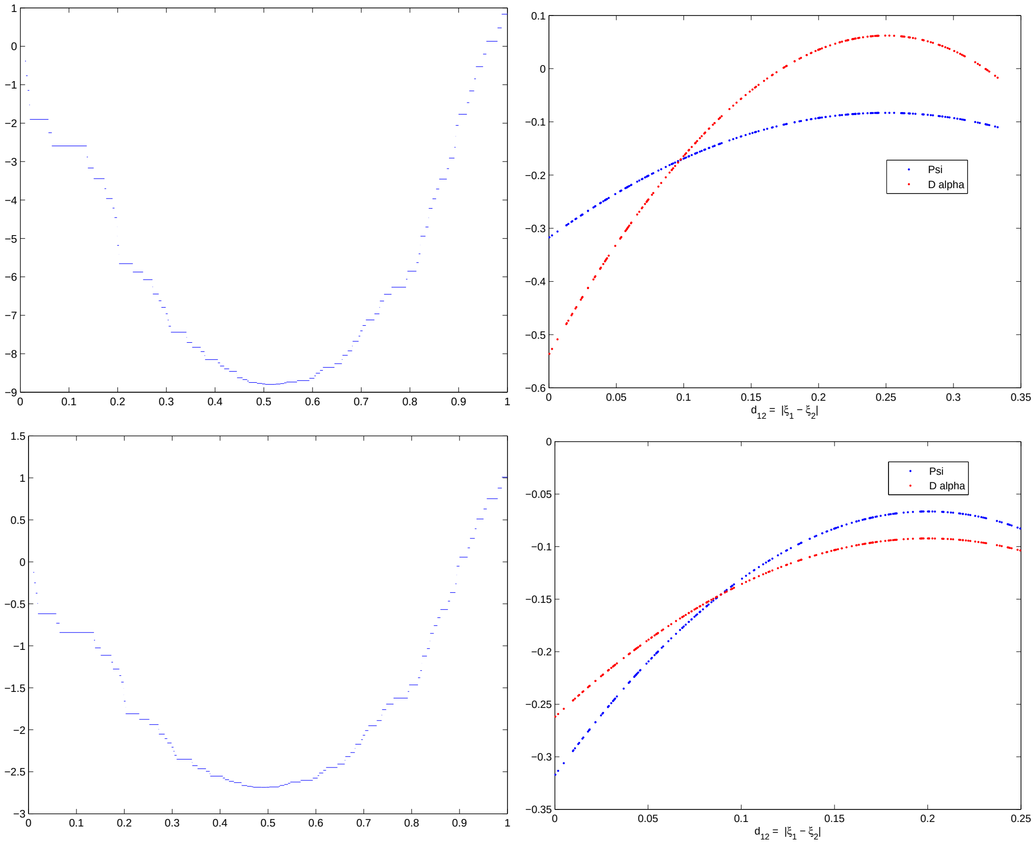

Figure 4: Values of w_i, $i = 1, \ldots, d$, against $\alpha = (\alpha_1, \ldots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for, from top to bottom, 3,4,5 and 6-point designs.

Weighted case

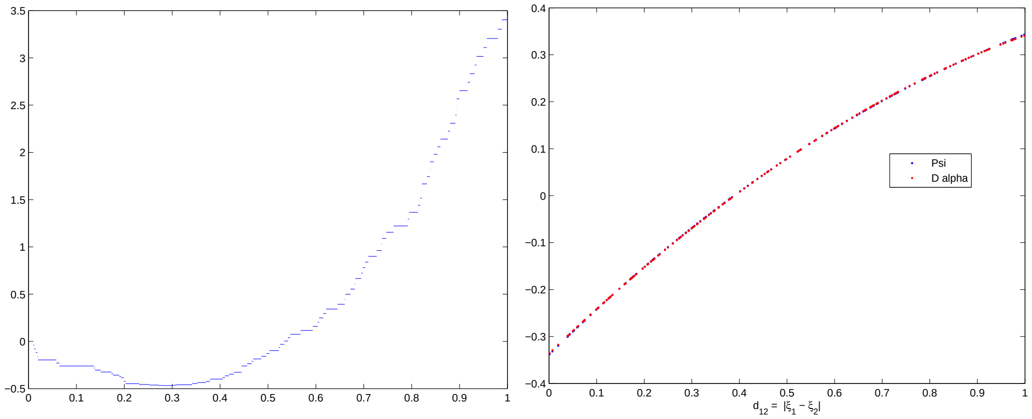

Figure 5: Values of w_i, $i = 1, \ldots, d$, against $\alpha = (\alpha_1, \ldots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for 2-point designs.

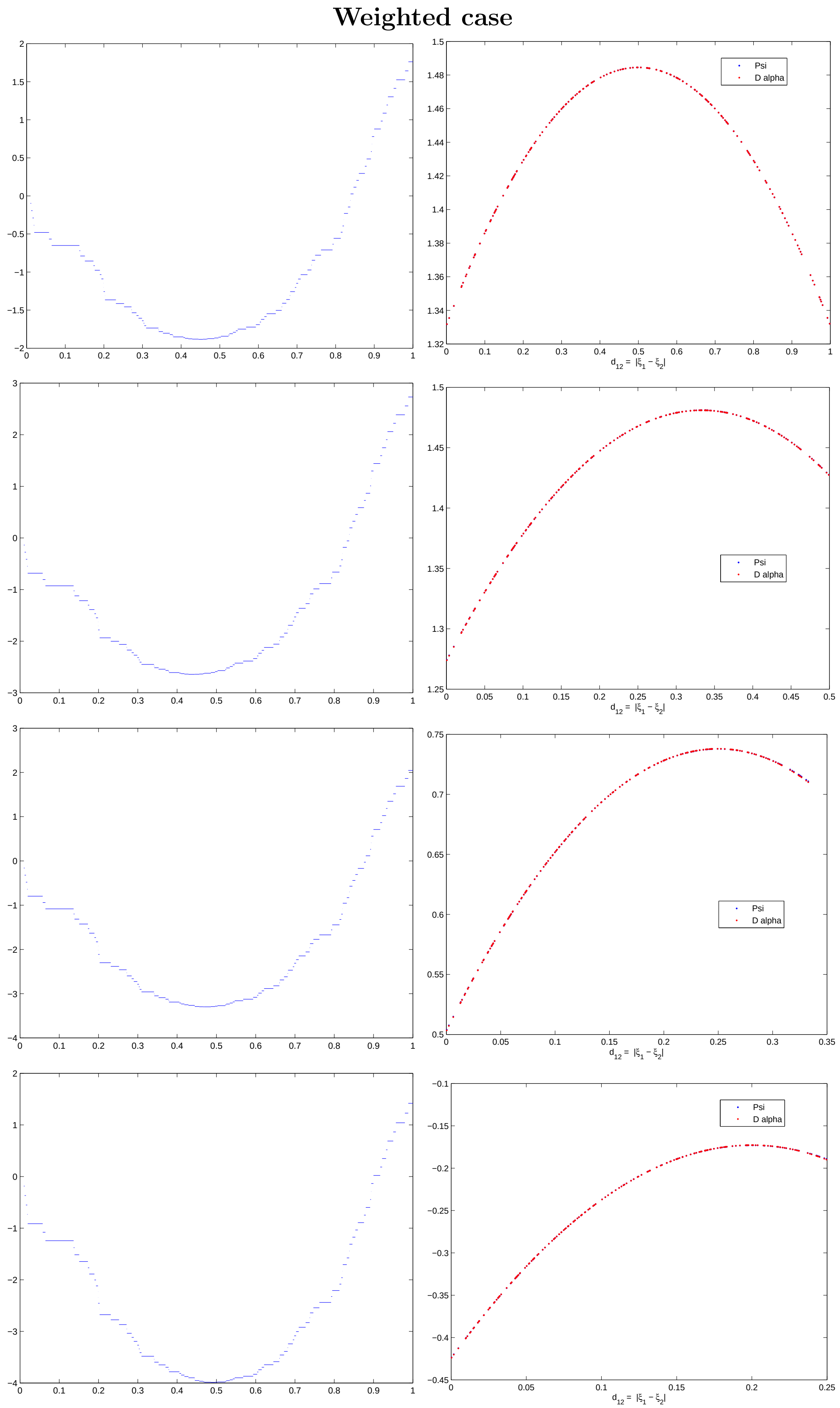

Figure 6: Values of w_i, $i = 1, \ldots, d$, against $\alpha = (\alpha_1, \ldots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for, from top to bottom, 3,4,5 and 6-point designs.

5.5 Results: Weighted and constrained case

Let us now illustrate the weighted and constrained case, see Figures 7 and 8.

Unlike in the simple constrained case, we were able to construct reasonable approximations when some weight was added to the objective function. Thus, our experiments showed that scaling factor $a = 0$ is not always optimal. Table 1 summarizes the optimal values of scaling factors for the different cases that we have considered.

Thus, introduction of the weight allowed us to approximate $\psi(\xi)$ even in the case when some constraints on $h(\alpha)$ were imposed.

Designs ξ	Unconstrained case	Constrained case
2-point designs	0.6612	1.1438
3-point designs	1.6488	1.4385
4-point designs	1.5914	0.9933
5-point designs	0.8209	0.3862
6-point designs	-0.1064	-0.3097

Table 1: Values of the optimal scaling factor a

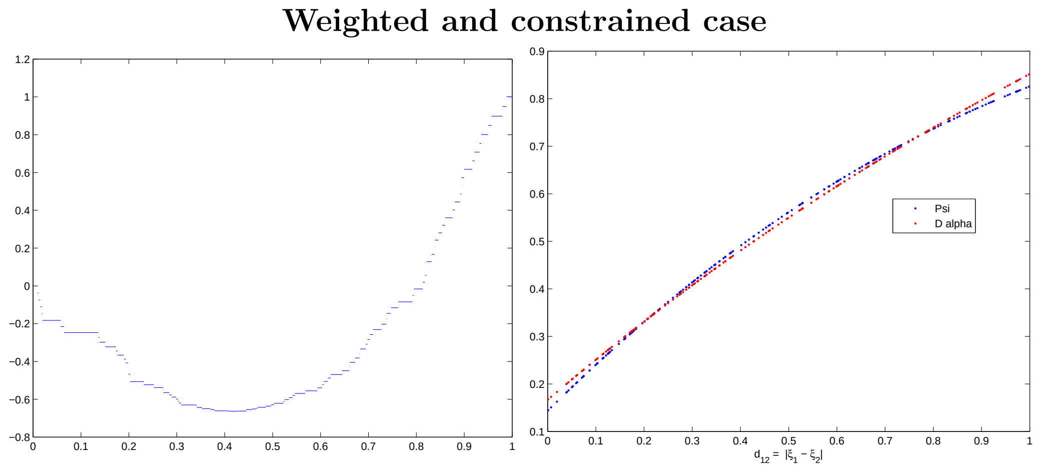

Figure 7: Values of w_i, $i = 1, \ldots, d$, against $\alpha = (\alpha_1, \ldots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for 2-point designs.

The value of the minimized function at the optimum can be treated as a relative distance between the two criteria, see Table 2. Smaller this distance is better the approximation.

5.6 Results:Degenerated case

In all the previous examples the partition $\alpha = (\alpha_1, \ldots, \alpha_d)$ was taking as a random vector. However, even if we take a partition α in a very simple form, say $\alpha = (0, \ldots, 0, 1)$, we still can obtain reasonable approximation of the criterion $\psi(\xi) = -IMSPE$. One of the possible explanations for this phenomenon is the relative

Weighted and constrained case

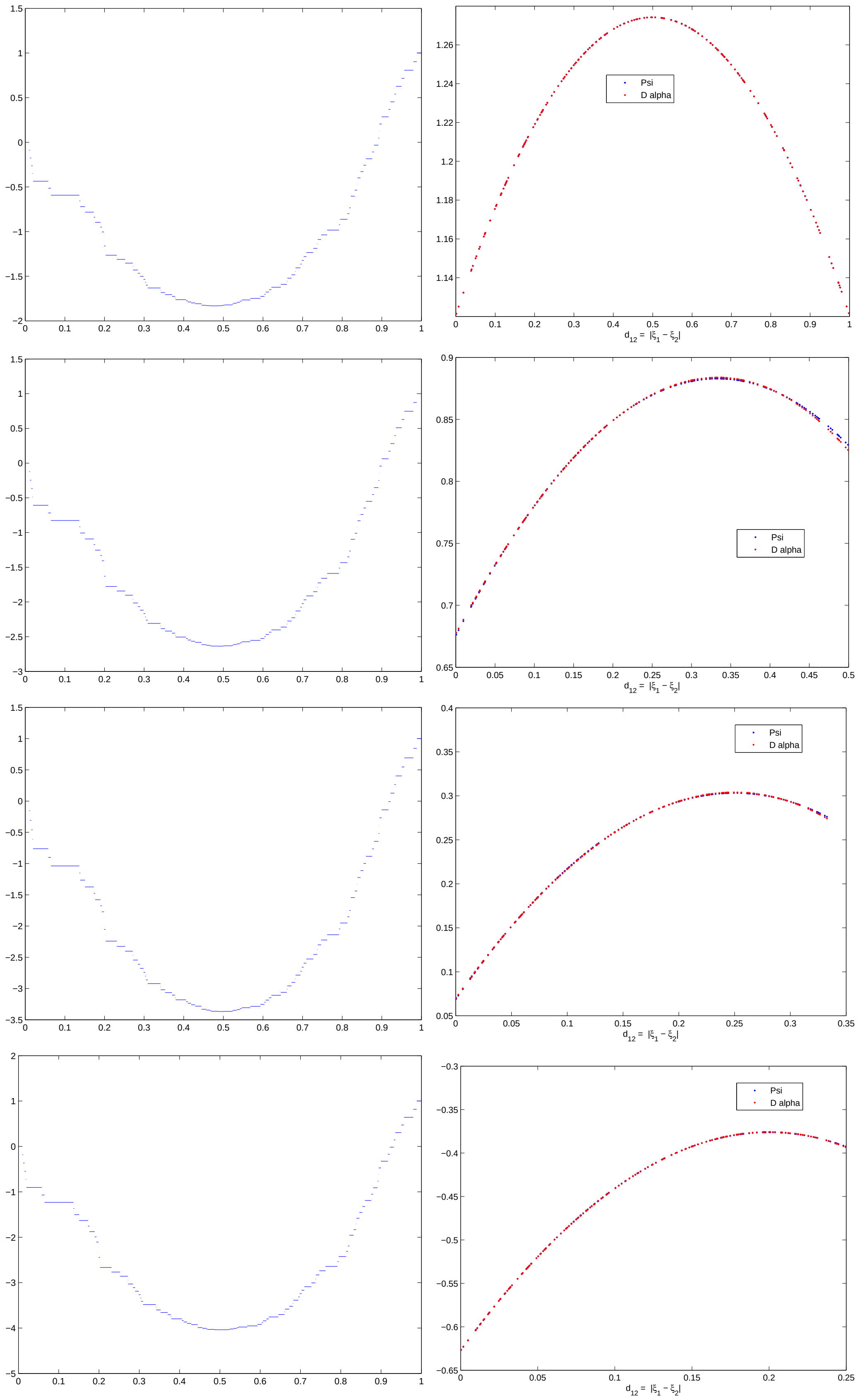

Figure 8: Values of w_i, $i = 1, \dots, d$, against $\alpha = (\alpha_1, \dots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for, from top to bottom, 3,4,5 and 6-point designs.

Designs ξ	No constraints	Weight	Constraints and Weight
2-point designs	0.0377	1.5450e-004	0.02033
3-point designs	1.2204e-005	1.2750e-008	2.1008e-007
4-point designs	0.0013	1.3660e-006	1.8261e-004
5-point designs	2.0676e-004	4.7275e-006	6.1369e-005
6-point designs	6.8409e-006	3.5032e-006	1.5682e-005

Table 2: Distances between the two criteria

simplicity of the chosen criterion $IMSPE$. Perhaps, when using more sophisticated criterion, say $MSPE$, a simple α will not be reliable. This aspect is left for further investigation.

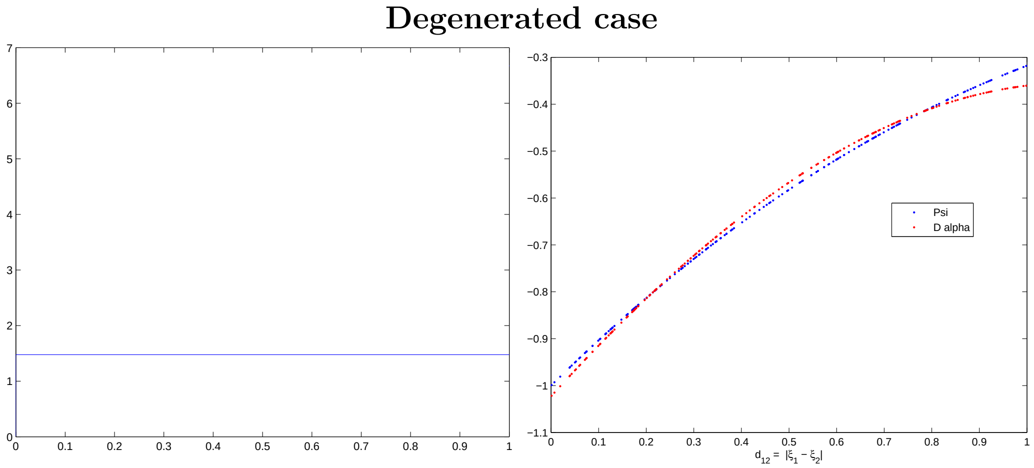

Figure 9: Values of w_i, $i = 1, \ldots, d$, against $\alpha = (\alpha_1, \ldots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for 2-point designs.

6 Conclusions and Discussion

In this paper we suggested the operator approximation for learning about relationship of IMSPE and compound D_α criterion introduced in (Müller and Stehlík (2010)). First we introduce integral relation between IMSPE and compounding kernel. We discuss a theoretical conditions for existence of solution for obtained Fredholm equation of the 1st kind. Then we provide numerical implementation and construct $h(\underline{\alpha})$. Several numerical examples illustrates the methods provided.

In general, it is a difficult task to make a link between two optimality criteria: one based on prediction, another based on parameters estimation, especially in the case of correlated processes. In the uncorrelated case it was done by (Kiefer and Wolfowitz (1960)). The correlated case, however, is not fully investigated. The first attempt to relate two criteria was done in (Müller and Pronzato (2009)). In this paper we concentrate on the correlated case and reformulate the conjecture of equivalence in an integral form.

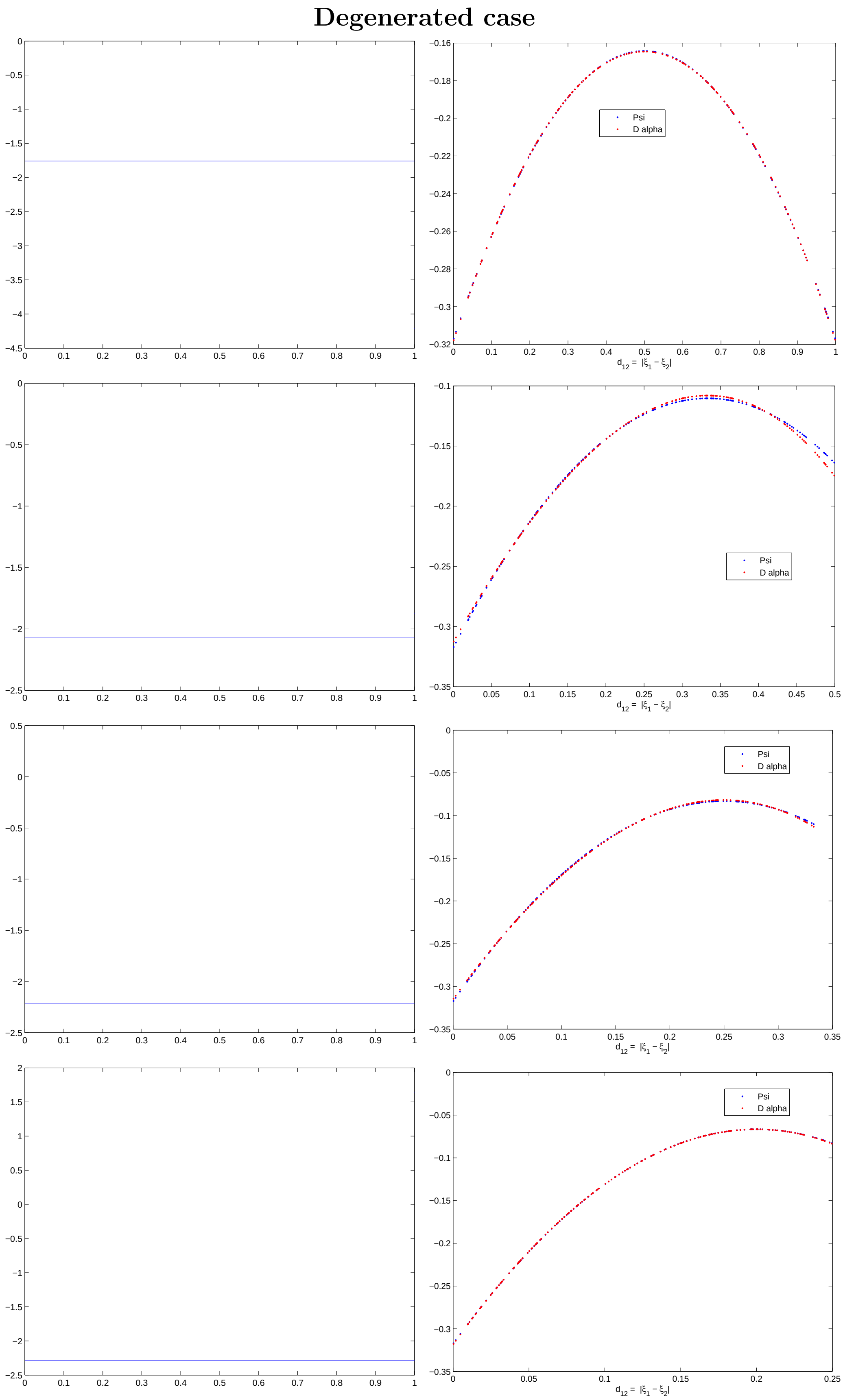

Figure 10: Values of w_i, $i = 1, \ldots, d$, against $\alpha = (\alpha_1, \ldots, \alpha_d)$ (left) and $\psi(\xi)$ and $D_\alpha(\xi)$ as a function of $d_{12} = |\xi_2 - \xi_1|$ (right) for, from top to bottom, 3,4,5 and 6-point designs.

Initially, we reformulated the problem of equivalence as a quadratic optimization problem. Then we solve this problem by using new gradient methods for quadratic optimization. These methods are advantageous for several reasons: they are much faster than standard gradient methods, have small computational cost, easily parallelized and monotonic at almost every iteration. We tried the proposed approach on several test problems and we present the obtained results in the paper. In terms of required computational effort it is less expensive to generate parameters estimation optimal designs rather than prediction optimal designs. Some particular structure of the relation between two criteria would allow to substitute a costly computation by much less costly one.

In this paper we concentrated on one dimensional case and as a prediction criterion we used IMSPE, leaving MPSE criterion for the future work. The approach also can be modified for the 2D case.

Also a thorough study of other regularizations will be of interest. As we have seen in section 4 we need regularization of $Lh = \psi$, as it is usual for Fredholm equations of the first kind. We have regularized by $||h|| = 1, h > 0$. This is a natural form of regularization, since we have h in the form of density. Regularization by maximum entropy ((Amato and Hughes (1991))) and regularization by reproducing kernel Hilbert space norm, (see e.g. (Wahba (1977))) would be worth further investigation.

In the present paper we considered the problem of solving $Lh = \psi$. The problem of solving (in h) $\max Lh = \max \psi$ is also of interest. That means, we search for h such that Lh and ψ have the same maximum. One idea is to employ the continuity property of L^p norm, i.e. $\lim_{p\to\infty} ||x||_p = ||x||_\infty = \max x(t)$. Thus we may solve $\lim_{p\to\infty} ||Lh||_p = \lim_{p\to\infty} ||\psi||_p$, hoping, that under certain regularities, we may approximate it by $||Lh||_p = ||\psi||_p$, for particularly large $p > 1$. For $||\psi||_p$ we may use MC strategies to evaluate this norm. For $||Lh||_p$ we may use that fact that $||Lh||_p^p = \int_0^1 ||\phi(., \alpha)||_p^p h(\alpha) d\alpha$. This will be worth further research.

Acknowledgements The authors are thankful for helpful comments of Werner G. Müller, Luc Pronzato and Joao Rendas. This work is supported by Amadee, Project Nr. FR 11/2010.

References

U. Amato and W. Hughes (1991), Maximum entropy regularization of Fredholm integral equations of the first kind, Inverse Problems 7 :793-808.

Baker et al. (1964), Numerical solution of Fredholm integral equations of first kind, The Computer Journal,1964; 7: 141-148.

A. Baldi-Antognini and M. Zagoraiou (2010), Exact optimal designs for computer experiments via Kriging metamodelling, Journal of Statistical Planning and Inference 140: 2607-2617.

S. Biedermann, N. Bissantz, H. Dette and E. Jones, Optimal Designs for Indirect Regression, Inverse Problems 27, doi:10.1088/0266-5611/27/10/105003.

J. Barzilai and J.M. Borwein (1988), Two-point step size gradient methods, IMA Journal of Numerical Analysis, 8: 141-148.

E. Bukina, L. Pronzato and A. Zhigljavsky, Gradient algorithms for solving linear equations with fast convergence rates, SIAM Journal on Numerical Analysis. Submitted,

S. B. Crary (2002), Design of Computer Experiments for Metamodel Generation, Analog Integrated Circuits and Signal Processing, 32, 7-16.

C. W. Groetsch (1984), The theory of Tikhonov regularization for Fredholm equations of the first kind, Boston, Pitman (section 1.3).

P. Ch. Hansen (1992), Numerical tools for analysis and solution of Fredholm integral equations of the first kind, Inverse Problems 8, 849-872.

J. Kiefer and J. Wolfowitz (1960), The equivalence of two extremum problems, Canad. J. Math., 12: 363-366.

J. Kiseľák and M. Stehlík (2008), Equidistant D-optimal designs for parameters of Ornstein-Uhlenbeck process, Statistics and Probability Letters 78, 1388-1396.

P. D. Lax, Functional analysis, Wiley-Interscience.

McGree, JM, Eccleston, JA and Duffull, SB. (1988). Compound optimal design criteria for nonlinear models, Journal Of Biopharmaceutical Statistics, 18(4): 646-661.

W. G. Müller and L. Pronzato, Towards an optimal design equivalence theorem for random fields, Ifas report Nr. 45.

W. G. Müller and M. Stehlík (2009), Issues in the Optimal Design of Computer Simulation Experiments. Applied Stochastic Models in Business and Industry, 25:163-177.

W. G. Müller and M. Stehlík (2010), Compound optimal spatial designs, Environmetrics, 21: 354-364.

F. Riesz and B. Sz.-Nagy (1955), Functional Analysis, Frederik Ungar, NY

J. Sacks, S. B. Schiller and W. J. Welch (1989), Design for Computer Experiments, Technometrics, 31, 1: 41-47.

A. N. Tikhonov and V. Y. Arsenin (1977), Solutions of Ill-pared Problems (New York: Wiley).

G. Wahba (1977) Practical approximate solutions to linear operalor equations when the data are noisy, SIAM J. Numm Anal 14: 651-67.

M. Zagoraiou and A. Baldi-Antognini (2009), Optimal designs for parameter estimation of the Ornstein-Uhlenbeck process, Appl. Stochastic Models Bus. Ind. 25:583-600.