Int. J. Appl. Math. Comput. Sci., 2020, Vol. 30, No. 4, 659–677
DOI: 10.34768/amcs-2020-0049

CONSTRUCTION OF CONSTRAINED EXPERIMENTAL DESIGNS ON FINITE SPACES FOR A MODIFIED E_K–OPTIMALITY CRITERION

DARIUSZ UCIŃSKI [a]

[a]Institute of Control and Computation Engineering
University of Zielona Góra
ul. Szafrana 2, 65-516 Zielona Góra, Poland
e-mail: d.ucinski@issi.uz.zgora.pl

A simple computational algorithm is proposed for minimizing sums of largest eigenvalues of the matrix inverse over the set of all convex combinations of a finite number of nonnegative definite matrices subject to additional box constraints on the weights of those combinations. Such problems arise when experimental designs aiming at minimizing sums of largest asymptotic variances of the least-squares estimators are sought and the design region consists of finitely many support points, subject to the additional constraints that the corresponding design weights are to remain within certain limits. The underlying idea is to apply the method of outer approximations for solving the associated convex semi-infinite programming problem, which reduces to solving a sequence of finite min-max problems. A key novelty here is that solutions to the latter are found using generalized simplicial decomposition, which is a recent extension of the classical simplicial decomposition to nondifferentiable optimization. Thereby, the dimensionality of the design problem is drastically reduced. The use of the algorithm is illustrated by an example involving optimal sensor node activation in a large sensor network collecting measurements for parameter estimation of a spatiotemporal process.

Keywords: constrained optimum experimental design, minimal sum of largest eigenvalues, generalized simplicial decomposition, optimal measurement selection.

1. Introduction

A fundamental question in optimum experimental design is how to allocate measurement resources in regression problems so that the experiment be maximally informative. Mathematically, this amounts to minimizing or maximizing some clearly defined real-valued function which quantifies the 'goodness' of the experiment. Most often, a suitable criterion is defined on the Fisher information matrix (FIM) being the inverse covariance matrix of the parameter estimates to be obtained from the data collected. A strength of this formulation is that the relevant design criterion can be extremized to provide an optimal allocation.

In the modern theory of regression design, feasible allocations are identified with probability measures on the design region, called continuous designs, or simply designs (Atkinson *et al.*, 2007; Fedorov and Leonov, 2014; Pronzato and Pàzman, 2013; Melas, 2006). In this way, the design problem is converted to that of extremizing a functional which depends on measures. Based on the appropriate equivalence theorems which constitute necessary and sufficient optimality conditions and are central in the theory of optimal design, it is sometimes possible to determine analytically an optimal measure. For general systems, however, it is usually the case that some iterative design procedure is required, for which equivalence theorems serve as methods for checking the optimality of any continuous design.

A standard sequential vertex-direction design algorithm embodies the idea of the following general feasible-direction method which is commonly used in nonlinear programming (Bertsekas, 1999): The current approximation $\xi^{(k)}$ to the optimal design $\xi^\star$ is updated by forming its convex combination with a (Dirac) measure δ_x putting unit mass at a single point x chosen in the design region so that the directional derivative of the design criterion at $\xi^{(k)}$ in the direction of δ_x is negative (resp. positive) if the design criterion is to be minimized (resp. maximized). A straightforward approach consisting in using an iterative nonlinear

optimization routine to solve for the optimal measurement locations and respective allocation masses in the resulting design is rather questionable, since the number of measurement locations is not known initially. Thus, efficient specialized algorithms have been invented which exploit the specific problem structure (see e.g., Pronzato and Zhigljavsky, 2014; Atkinson *et al*., 2007; Pronzato and Pàzman, 2013; Fedorov and Leonov, 2014). They produce a sequence of measures whose accumulation point solves the design problem.

An alternative strategy is to cover the design region with a suitable network, $\mathcal{N}$, of points which should be rich enough to contain close approximations to the points likely to have positive mass in the optimal design, and to focus solely on optimizing the masses associated with elements in $\mathcal{N}$. Such a formulation turns out to be extremely convenient since we deal with minimization (resp. maximization) of a convex (resp. concave) function over a nice constraint set, a canonical simplex $\mathcal{S}_n$, and this can be tackled by standard numerical methods, e.g., the gradient projection method or the conditional gradient method (Botkin and Stoer, 2005; Wu, 1978). (Note that the projection operation on $\mathcal{S}_n$ is almost as simple as a closed-form solution (cf. Maculan *et al*., 2003).) The vertex-direction method can also be used here, but it is usually extremely slow. That is why its various improvements were proposed, e.g., the vertex-exchange method (Böhning, 1986).

A highly competitive idea for a moderate cardinality of $\mathcal{N}$ is to employ extremely powerful algorithms for convex optimization based on semidefinite programming. Results of their successful application to D-, E- and E-optimum designs were reported by Joshi and Boyd (2009), Chepuri and Leus (2015) and Lu and Pong (2013). These methods have polynomial worst-case complexity and perform well in practice, rapidly computing the global optima with nonheuristic stopping criteria using interior-point algorithms.

For some criteria, however, even more specialized algorithms can be invented, which exploit specific problem structures. The most representative example is the extremely simple multiplication algorithm devised and analysed by Silvey *et al*. (1978), Pázman (1986), and Yu (2010). An interesting feature of this scheme is that it is globally convergent and that the successive values of the design criterion form a monotonic sequence. Although slow convergence is sometimes reported, some neat methods of removing nonoptimal support points can substantially accelerate it (Pronzato, 2003; Harman and Pronzato, 2007). What is more, the ease with which we can implement it constitutes a decided advantage. Yu (2011) combined vertex-direction, vertex-exchange and multiplicative algorithms for D-optimality in the so-called cocktail algorithm, which increases the speed while preserving convergence.

Recently, Harman *et al*. (2020) set forth a very effective randomized exchange algorithm (REX) which can be interpreted as a combination of both the vertex-exchange algorithm and the KL exchange algorithm commonly used to determine exact designs.

Some studies have been undertaken in order to extend the appealing framework of the design on the finite design space $\mathcal{N}$ to more complex settings which are encountered in applications. Thus, equality constraints on the design weights are sometimes considered, as discussed by Torsney and Mandal (2001). In practice, however, various inequality constraints must more frequently be considered, which are due to cost limitations, required design measure space restrictions for achieving certain robustness properties, or restrictions on the experimental space. The incorporation of additional linear constraints on the weights is necessary if various resource constraints have to be taken into account. For example, when measurements at different points incur different costs, an important requirement might be that the total cost of the experiment must not exceed a given budget. The constrained setting also appears naturally in relaxed formulations of the sensor selection problems, which consists of selecting a given number of gauged sites from among a much larger number of candidate ones (Joshi and Boyd, 2009). Although much work has been done in this respect as regards theory (e.g., Cook and Fedorov, 1995), the number of publications on the algorithmic aspects of constrained optimization of experimental design is still very limited.

Harman and Benková (2017) proposed a nontrivial and efficient barycentric algorithm, which draws on the idea of the multiplicative algorithm and is specialized in two linear inequality constraints (the size and cost constraints) on the weights. Problems with larger numbers of linear equality/inequality constraints can be treated by employing either interior-point methods (Joshi and Boyd, 2009; Chepuri and Leus, 2015; Lu and Pong, 2013), most often using existing SDP solvers, or the simplicial decomposition (SD), an inner-linearization polyhedral approximation method (Bertsekas, 2015). The multi-aspect work by Esteban-Bravo *et al*. (2017), where Newton-type methods were exploited to attack problems subject to nonlinear constraints, is truly representative of the former. The latter turns out to be extremely valuable for large cardinalities of $\mathcal{N}$, as it drastically reduces the problem dimensionality, without even saying about the striking simplicity of its implementation (it alternates between solving an LP problem and extremizing the original design criteria over the convex hull of a finite number of given nonnegative matrices, which can easily be solved using vertex-direction, vertex-exchange or multiplicative algorithms discussed previously), cf. (Uciński and Patan, 2007; Uciński, 2015).

In addition to that, a very auspicious, versatile

and efficient technique exploiting second-order cone programming for D-, A-, G- and I-optimality criteria was put forward by Sagnol and Harman (2015), who substantially refined the technically sound approach outlined by Sagnol (2011) for c-optimality. Finally, recent research reported by Duarte *et al*. (2020) points to a lot of potential for handling various intricate constraints presented by mixed integer nonlinear programming (MINLP) formulations when supported by quite efficient state-of-the-art numerical solvers.

Nevertheless, algorithmic construction of optimum designs on finite design spaces has not reached the maturity of its unconstrained counterpart yet and still waits for effective and efficient algorithms with good performance for large numbers (sometimes many thousands) of candidate support points. These algorithms should also cover optimality criteria which are not as well behaved as common ones, e.g., those of D- and A-optimality. Specifically, the communications on nonsmooth criteria are limited.

An important nonsmooth optimality criterion is the minimum eigenvalue of the FIM, called the E-optimality criterion. Maximizing it, we minimize the length of the largest axis of the uncertainty ellipsoid for the estimates. This criterion turns out to be nondifferentiable when the minimum eigenvalue of the FIM is multiple. It is easy to show (Joshi and Boyd, 2009) that the corresponding relaxed problem can be reformulated in terms of an LMI-constrained convex minimization problem, but when the cardinality of $\mathcal{N}$ grows, solving SDP problems becomes computationally demanding and interior-point solvers quickly run into time and memory issues on mediocre computers.

Instead, Pronzato and Pàzman (2013, p. 236) made use of an equivalent formulation in terms of a semi-infinite programming (SIP) problem and then the method of outer approximations of Shimizu and Aiyoshi (1980) was applied, which resulted in an extremely simple computational scheme alternating between determining the minimal eigenvalue of the current FIM and solving an LP problem. This simplicity makes the algorithm well suited for large cardinalities of $\mathcal{N}$.

The above LP-based approach to E-optimality was then extended by Burclová and Pázman (2016) for the case of the E_k-optimality criterion (the sum of k smallest eigenvalues of the FIM). This criterion, put forward and thoroughly analyzed by Harman (2004), is a generalization of both E-optimality (it corresponds to $k = 1$) and trace optimality (when k equals the number of estimated parameters, the E_k-optimality criterion is just the trace of the FIM). Its distinguishing feature is that the minimal efficiency of a design over the class of all orthogonally invariant design criteria coincides with its minimal efficiency over the finite class of all E_k-optimality criteria. Recall that for orthogonally invariant criteria the quality of a design depends solely on the shape of the corresponding confidence ellipsoid for the estimates and not on its orthogonal rotations, and that this class includes most design criteria used in practice. Thus, the search for a design maximum-efficient, i.e., the most efficiency stable, for all orthogonally invariant criteria, can be replaced by a drastically simpler exploration of the maximum efficient designs among the E_k-optimality criteria with k ranging from 1 to the number of estimated parameters.

Apart from application of the E_k-optimality criteria as accessory criteria when constructing criterion-robust designs, practitioners may be tempted by using the E_k-optimality criterion for an arbitrarily fixed $k > 1$. A rationale could be its dependence on more than a single eigenvalue of the FIM, accompanied by the simplicity of the LP-based method of its numerical construction. But in this setting utmost care should be taken to avoid the potential pitfalls. Specifically, a major drawback of the E_k-optimality criterion is that for $k > 1$ there is no guarantee that at an optimal design the information matrix will be nonsingular. Indeed, a very large sum of k smallest eigenvalues of the FIM does not necessarily mean that all of them are nonzero. Thus, identifiability may be lost at an E_k-optimum design. Another disadvantage is that the shape of the ellipsoid of concentration for the estimates is only indirectly influenced by the eigenvalues of the FIM. In fact, the squared lengths of the axes of the confidence ellipsoid are proportional to the eigenvalues of the inverse of the FIM. This constitutes a motivation behind a revision to the form of the E_k-optimality criterion.

The first aim of this paper is to investigate properties of the alternative form of the E_k-optimality criterion, being defined as the sum of k largest eigenvalues of the FIM. A design minimizing this criterion makes the sum of the squared lengths of k largest axes of the confidence ellipsoid minimal. Equivalently, it suppresses the average of k largest variances of the estimates. It constitutes a generalization of both E- and A-optimality, and it will be shown here that it possesses numerous desirable features, such as convexity, antitonicity or orthogonal invariance. What is more, there is no risk of ending up with a singular FIM provided that we deal with a problem in which the parameters are identifiable.

The selfimposed form of the modified E_k-optimality criterion is in marked contrast to the lack of communications on its use in the literature. A plausible explanation is its nonsmoothness and a less convenient form than for the genuine E_k-optimality criterion due to the replacement of the FIM by its inverse. As a result, it is impossible to employ the LP-based algorithm of Pronzato and Pàzman (2013) to numerically construct optimal designs. That is why the second objective of this paper is to develop an efficient computational algorithm for construction of the corresponding optimum designs.

The proposed algorithm retains the basic structure of the method of outer approximations employed by Burclová and Pázman (2016), i.e., it alternates between computing k smallest eigenvalues of the FIM along with their associated eigenvectors, and solving a finite min-max problem. The marked difference is that the latter is strongly nonlinear in the design weights and, therefore, LP solvers are of no use here. They key and original idea in attacking this possibly large-scale problem is to apply generalized simplicial decomposition (GSD), proposed by Bertsekas and Yu (2011), a viable counterpart of ordinary SD, which is directed towards nonsmooth convex minimax problems. Bertsekas and Yu (2011) only sketched the idea as an application of a more general approach stemming from extended monotropic programming. Successful attempts to use it in the context of optimum sensor selection were reported by Patan and Uciński (2019). Here we adopt it to the specificity of the constrained design problem in question, provide a separability form of optimality conditions, which can be easily applied to terminate the algorithm, and discuss specifics of our implementation. A nontrivial computational example is used to validate the proposed technique.

The paper is organized as follows. In Section 2 the general problem of optimum experimental design is defined with special emphasis on E_k-optimality. In Section 3 the proposed modified E_k-optimality criterion is introduced and its properties are discussed. Section 4 reformulates the problem in terms of continuous designs. In Section 5 the method of outer approximations is set forth as the tool to determine numerical approximations to the respective optimal designs. In Section 6 generalized simplicial decomposition is employed to implement the step involving the solution of a finite min-max problem. Section 7 reports computational results for a nontrivial problem of sensor selection. In Section 8 some concluding remarks are made. Finally, three appendices contain some accessory results or proofs of theoretical results.

Notation. Throughout the paper, $\mathbb{R}_+$ and $\mathbb{R}_{++}$ stand for the sets of nonnegative and positive real numbers, respectively. We adopt the convention that all vectors have column form. The set of real $m \times n$ matrices is denoted by $\mathbb{R}^{m\times n}$. We use $\mathbb{S}^m$ to denote the set of symmetric $m \times m$ matrices, $\mathbb{S}^m_+$ to denote the set of symmetric nonnegative definite $m \times m$ matrices, and $\mathbb{S}^m_{++}$ to denote the set of symmetric positive definite $m \times m$ matrices. The curled inequality symbol $\succeq$ (resp. $\succ$) is used to denote generalized inequalities. More precisely, between vectors, it represents a componentwise inequality, and between symmetric matrices, it represents the Loewner ordering: given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^m$, $\boldsymbol{A} \succeq \boldsymbol{B}$ (resp. $\boldsymbol{A} \succ \boldsymbol{B}$) means that $\boldsymbol{A} - \boldsymbol{B}$ is nonnegative (resp. positive) definite. The symbols $\mathbf{1}$ and $\mathbf{0}$ denote vectors whose all components are ones and zeros, respectively. The context makes their lengths clear. By analogy, $\mathbf{0}$ and $\boldsymbol{I}$ stand for the zero and identity matrices of appropriate dimensions, respectively. However, we shall occasionally write $\boldsymbol{I}_k$ for the $k \times k$ identity matrix to accentuate its dimensions. Given two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, their scalar product is denoted either by $\boldsymbol{a}^\mathsf{T}\boldsymbol{b}$ or $\boldsymbol{a} \cdot \boldsymbol{b}$.

Given a set of points A, $\mathrm{conv}(A)$ stands for its convex hull, i.e., the set of all convex combinations of elements of A. If A is convex, $\mathrm{ri}(A)$ signifies its relative interior. The probability (or canonical) simplex in $\mathbb{R}^n$ is defined as

$$\mathcal{S}_n = \mathrm{conv}\big(\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}\big) = \big\{\boldsymbol{p} \in \mathbb{R}^n_+ \mid \mathbf{1}^\mathsf{T}\boldsymbol{p} = 1\big\},$$

where $\boldsymbol{e}_j$ is the usual unit vector along the j-th coordinate of $\mathbb{R}^n$.

For any $\boldsymbol{A} \in \mathbb{S}^m$, let $\lambda_{\max}(\boldsymbol{A}) = \lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_m(\boldsymbol{A}) = \lambda_{\min}(\boldsymbol{A})$ denote the eigenvalues of $\boldsymbol{A}$ in decreasing order.

2. Optimum experimental design problem

Consider observations $\boldsymbol{y}_{ij}$ of a d-dimensional vector $\boldsymbol{y}$ of response variables, performed at fixed values $\boldsymbol{x}_i$ of the l-dimensional vector $\boldsymbol{x}$ of explanatory (or independent) variables (e.g., time, temperature, spatial location, drug doses, etc.), which follow the parametric model structure (Seber and Wild, 1989, p. 529)

$$\boldsymbol{y}_{ij} = \boldsymbol{\eta}(\boldsymbol{x}_i, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_{ij}, \qquad \begin{cases} j = 1, \ldots, r_i, \\ i = 1, \ldots, n. \end{cases}$$

We assume that $\boldsymbol{x}_i \neq \boldsymbol{x}_k$ whenever $i \neq k$. The additional index j is necessary when the observations are replicated $r_i > 1$ times for the setting $\boldsymbol{x}_i$. Then the total number of experimental runs is $N = \sum_{i=1}^n r_i$. Here the regression function $\boldsymbol{\eta} : \mathbb{R}^{l+m} \to \mathbb{R}^d$ is given *a priori* and $\boldsymbol{\theta}$ constitutes an m-dimensional vector of constant but unknown parameters. The d-dimensional vectors of additive random errors $\boldsymbol{\varepsilon}_{ij}$ disturbing the model are assumed to be sampled from a multivariate normal distribution satisfying

$$\mathbb{E}(\boldsymbol{\varepsilon}_{ij}) = \mathbf{0},$$
$$\mathbb{E}(\boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{k\ell}^\mathsf{T}) = \delta_{ik}\delta_{j\ell}\boldsymbol{V}(\boldsymbol{x}_i),$$

where the dispersion matrices $\boldsymbol{V}(\boldsymbol{x}_i) \in \mathbb{S}^d_{++}$, $i = 1, \ldots, n$ are known, possibly up to a common constant multiplier, and δ_{ij} signifies the Kronecker delta. This means that observations at different experimental conditions are uncorrelated, but we allow for correlations between individual responses.

For a linear functional form of $\boldsymbol{\eta}$, we have

$$\boldsymbol{\eta}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{F}(\boldsymbol{x}_i)^\mathsf{T}\,\boldsymbol{\theta},$$

where $\boldsymbol{F}(\boldsymbol{x}_i) \in \mathbb{R}^{m\times d}$, $i = 1,\dots,n$ are known. Let $\boldsymbol{M}_i \in \mathbb{S}_+^m$ be given by

$$\boldsymbol{M}_i = \boldsymbol{F}(\boldsymbol{x}_i)\boldsymbol{V}^{-1}(\boldsymbol{x}_i)\boldsymbol{F}^{\mathsf{T}}(\boldsymbol{x}_i), \quad i = 1,\dots,n.$$

If the matrix

$$\boldsymbol{M} = \sum_{i=1}^{n} r_i \boldsymbol{M}_i,$$

called the *Fisher information matrix* (FIM), has full rank, then the weighted least squares estimator of $\boldsymbol{\theta}$ is given by (Fedorov and Leonov, 2014, p. 15)

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{M}^{-1} \sum_{i=1}^{n} r_i \boldsymbol{F}(\boldsymbol{x}_i)\boldsymbol{V}^{-1}(\boldsymbol{x}_i)\bar{\boldsymbol{y}}_i,$$

where

$$\bar{\boldsymbol{y}}_i = \frac{1}{r_i}\sum_{j=1}^{r_i} \boldsymbol{y}_{ij}.$$

It is unbiased and efficient, and has the multivariate normal distribution with

$$\mathrm{Cov}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{M}^{-1}.$$

On some mild assumptions (Pronzato and Pàzman, 2013), these properties hold asymptotically for a nonlinear function $\boldsymbol{\eta}$ if we define $\widehat{\boldsymbol{\theta}}$ as the maximum-likelihood estimator of $\boldsymbol{\theta}$ and replace $\boldsymbol{F}^{\mathsf{T}}(\boldsymbol{x}_i)$ by $\partial\boldsymbol{\eta}(\boldsymbol{x}_i,\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}\big|_{\boldsymbol{\vartheta}=\boldsymbol{\theta}}$.

We assume that the values of $\boldsymbol{x}_i$, $i = 1,\dots,n$ are fixed and may not be altered, but we have full control over the corresponding numbers of replications r_i, $i = 1,\dots,n$. The focus here will be on choosing the latter values in some optimal way to enhance the process of estimating $\boldsymbol{\theta}$. To form a basis for the comparison of different experiments, a number of real-valued criteria defined on the FIM have been proposed (Atkinson *et al.*, 2007; Pronzato and Pàzman, 2013; Fedorov and Leonov, 2014). They are most often related to the confidence ellipsoid, i.e., a highest probability density region for the parameters. The most common options include (we use symbols Ψ and Φ for criteria which are supposed to be minimized and maximized, respectively):

(i) the D-optimality criterion,

$$\Phi_{\mathrm{D}}(\boldsymbol{M}) = \det{}^{1/m}(\boldsymbol{M}),$$

maximization of which amounts to minimizing the volume of the confidence ellipsoid;

(ii) the A-optimality criterion,

$$\Psi_{\mathrm{A}}(\boldsymbol{M}) = \mathrm{trace}(\boldsymbol{M}^{-1}),$$

minimization of which is equivalent to minimizing the sum of the squared lengths of the axes of the confidence ellipsoid;

(iii) the E-optimality criterion,

$$\Phi_{\mathrm{E}}(\boldsymbol{M}) = \lambda_{\min}(\boldsymbol{M}),$$

maximization of which leads to minimizing the length of the largest axis of the confidence ellipsoid;

(iv) the trace-optimality criterion,

$$\Phi_{\mathrm{tr}}(\boldsymbol{M}) = \mathrm{trace}(\boldsymbol{M}), \tag{1}$$

maximization of which makes the sum of the diagonal elements of the FIM maximal.

Different criteria yield slightly different optimal designs and the choice of a particular criterion is dictated by a specific application and ease of computations. The criterion (1) is occasionally used due to its simplicity, but it may lead to serious problems with identifiability. Indeed, it is employed in the hope that large diagonal elements of the FIM will translate into "small" elements of its inverse, but this may not necessarily be the case. Taken to extremes, the use of this criterion may result in a singular FIM (Zarrop and Goodwin, 1975).

In the search for a flexible and general criterion, Harman (2004) advocated maximization of the E_k-optimality criterion

$$\Phi_{\mathrm{E}_k}(\boldsymbol{M}) = \sum_{\ell=m-k+1}^{m} \lambda_\ell(\boldsymbol{M}),$$

i.e., the sum of k smallest eigenvalues of the FIM for an arbitrarily selected $k \in \{1,\dots,m\}$. In fact, it defines a family of criteria which range from E-optimality (for $k = 1$) to trace-optimality (for $k = m$). A distinctive feature of this criterion is orthogonal invariance, i.e.,

$$\Phi_{\mathrm{E}_k}(\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}}) = \Phi_{\mathrm{E}_k}(\boldsymbol{M})$$

for any orthonormal (i.e., satisfying $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}_m$) matrix $\boldsymbol{U} \in \mathbb{R}^{m\times m}$.

Burclová and Pázman (2016) developed a relatively simple algorithm to compute E_k-optimum designs. It makes use of the formulation in terms of an SIP problem and applies the method of outer approximations exposed by Shimizu and Aiyoshi (1980). This is an extension of the idea set forth by Pronzato and Pàzman (2013, p. 326) for the E-optimality criterion. In general, SIP problems are markedly hard to solve, especially when the lower-level program is not convex (and this is the case here). Specifically, a formidable challenge is that to establish feasibility of any value of the upper-level decision variable means to find a global extremum in the lower-level problem. This imposes a heavy computational burden, which is even more pronounced in generalized SIP problems receiving particular attention recently (Djelassi *et al.*, 2019), in which the set of

constraints depends on the upper-level decision variables. But Burclová and Pázman (2016) exploit the relative simplicity of their SIP formulation to the maximum and their computational scheme alternates between computing k smallest eigenvalues of an $m \times m$ matrix accompanied by the corresponding eigenvectors and solving an LP subproblem. The simplicity of the approach makes the algorithm well-suited for very large cardinalities of the set of candidate design points.

Unfortunately, for $k > 1$ the E_k-optimality criterion inherits the main drawback of the trace-optimality criterion, i.e., it does not prevent optimal information matrices from being singular. Indeed, a large sum of k smallest eigenvalues of the FIM does not imply that the minimal eigenvalue is positive.

Another severe disadvantage is that manipulating the sum of smallest eigenvalues of the FIM influences the axes of the uncertainty ellipsoid only indirectly and in a rather awkward manner. A high value of this criterion might correspond to an ellipsoid which is extremely elongated, at least in one direction, which means an excessive variability of the estimates and contradicts the main objective of optimum design.

3. Modified E_k-optimality

In order to remove the above disadvantages of the E_k-optimality criterion while retaining its ability to embrace a range of optimality criteria, the following modified E_k-optimality criterion is proposed:

$$\Psi_{\mathrm{E}_k^{\mathrm{inv}}}(\boldsymbol{M}) = \begin{cases} \displaystyle\sum_{\ell=1}^{k} \lambda_\ell(\boldsymbol{M}^{-1}) & \text{if } \boldsymbol{M} \succ \boldsymbol{0}, \\ +\infty & \text{otherwise.} \end{cases} \tag{2}$$

As it is the sum of k largest eigenvalues of the inverse of the FIM, its minimization will suppress the sum of the squared lengths of k largest axes of the confidence ellipsoid. Its finite values always correspond to a nonsingular FIM, which implies identifiability. On the one hand, $\mathrm{E}_k^{\mathrm{inv}}$-optimum designs are E-optimum ones for $k = 1$ and, on the other, they constitute A-optimum designs for $k = m$. This means that the family of optimality criteria so defined for $k \in \{1, \ldots, m\}$ has a better interpretability in terms of the estimation accuracy than the same family for the genuine E_k-optimality criterion.

While analyzing the properties of the $\mathrm{E}_k^{\mathrm{inv}}$-optimality criterion and setting up the numerical algorithm of Section 5, the following equivalent form is essential:

$$\Psi_{\mathrm{E}_k^{\mathrm{inv}}}(\boldsymbol{M}) = \max_{\boldsymbol{Q} \in \mathcal{Q}} \operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}} \boldsymbol{M}^{-1} \boldsymbol{Q}) \tag{3}$$

for $\boldsymbol{M} \in \mathbb{S}_{++}^m$, where

$$\mathcal{Q} = \{\boldsymbol{Q} \in \mathbb{R}^{m \times k} : \boldsymbol{Q}^{\mathsf{T}} \boldsymbol{Q} = \boldsymbol{I}_k\},$$

cf. Theorem A1. The set $\mathcal{Q}$ is not convex, but it is compact. (The Frobenius norm of any $\boldsymbol{Q} \in \mathcal{Q}$ equals $\sqrt{k}$.)

For notational simplicity, in what follows we write Ψ instead of $\Psi_{\mathrm{E}_k^{\mathrm{inv}}}$.

The result below lists the main analytical properties of the proposed criterion. Its proof is included in Appendix B.

Theorem 1. *In $\mathbb{S}_+^m$ the design criterion Ψ is*

(i) antitonic, i.e., reversing the Loewner ordering,

$$\boldsymbol{M}_1 \preceq \boldsymbol{M}_2 \Longrightarrow \Psi(\boldsymbol{M}_1) \geq \Psi(\boldsymbol{M}_2),$$

(ii) positively homogeneous of degree -1, i.e.,

$$\Psi(\alpha \boldsymbol{M}) = \frac{1}{\alpha} \Psi(\boldsymbol{M}), \quad \forall \alpha > 0,$$

(iii) convex,

(iv) orthogonally invariant.

A serious problem presented by the criterion (2) is its potential nondifferentiability for $k < m$, which may happen in case its matrix argument has multiple eigenvalues. As a very simple example, set $\boldsymbol{M} = \boldsymbol{I}$ and observe that then the partial derivatives of $\mathrm{E}_k^{\mathrm{inv}}$ with respect to the diagonal elements of $\boldsymbol{M}$ fail to exist. In general, we must cope not only with the likelihood of a nondifferentability at a minimizing point, but also with the fact that our highly nonlinear objective function has no simple analytical expression. Fortunately, the convexity of the design criterion makes it possible to exploit some of the machinery of convex analysis. In Section 5 we demonstrate that the alternative form of this criterion, Eqn. (3), is computationally tractable and its minimization can be made into a convergent and implementable procedure.

4. Relaxed formulation using continuous designs

The resulting optimization problem constitutes a classical discrete resource allocation problem (Katoh, 2001): Given a total amount of N observations, we wish to allocate it to n measurement settings so that the objective value (cost) $\Psi[\boldsymbol{M}(r_1, \ldots, r_n)]$ is minimized. Its combinatorial nature implies that calculus techniques cannot be exploited in the solution and, with a long list of candidate settings $\boldsymbol{x}_i$ and a large N, complicated search algorithms can readily consume appreciable computer time and space.

A commonly used device for this problem is to extend the definition of the solution (Atkinson *et al.*, 2007; Fedorov and Leonov, 2014; Pronzato and Pàzman, 2013). To this end, it is convenient to operate on the frequencies

of observations $p_i = r_i/N$, called *weights*, in place of the design variables r_i, $i = 1, \ldots, n$, and on the so-called *normalized* FIM

$$\widetilde{\boldsymbol{M}}(\boldsymbol{p}) = \sum_{i=1}^{n} p_i \boldsymbol{M}_i$$

in lieu of $\boldsymbol{M}(r_1, \ldots, r_n)$. Note that $n \leq N$ and $\boldsymbol{p} = (p_1, \ldots, p_n)$ satisfies

$$\mathbf{1}^{\mathsf{T}}\boldsymbol{p} = 1, \quad \boldsymbol{p} \succeq \mathbf{0}. \tag{4}$$

When N is large, we may dispense with the restriction that the feasible weights are integer multiples of $1/N$ and seek a solution in a wider class of *approximate designs*, i.e., the ones in which the weights are any real numbers satisfying (4). This relaxation makes the optimization problem more tractable and we adhere to it in what follows. Moreover, owing to antitonicity, a $\boldsymbol{p}$ which minimizes $\Psi[\widetilde{\boldsymbol{M}}(\boldsymbol{p})]$ yields multiplicities $r_1, \ldots, r_n$ which also minimize $\Psi[\boldsymbol{M}(r_1, \ldots, r_n)]$. Consequently, minimization of this slightly modified design criterion $\Psi[\widetilde{\boldsymbol{M}}(\,\cdot\,)]$ is further considered. For simplicity of notation, we will also drop the tilde over $\boldsymbol{M}(\,\cdot\,)$. The settings $\boldsymbol{x}_i$ associated with positive weights p_i are said to be *support points*.

One of the common criticisms of the mathematical optimal design is that the produced designs concentrate at a relatively small number of support points, rather than spreading the measurement effort around appropriately, which many practicing statisticians tend to do (Cook and Fedorov, 1995). Indeed, the results reported by various authors indicate that the number of support points with nonzero weights is most often close to the number of the estimated parameters. This gave rise to investigations aiming at imposing the appropriate limitations on the form of the optimal designs. Following this line of research, in the remainder of the paper, we are interested in solving the linearly constrained design problem stated as follows.

Problem 1. Given a vector $\boldsymbol{b} \in \mathbb{R}^n_{++}$ satisfying $\mathbf{1}^{\mathsf{T}}\boldsymbol{b} > 1$, find a vector of weights $\boldsymbol{p} = (p_1, \ldots, p_n)$ to minimize

$$J(\boldsymbol{p}) = \Psi\big(\boldsymbol{M}(\boldsymbol{p})\big)$$

over the set $\mathcal{P} = \big\{\boldsymbol{p} \in \mathbb{R}^n \mid \mathbf{0} \preceq \boldsymbol{p} \preceq \boldsymbol{b}, \mathbf{1}^{\mathsf{T}}\boldsymbol{p} = 1\big\}$.

In this formulation, we prevent spending the overall experimental effort at few points by directly bounding the frequencies of observations from above. Problems of this type have received close attention in the general framework of optimal design with bounded density (Cook and Fedorov, 1995; Fedorov, 1989; Sahm and Schwabe, 2001). In the specific context considered here, this formulation possesses a number of notable features which, in theory, should make its solution straightforward. First of all, note that the performance index Ψ is convex over the canonical simplex $\mathcal{S}_n$ due to the convexity of Ψ and the linear dependence of the FIM on the weights. Moreover, the constraint set $\mathcal{P}$ constitutes the intersection of S_n and the box $B = \big\{\boldsymbol{p} \in \mathbb{R}^n \mid \mathbf{0} \preceq \boldsymbol{p} \preceq \boldsymbol{b}\big\}$, which is a rather nice convex set.

Write $\mathcal{P}_+ = \big\{\boldsymbol{p} \in \mathcal{P} \mid \boldsymbol{M}(\boldsymbol{p}) \succ \mathbf{0}\big\}$. For abbreviation, we set

$$f(\boldsymbol{p}, \boldsymbol{Q}) = \begin{cases} \operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{Q}) & \text{if } \boldsymbol{p} \in \mathcal{P}_+, \\ +\infty & \text{otherwise.} \end{cases}$$

5. Algorithm for construction of $\mathrm{E}_k^{\mathrm{inv}}$-optimum designs

5.1. Regularization of the optimality criterion. Before proceeding further, observe that we need to restate Problem 1 in order to make the numerical search for an optimum design well conditioned. This is due to the possible bad behaviour of the mapping f whenever $\boldsymbol{M}(\boldsymbol{p})$ tends to become singular. This is illustrated by the following example.

Example 1. Assume that $\boldsymbol{M}_1 = \operatorname{diag}([1, 0, 0])$, $\boldsymbol{M}_2 = \operatorname{diag}([0, 1, 0])$, $\boldsymbol{M}_3 = \operatorname{diag}([0, 0, 1])$. Consider the weights parameterized as

$$\boldsymbol{p}(\gamma) = \big(\tfrac{1}{2}(1-\gamma), \tfrac{1}{2}(1-\gamma), \gamma\big)$$

by $\gamma \in [0, 1]$. For $\gamma \in (0, 1)$ we have that $\boldsymbol{M}^{-1}(\boldsymbol{p}(\gamma)) = \operatorname{diag}([2/(1-\gamma), 2/(1-\gamma), 1/\gamma])$. Now consider

$$\boldsymbol{Q}_1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{Q}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

It follows that

$$f(\boldsymbol{p}(\gamma), \boldsymbol{Q}_1) = \frac{2}{1-\gamma} + \frac{1}{\gamma} \to +\infty,$$
$$f(\boldsymbol{p}(\gamma), \boldsymbol{Q}_2) = \frac{4}{1-\gamma} \to 4$$

as $\gamma \to 0$. This means that numerical exploration of the design criterion through f is in danger of approaching the singular matrix $\boldsymbol{M}(\boldsymbol{p}(0)) = \operatorname{diag}([1/2, 1/2, 0])$ just by odd coincidence. A consistent behaviour of the function f is to be expected here, i.e., $f(\boldsymbol{p}, \boldsymbol{Q})$ should grow unboundedly whenever the FIM tends to be singular. This singularity is a symptom of the loss of identifiability (Pronzato and Pàzman, 2013; Coll and Sánchez, 2019) and it is quite natural to introduce a safeguard against this undesirable situation built into the computational procedure. ♦

In order to avoid a potentially degenerate numerical minimization problem, a penalty for solutions yielding $\boldsymbol{M}(\boldsymbol{p})$ close to being singular is incorporated in the design

criterion. Specifically, setting β as an arbitrary small positive real, we replace Ψ by

$$\widetilde{\Psi}(\boldsymbol{M}) \begin{cases} \max\limits_{\boldsymbol{Q}\in\mathcal{Q}} \operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}\boldsymbol{Q}) \\ -\beta\log\det(\boldsymbol{M}) & \text{if } \boldsymbol{M}\succ\boldsymbol{0}, \\ +\infty & \text{otherwise.} \end{cases} \tag{5}$$

Clearly, the term $-\beta\log\det(\boldsymbol{M})$ grows unboundedly as $\boldsymbol{M}$ tends to be singular. Since $\Psi(\boldsymbol{M}) = \sum_{\ell=1}^{k}\lambda_\ell(\boldsymbol{M}^{-1}) \geq k\lambda_{\min}(\boldsymbol{M}^{-1}) = k/\lambda_{\max}(\boldsymbol{M}) > 0$, the same behaviour of $\widetilde{\Psi}(\boldsymbol{M})$ is guaranteed. What is more, the concavity of the log-determinant on $\mathbb{S}^m_{++}$ implies that the convexity of the design criterion is retained.

The value of the adjustable parameter β should be small, so as not to pull the computed design away from minimizing $\Psi(\boldsymbol{M})$ in favour of maximizing $\det(\boldsymbol{M})$. In computer experiments reported in Section 7 its choice has been made via a trial-and-error procedure so as to make the ultimate value of the penalty component remain within a margin of several per cent of the value of the compound criterion (5). A similar strategy proved to perform quite well in a similar design problem for correlated observations investigated by Uciński (2020).

Continuing in this fashion, we introduce the function

$$\widetilde{f}(\boldsymbol{p},\boldsymbol{Q}) = \begin{cases} \operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{Q}) \\ -\beta\log\det(\boldsymbol{M}) & \text{if } \boldsymbol{p}\in\mathcal{P}_+, \\ +\infty & \text{otherwise,} \end{cases}$$

and the corresponding regularized optimality criterion $\widetilde{J}(\boldsymbol{p}) = \widetilde{\Psi}(\boldsymbol{M}(\boldsymbol{p}))$ in lieu of f and J, respectively.

For notational simplicity, from now on we will write $\widetilde{J}$, $\widetilde{\Psi}$ and $\widetilde{f}$ simply as J, Ψ and f, respectively.

5.2. Method of outer approximations. The $\mathrm{E}_k^{\mathrm{inv}}$-optimality criterion is convex, but nondifferentiable, which suggests that its minimization may be nontrivial. For $k = 1$, i.e., the E-optimality criterion, minimizers can sometimes be determined in closed form. This is the case of polynomial or trigonometric regression (Pukelsheim, 1993; Melas, 2006); see also the work of Harman (2004), where E_k-optimal designs are found explicitly for polynomial regression. In general, however, it is rather hard to construct designs of this type without resorting to numerical optimization.

Observe that Problem 1 can equivalently be formulated as the following semi-infinite programming (SIP) problem (Hettich and Kortanek, 1993; Polak, 1987; Reemtsen and Görner, 1998): Determine $\boldsymbol{p}\in\mathcal{P}$ and $\alpha\in\mathbb{R}$ to minimize α subject to the constraints

$$f(\boldsymbol{p},\boldsymbol{Q}) \leq \alpha, \quad \boldsymbol{Q}\in\mathcal{Q}. \tag{6}$$

The reason why such problems are called semi-infinite is that the design vector $\boldsymbol{p}$ is finite dimensional but the number of constraints (6) is infinite. This type of conversion has been used frequently in optimum experimental design (cf., e.g., Duarte and Wong, 2014), especially while constructing robust designs.

A very convenient technique to solve the above SIP problem is the method of outer approximations (Polak, 1997, p. 460) which reduces computations to solving a sequence of simpler finite min-max problems. Strictly speaking, in each step, the constraints (6) are replaced by

$$f(\boldsymbol{p},\boldsymbol{Q}) \leq \alpha, \quad \boldsymbol{Q}\in\mathcal{R}.$$

for a finite set $\mathcal{R}\subset\mathcal{Q}$ consisting of the most representative values of $\mathcal{Q}$. Thus, the constraint set $\{(\boldsymbol{p},\alpha)\in\mathcal{P}\times\mathbb{R} : f(\boldsymbol{p},\boldsymbol{Q})\leq\alpha, \forall\boldsymbol{Q}\in\mathcal{Q}\}$ is a subset of $\{(\boldsymbol{p},\alpha)\in\mathcal{P}\times\mathbb{R} : f(\boldsymbol{p},\boldsymbol{Q})\leq\alpha, \forall\boldsymbol{Q}\in\mathcal{R}\}$, i.e., the latter constitutes an "outer approximation" to the former, which accounts for the name of the method.

From now on, we write

$$J_{\mathcal{R}}(\boldsymbol{p}) = \max_{\boldsymbol{Q}\in\mathcal{R}} f(\boldsymbol{p},\boldsymbol{Q}).$$

Obviously, we always have $J(\boldsymbol{p}) \geq J_{\mathcal{R}}(\boldsymbol{p})$, $\forall\boldsymbol{p}\in\mathcal{P}$.

The algorithm implementing the method of outer approximations is outlined as Algorithm 1. Note that minimizers and maximizers in Steps 1 and 2, respectively, should be global.

Global minimizers of J are the only accumulation points of the sequence $\{\boldsymbol{p}^{(\kappa)}\}$ and the termination

Algorithm 1. Method of outer approximations.

Step 0. (*Initialization*) Guess initial weights $\boldsymbol{p}^{(0)}\in\mathcal{P}_+$. Set $\kappa = 0$, compute $\boldsymbol{Q}^{(0)} = \arg_{\boldsymbol{Q}\in\mathcal{Q}}\max f(\boldsymbol{p}^{(0)},\boldsymbol{Q})$, and set $\mathcal{Q}^{(0)} = \{\boldsymbol{Q}^{(0)}\}$. Choose $0<\epsilon\ll 1$, a parameter used in the stopping rule.

Step 1. (*Solution of the finite min-max problem*) Compute

$$\boldsymbol{p}^{(\kappa+1)} = \arg\min_{\boldsymbol{p}\in\mathcal{P}} J_{\mathcal{Q}^{(\kappa)}}(\boldsymbol{p}). \tag{7}$$

Step 2. (*Determination of the next representative orthogonal matrix*) Compute

$$\boldsymbol{Q}^{(\kappa+1)} = \arg\max_{\boldsymbol{Q}\in\mathcal{Q}} f(\boldsymbol{p}^{(\kappa+1)},\boldsymbol{Q}). \tag{8}$$

Step 3. (*Termination check*) If

$$f(\boldsymbol{p}^{(\kappa+1)},\boldsymbol{Q}^{(\kappa+1)}) \leq J_{\mathcal{Q}^{(\kappa)}}(\boldsymbol{p}^{(\kappa+1)})(1+\epsilon) \tag{9}$$

then STOP and $\boldsymbol{p}^{(\kappa+1)}$ is optimal. Otherwise, set $\mathcal{Q}^{(\kappa+1)} = \mathcal{Q}^{(\kappa)}\cup\{\boldsymbol{Q}^{(\kappa+1)}\}$, replace κ by $\kappa+1$, and go to Step 1.

condition of Step 3 will be satisfied in a finite number of iterations (cf. Shimizu and Aiyoshi, 1980, Thm. 3; Polak, 1997, Thm. 3.5.20, p. 462). Basically, these convergence results require the continuity of f on $\mathcal{P} \times \mathcal{Q}$ and the compactness of $\mathcal{P}$ and $\mathcal{Q}$. (Note that the convexity of f, $\mathcal{P}$ and $\mathcal{Q}$ is not required.) Here this function may be unbounded, but those results still apply. This is because for the needs of the proof of convergence we could replace f by the composition $g \circ f$, where g is any strictly increasing continuous mapping transforming $[0, +\infty]$ onto a compact interval, e.g., we could set $g(v) = \arctan(v)$. It is easy to see that such a transformation does not alter the minimizers.

Observe that the implementation of Step 2 is straightforward. Theorem A1 yields

$$\begin{aligned}\boldsymbol{Q}^{(\kappa)} &= \left[\ \boldsymbol{v}_1(\boldsymbol{M}^{-1}(\boldsymbol{p}^{(\kappa)})) \ \middle|\ \cdots \ \middle|\ \boldsymbol{v}_k(\boldsymbol{M}^{-1}(\boldsymbol{p}^{(\kappa)})) \ \right] \\ &= \left[\ \boldsymbol{v}_m(\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})) \ \middle|\ \cdots \ \middle|\ \boldsymbol{v}_{m-k+1}(\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})) \ \right],\end{aligned} \tag{10}$$

where $\boldsymbol{v}_\ell$ stands for the normalized eigenvector corresponding to λ_ℓ, $\ell = 1, \ldots, m$.

Note that these eigenvectors are required to be orthonormal, which is guaranteed if a system of orthonormal eigenvectors of $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$ is available. The existence of such a system results from the symmetry of $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$ (Harville, 1997, Cor. 21.5.9, p. 534), but standard numerical solvers may fail to produce it in case $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$ has multiple eigenvalues (usually, they only guarantee the linear independence of the eigenvectors corresponding to the same eigenvalue). This problem can, however, be easily addressed using the Schur decomposition of $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$, the implementations of which are provided by most numerical libraries. Indeed, the spectrum of $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$ is real, which means that there exist an orthogonal matrix $\boldsymbol{V} \in \mathbb{R}^{m \times m}$ and an upper triangular matrix $\boldsymbol{B} \in \mathbb{R}^{m \times m}$ whose diagonal elements are not-necessary-distinct eigenvalues of $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$ (in arbitrary order) such that $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)}) = \boldsymbol{V}\boldsymbol{B}\boldsymbol{V}^{\mathsf{T}}$ (cf. Bernstein, 2005, Cor. 5.4.3, p. 172; Harville, 1997, Thm. 21.5.11, p. 536). But the symmetry of $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$ additionally yields $\boldsymbol{B}^{\mathsf{T}} = (\boldsymbol{V}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{V})^{\mathsf{T}} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{M}\boldsymbol{V} = \boldsymbol{B}$, which means that $\boldsymbol{B}$ is actually diagonal and then the columns of $\boldsymbol{V}$ form the desired system of orthonormal eigenvectors of $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$.

Clearly, for $\boldsymbol{M}(\boldsymbol{p}^{(\kappa)})$ with multiple eigenvalues the eigenvectors $\boldsymbol{v}_\ell$ are thus defined up to possible permutations within the groups of the eigenvectors corresponding to the same eigenvalues, but Theorem A1 implies that any arbitrary choice of the $\boldsymbol{v}_\ell$'s forces (8) provided that $\boldsymbol{Q}^{(\kappa+1)}$ is defined by (10). (The method of outer approximations does not require the uniqueness of the global minimizers and maximizers in (7) and (8), respectively.)

As for Step 3, in (9) we make use of

$$\begin{aligned}f(\boldsymbol{p}^{(\kappa)}, \boldsymbol{Q}^{(\kappa)}) &= \sum_{\ell=1}^{k} \lambda_\ell(\boldsymbol{M}^{-1}(\boldsymbol{p}^{(\kappa)})) \\ &= \sum_{\ell=m-k+1}^{m} \lambda_\ell^{-1}(\boldsymbol{M}(\boldsymbol{p}^{(\kappa)}))\end{aligned}$$

However, the solution of the finite min-max problem (7) of Step 1 is not obvious. At first sight, it does not appear involved since the constraint set $\mathcal{P}$ is the intersection of a hyperbox and a hyperplane, which is a nice polyhedral set. Therefore, by introducing an additional scalar variable α, we could rewrite it as follows.

Problem 2. Find a pair $(\boldsymbol{p}^\star, \alpha^\star)$ to minimize α subject to

$$\begin{gathered}f_0(\boldsymbol{p}) \le \alpha, \quad \ldots, \quad f_\kappa(\boldsymbol{p}) \le \alpha, \\ (\boldsymbol{p}, \alpha) \in \mathcal{P} \times \mathbb{R},\end{gathered}$$

where $f_j(\boldsymbol{p}) = f(\boldsymbol{p}, \boldsymbol{Q}^{(j)})$, $j = 0, \ldots, \kappa$.

Basically, this is a smooth convex optimization problem which could be numerically solved using Newton-like methods, e.g., SQP. Unfortunately, application of this clear idea is complicated by a possibly very large value of $n+1$, the number of decision variables. An alternative technique is therefore badly needed. In what follows it is demonstrated that, with a little turning up, it is possible to retain the simplicity of Problem 2 and efficiently solve it.

6. General simplicial decomposition for the finite min-max problem

6.1. Generalized simplicial decomposition. Uciński and Patan (2007), Patan and Uciński (2008), Uciński (2012; 2015), as well as Herzog *et al.* (2018) demonstrated that for differentiable convex design criteria the inner linearization algorithm called simplicial decomposition (SD) (Bertsekas, 2015; Patriksson, 2001) proves extremely effective at drastically reducing the problem dimensionality and exploiting special structure present in common design criteria.

Specifically, $\mathcal{P}$ is approximated with the convex hull of an ever expanding set $\mathcal{P}^{(\tau)}$ that consists of extreme points of $\mathcal{P}$ plus an arbitrary starting point $\boldsymbol{p}^{(0)} \in \mathcal{P}$. The method alternates between minimization of the design criterion over $\mathrm{conv}(\mathcal{P}^{(\tau)})$ (this set has a relatively low number of extreme points and this is where a substantial dimensionality reduction emerges) and addition of a new extreme point $\widetilde{\boldsymbol{p}}_\tau \in \mathcal{P}$ so as to guarantee a cost improvement when it is minimized over $\mathrm{conv}(\mathcal{P}^{(\tau+1)})$, where $\mathcal{P}^{(\tau+1)} = \mathcal{P}^{(\tau)} \cup \{\widetilde{\boldsymbol{p}}_\tau\}$ (this is done by minimizing

the linearized design criterion over $\mathcal{P}$, which boils down to solving a simple LP problem).

Unfortunately, there is no direct way to extend the applicability of SD to nondifferentiable convex cost functions. Larsson *et al.* (2015; 1998) exposed ergodic sequences of subgradients and a conditional subgradient method, but the appealing simplicity of the SD algorithm was thereby lost. Bertsekas and Yu (2011), however, advanced generalized simplicial decomposition (GSD), which retains some key advantages of SD: it involves a solution of linear programs, called the column generation problems (CGPs), to generate new extreme points of $\mathcal{P}$ and a solution of typically low-dimensional nonlinear convex programs over $\mathrm{conv}(\mathcal{P}^{(\tau)})$, called the restricted master problems (RMP). The latter programs have low dimensions as long as $\mathcal{P}^{(\tau)}$ has a relatively small cardinality. Moreover, the sequence of the solutions produced by GSD tends to a solution of the original problem in a finite number of iterations (due to the polyhedral form of $\mathcal{P}$) steadily decreasing the objective function.

Adapting the universal GSD scheme to our needs, we obtain Algorithm 2. In the sequel, its consecutive steps will be discussed in turn.

6.2. Initialization. For simplicity, we assume that $\mathcal{P}^{(0)}$ consists of only two points, i.e., $\mathcal{P}^{(0)} = \{\boldsymbol{p}^1, \boldsymbol{p}^2\}$, where $\boldsymbol{p}^1 \in \mathrm{ri}(\mathcal{P})$. Since $\mathbf{1}^{\mathsf{T}}\boldsymbol{b} > 1$ and $\boldsymbol{b} \succ \mathbf{0}$, we set

$$\boldsymbol{p}^1 = \frac{1}{\mathbf{1}^{\mathsf{T}}\boldsymbol{b}}\boldsymbol{b}.$$

(If all the elements of $\boldsymbol{b}$ are equal to each other, $\boldsymbol{p}^1$ is then simply the centre of $\mathcal{S}_n$.) Then $\boldsymbol{p}^2$ is selected as a point minimizing the value of $J'_{\mathcal{Q}^{(\kappa)}}(\boldsymbol{p}^1; \boldsymbol{p}-\boldsymbol{p}^1)$, the directional derivative of $J_{\mathcal{Q}^{(\kappa)}}$ at $\boldsymbol{p}^1$ in the direction $\boldsymbol{p}-\boldsymbol{p}^1$, over $\boldsymbol{p} \in \mathcal{P}$.

By Danskin's theorem (Bertsekas, 1999, p. 717), we have

$$J'_{\mathcal{Q}^{(\kappa)}}(\boldsymbol{p}^1; \boldsymbol{p}-\boldsymbol{p}^1) = \max_{j: f_j(\boldsymbol{p}^1) = J_{\mathcal{Q}^{(\kappa)}}(\boldsymbol{p}^1)} (\boldsymbol{p}-\boldsymbol{p}^1)^{\mathsf{T}} \nabla f_j(\boldsymbol{p}^1).$$

Therefore, $\boldsymbol{p}^2$ is the first component of the solution to the following LP problem: Find a pair $(\boldsymbol{p}^\star, \omega^\star)$ to minimize ω subject to

$$\begin{aligned} \left[\nabla f_j(\boldsymbol{p}^1)\right]^{\mathsf{T}} \boldsymbol{p} &\le \left[\nabla f_j(\boldsymbol{p}^1)\right]^{\mathsf{T}} \boldsymbol{p}^1 + \omega, \\ &\forall j : f_j(\boldsymbol{p}^1) = J_{\mathcal{Q}^{(\kappa)}}(\boldsymbol{p}^1), \\ (\boldsymbol{p}, \omega) &\in \mathcal{P} \times \mathbb{R}. \end{aligned}$$

Algorithm 2. Solving Problem 2 via GSD.

Step 0. (*Initialization*) Guess an initial finite set $\mathcal{P}^{(0)} \subset \mathcal{P}$ containing a point in $\mathrm{ri}(\mathcal{P})$ and such that $\mathrm{card}(\mathcal{P}^{(0)}) \ge 2$. Set $\tau = 0$.

Step 1. (*Solution of the restricted master problem*) Find a pair $(\boldsymbol{p}^{(\tau)}, \alpha^{(\tau)})$ to minimize α subject to

$$\begin{aligned} f_0(\boldsymbol{p}) \le \alpha, \quad \ldots, \quad f_\kappa(\boldsymbol{p}) \le \alpha, \\ (\boldsymbol{p}, \alpha) \in \mathrm{conv}(\mathcal{P}^{(\tau)}) \times \mathbb{R}, \end{aligned} \tag{11}$$

along with the appropriate values of the dual optimal variables $\mu_j^{(\tau)} \ge 0$, $j = 0, \ldots, \kappa$ corresponding to the inequality constraints (11).

Step 2. (*Termination check*) If there exist some real numbers c and η_j, $j = 0, \ldots, \kappa$ satisfying

$$\sum_{j=0}^{\kappa} \eta_j \frac{\partial f_j(\boldsymbol{p}^{(\tau)})}{\partial p_i} \begin{cases} \ge c & \text{if } p_i^{(\tau)} = 0, \\ = c & \text{if } 0 < p_i^{(\tau)} < b_i, \\ \le c & \text{if } p_i^{(\tau)} = b_i \end{cases} \quad \text{for } i = 1, \ldots, n, \tag{12}$$

$$\eta_j \begin{cases} = 0 & \text{if } f_j(\boldsymbol{p}^{(\tau)}) < \alpha^{(\tau)}, \\ \ge 0 & \text{otherwise} \end{cases} \quad \text{for } j = 0, \ldots, \kappa, \tag{13}$$

$$\sum_{j=0}^{\kappa} \eta_j = 1, \tag{14}$$

then STOP and $\boldsymbol{p}^{(\tau)}$ is optimal.

Step 3. (*Solution of the column generation problem*) Set

$$\gamma^{(\tau)} = \sum_{j=0}^{\kappa} \mu_j^{(\tau)} \nabla f_j(\boldsymbol{p}^{(\tau)}).$$

Generate an extreme point $\widetilde{\boldsymbol{p}}_\tau \in \mathcal{P}$ as an optimal solution to the LP problem

$$\begin{aligned} &\text{minimize} && (\gamma^{(\tau)})^{\top} \boldsymbol{p} && (15) \\ &\text{subject to} && \boldsymbol{p} \in \mathcal{P}. && (16) \end{aligned}$$

Set $\mathcal{P}^{(\tau+1)} = \mathcal{P}^{(\tau)} \cup \{\widetilde{\boldsymbol{p}}_\tau\}$. Increment τ and go back to Step 1.

6.3. Solution of the restricted master problem. Standard optimality conditions (Bertsekas, 1999, Chapter 5) imply that $(\boldsymbol{p}^{(\tau)}, \alpha^{(\tau)})$ together with the nonnegative dual variable $\boldsymbol{\mu}^{(\tau)} = (\mu_0^{(\tau)}, \ldots, \mu_\kappa^{(\tau)})$ satisfy the Langrangian optimality condition

$$(\boldsymbol{p}^{(\tau)}, \alpha^{(\tau)}) \in \arg \min_{\substack{\boldsymbol{p} \in \mathrm{conv}(\mathcal{P}^{(\tau)}) \\ \alpha \in \mathbb{R}}} \mathcal{L}(\boldsymbol{p}, \alpha, \boldsymbol{\mu}^{(\tau)}),$$

where

$$\mathcal{L}(\boldsymbol{p},\alpha,\boldsymbol{\mu}) = \alpha + \sum_{j=0}^{\kappa} \mu_j (f_j(\boldsymbol{p}) - \alpha)$$

constitutes the Lagrangian. Clearly, the complementary slackness conditions imply that μ_j can be positive only when the j-th constraint in (11) is active, i.e., $f_j(\boldsymbol{p}^{(\tau)}) = \alpha^{(\tau)}$. What is more, we necessarily have $\sum_{j=0}^{\kappa} \mu_j^{(\tau)} = 1$ and this condition, together with the nonnegativity of the multipliers, defines the canonical simplex in $\mathbb{R}^{\kappa}$ as the domain of the dual function.

The typically low-dimensional nonlinear RMP can be solved using fast Newton-like methods, such as SQP, and their implementations usually return, as a by-product, the values of the required Lagrange multipliers $\mu_j^{(\tau)}$.

Note that if $\mathcal{P}^{(\tau)} = \{\boldsymbol{p}^1, \ldots, \boldsymbol{p}^r\}$, the RMP takes the following form: Find a sequence of weights $\boldsymbol{q}^{(\tau)} = (q_1^{(\tau)}, \ldots, q_r^{(\tau)})$ and a scalar $\alpha^{(\tau)}$ so as to minimize α subject to

$$f_j(\boldsymbol{N}\boldsymbol{q}) \le \alpha, \quad j = 0, \ldots, \kappa,$$
$$(\boldsymbol{q}, \alpha) \in \mathcal{S}_r \times \mathbb{R},$$

where

$$\boldsymbol{N} = \left[\begin{array}{c|c|c} \boldsymbol{p}^1 & \ldots & \boldsymbol{p}^r \end{array} \right].$$

Having found $\boldsymbol{q}^{(\tau)}$ and $\alpha^{(\tau)}$, we set $\boldsymbol{p}^{(\tau)} = \boldsymbol{N}\boldsymbol{q}^{(\tau)}$.

The formulae for the gradients and Hessians of f_j, which are needed in an efficient implementation of this step using general-purpose NLP solvers, are given in Appendix C.

6.4. Optimality conditions. The termination conditions (12)–(14) can be easily deduced from the Kuhn–Karush–Tucker characterization of $\boldsymbol{p}^{\star}$, an optimal solution to Problem 2; see, e.g., the works of Uciński and Patan (2007) or Uciński (2012) for similar derivations. Note that this characterization says that the components of a convex combination of $\nabla f_0(\boldsymbol{p}^{\star}), \ldots, \nabla f_{\kappa}(\boldsymbol{p}^{\star})$ corresponding to weights p_i between 0 and b_i should equal the same value, whereas the ones corresponding to zero and upper bounds b_i should respectively be no less and no greater than this value, respectively. In practice, this condition is easy to check using an LP solver. We merely set the objective function to zero, treat $\boldsymbol{\eta} \in \mathbb{R}^{\kappa+1}$ and $c \in \mathbb{R}$ as variables and (12)–(14) as constraints, and verify whether or not this linear program is solvable.

Note that in this primal LP problem the number of constraints may be by far larger than the number of variables (this is because, typically, the number of active constraints from among (11) is low). Therefore, to significantly reduce the time spent on this step, it is much more reasonable to solve the dual problem. The primal has a finite optimal solution if, and only if, so does the dual. In turn, the primal is infeasible if, and only if, the dual is unbounded.

6.5. Solution of the column generation problem. The simple form of the constraints (16) makes it possible to develop an algorithm to solve this LP problem, which is almost as simple as a closed-form solution. The key idea is to make use of the following assertion being a direct consequence of Lemma 1 by Uciński (2012).

Theorem 2. *A vector $\boldsymbol{q} \in \mathcal{P}$ constitutes a global solution to the problem* (15)–(16) *if, and only if, there exists a scalar ρ such that*

$$\gamma_i^{(\tau)} \begin{cases} \ge \rho & \text{if } p_i = 0, \\ = \rho & \text{if } 0 < p_i < b_i, \\ \le \rho & \text{if } p_i = b_i \end{cases}$$

for $i = 1, \ldots, n$.

We thus see that it is sufficient to pick the consecutive lowest components of $\boldsymbol{\gamma}^{(\tau)}$ and set the corresponding weights p_i as their maximal allowable values b_i. The process is repeated until the sum of the assigned weights exceeds one. Then the value of the last weight which was set in this manner should be corrected so as the sum of the already processed weights be one, and the remaining (i.e., unassigned) weights are then set as zeros. This straightforward scheme is implemented as Algorithm 3. Note that its correctness requires satisfaction of the condition $\boldsymbol{b} \preceq \boldsymbol{1}$, which is by no means restrictive.

Algorithm 3. Algorithm model for solving the CGP.

Step 0. (*Initialization*) Set $j = 0$ and $v^{(0)} = 0$.

Step 1. (*Sorting*) Sort the elements of $\boldsymbol{\gamma}^{(\tau)}$ in nondecreasing order, i.e., find a permutation π on the index set $I = \{1, \ldots, n\}$ such that

$$\gamma_{\pi(i)}^{(\tau)} \le \gamma_{\pi(i+1)}^{(\tau)}, \quad i = 1, \ldots, n-1.$$

Step 2. (*Identification of nonzero weights*)
Step 2.1. If $v^{(j)} + b_{\pi(j+1)} < 1$ then set

$$v^{(j+1)} = v^{(j)} + b_{\pi(j+1)}.$$

Otherwise, go to Step 3.
Step 2.2. Increment j by one and go to Step 2.1.

Step 3. (*Form the ultimate solution*) Set

$$p_{\pi(i)} = \begin{cases} b_{\pi(i)} & \text{for } i = 1, \ldots, j, \\ 1 - v^{(j)} & \text{for } i = j+1, \\ 0 & \text{for } i = j+2, \ldots, n. \end{cases}$$

7. Simulation example

As a nontrivial device to check the viability of the proposed technique, the problem of sensor selection for a distributed parameter system (DPS) will be used (cf. Uciński, 2005; Patan and Kowalów, 2018). Here the response y is defined implicitly as the solution of a partial differential equation (PDE). Specifically, consider the contaminating mobile source identification problem which is of paramount interest in security, environmental and industrial monitoring, or pollution control (Khapalov, 2010; Beddiaf *et al.*, 2016). In a typical scenario, after some chemical contamination has occurred, there is a developing plume of dangerous or toxic material. Its evolution is affected by weather conditions and the surrounding geography. What is more, the contamination source itself can be mobile. Emergency services wish to quickly know where the source of the plume is located and what its velocity is.

To set up a specific scenario of this type, consider the spatiotemporal dynamics of the contaminant dispersion over the spatial domain $\Omega = [0,1]^2$ and the time horizon $T = (0,1]$, which is modelled by the advection-diffusion equation

$$\frac{\partial y}{\partial t} + \boldsymbol{v} \cdot \nabla y - \nabla \cdot (a \nabla y) = u, \tag{17}$$

where $y = y(\boldsymbol{x}, t)$ is the contaminant concentration at spatial point $\boldsymbol{x} \in \Omega$ and time instant $t \in T$, $\boldsymbol{v} = (1,1)$ is the known wind velocity, and ∇ stands for the spatial gradient. In this description $u = u(\boldsymbol{x}, t)$ and $a = a(\boldsymbol{x})$ are the forcing term modelling the contamination source and the diffusion coefficient, respectively.

Equation (17) is closely related to a physical model of the mesoscale atmospheric motion (Jacobson, 1999). It is complemented with the boundary conditions

$$y = 0 \quad \text{on } \partial\Omega^- \times T, \tag{18}$$

$$\frac{\partial y}{\partial \boldsymbol{n}} = 0 \quad \text{on } \partial\Omega^+ \times T, \tag{19}$$

and the initial condition

$$u\big|_{t=0} = 0 \quad \text{in } \Omega. \tag{20}$$

Here $\partial\Omega$ is the boundary of Ω, $\partial\Omega^- = \big\{\boldsymbol{x} \in \partial\Omega : \boldsymbol{v}\cdot\boldsymbol{n} < 0\big\} = \{0\}\times[0,1] \cup [0,1]\times\{0\}$, $\partial\Omega^+ = \big\{\boldsymbol{x} \in \partial\Omega : \boldsymbol{v}\cdot\boldsymbol{n} \geq 0\big\} = \{1\} \times [0,1] \cup [0,1] \times \{1\}$, where $\partial y/\partial \boldsymbol{n}$ stands for the derivative of y in the direction of the outward normal of $\partial\Omega$, $\boldsymbol{n}$.

Both the terms $u = u(\boldsymbol{x}, t)$ and $a = a(\boldsymbol{x})$ are approximated by some functions which are known up to a vector of unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^m$. In the scenario considered, the mobile source with a known emission intensity of 70, initially located at an unknown spatial point $\boldsymbol{z} = (z_1, z_2)$ and moving in uniform motion with constant velocity $\boldsymbol{s}$ parallel to the x_2-axis, i.e., $\boldsymbol{s} = (0, v)$ with unknown v, is described by

$$u(\boldsymbol{x}, t) = 70 \exp(-100\|\boldsymbol{x} - (\boldsymbol{z} + \boldsymbol{s}t)\|).$$

This emulates the action of a mobile pointwise source (Dirac's delta is approximated here by a slender Gaussian function).

In turn, the diffusion coefficient is modelled as a linear-in-parameters function

$$a(\boldsymbol{x}) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_1 x_2 + a_5 x_2^2,$$

where the values of a_0 to a_5 are unknown.

All the unknown coefficients can then be collected in a vector

$$\boldsymbol{\theta} = (a_0, a_1, a_2, a_3, a_4, a_5, z_1, z_2, v),$$

i.e., $m = 9$. We shall use the notation $y(\boldsymbol{x}, t; \boldsymbol{\theta})$ to emphasize the dependence of the solution on specific values of $\boldsymbol{\theta}$. For simulation purposes, the following nominal value of this vector was adopted:

$$\boldsymbol{\theta}^0 = (0.2, -0.05, 0.2, -0.1, 0.05, 0.2, 0.1, 0.1, 0.8).$$

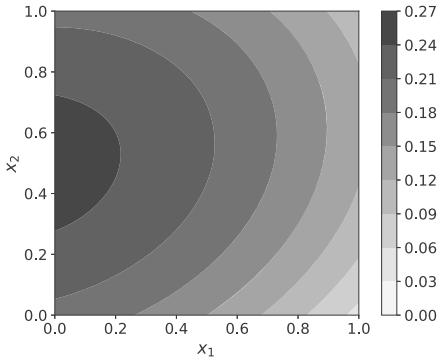
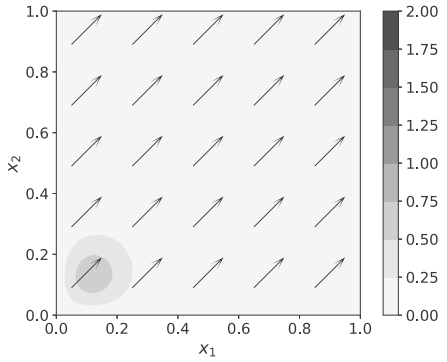


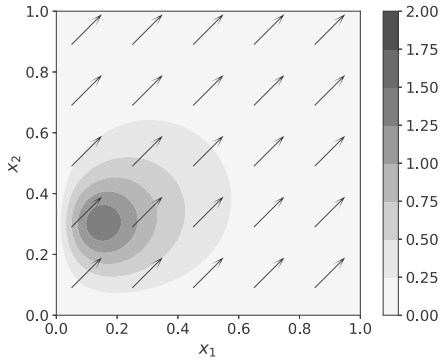
Fig. 1. Isolines of the unknown diffusion coefficient.

Figure 1 shows the isolines of the spatially-varying diffusion coefficient determined by the first three components of $\boldsymbol{\theta}^0$. The evolution of the contaminant concentration is displayed in Fig. 2. The dispersion plume of the contaminant primarily spreads over the entire spatial domain Ω, reflecting a complex combination of advection and diffusion processes. It is strongly influenced by the direction of the wind being the dominant transport factor.

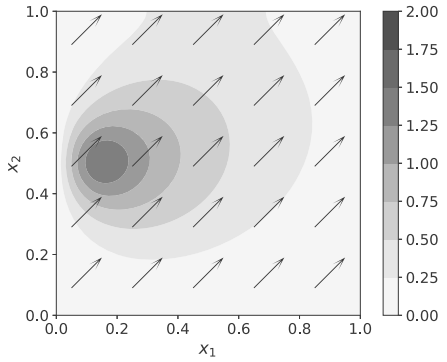
Now assume that a given number $r = 150$ of available sensors can be deployed in $\bar{\Omega} = \Omega \cup \partial\Omega$ to measure the contaminant concentration at a sequence of given time instants $t_\ell = 0.1\ell$, $\ell = 1, \ldots, 10$. This means that the response is one-dimensional ($d = 1$). The goal is



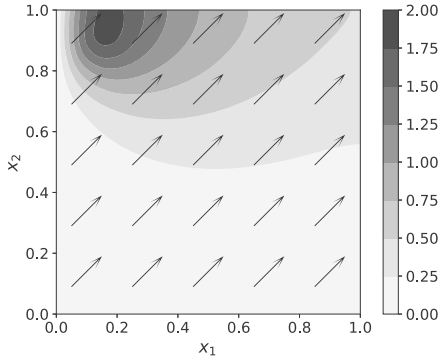
(a) $t = 0.025$



(b) $t = 0.25$



(c) $t = 0.5$



(d) $t = 1.0$

Fig. 2. Isolines of the contaminant concentration at consecutive time instants for a source moving in uniform motion from point $(0.1, 0.1)$ to point $(0.1, 0.9)$. The arrows reflect the wind direction.

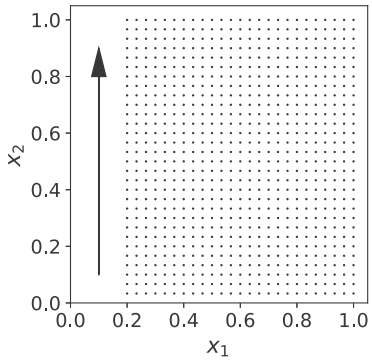


Fig. 3. Trajectory of the contamination source (the arrow) and a finite set of candidate points for sensor location (represented by points).

to use these observations to estimate the parameter vector $\boldsymbol{\theta}$. The model (17)–(20) calibrated in this way can then be used, e.g., to predict the plume envelope evolution.

The sensors usually cannot be placed at arbitrary positions owing to limited access to specific spatial areas. Here we introduce a limitation of this type by assuming that no measurements can be made in the region $[0, 0.2) \times [0, 1]$. The actual sensor location should be selected from among a given finite (but possibly large) set of candidate locations, see Fig. 3. Here there are $n = 750$ candidate locations with coordinates $(0.2 + i/30, j/30)$ for $i = 0, \ldots, 24$ and $j = 1, \ldots, 30$. This selection should be made so as to collect the most valuable information about the unknown parameters as quantified by the $\mathrm{E}_k^{\mathrm{inv}}$-optimality criterion. The measurements are disturbed by uncorrelated noise with zero mean and a constant variance (note that the value of the variance does not influence the sensor locations).

For the design, the elements of the row vector of the sensitivity coefficients $\partial y(\boldsymbol{x}, t; \boldsymbol{\vartheta})/\partial \boldsymbol{\vartheta}$ at admissible sites are indispensable in order to determine matrices $\boldsymbol{M}_i$ which are required to evaluate optimality criteria. As y depends on $\boldsymbol{\theta}$ nonlinearly, the response is linearized with respect to $\boldsymbol{\theta}$ around the nominal vector $\boldsymbol{\theta}^0$. It is easy to check (Uciński, 2005) that this strategy of taking measurements adheres to the framework considered here on setting

$$\boldsymbol{M}_i = \sum_{\ell=1}^{10} \left(\frac{\partial y(\boldsymbol{x}_i, t_\ell; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \right)^{\mathsf{T}} \left(\frac{\partial y(\boldsymbol{x}_i, t_\ell; \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \right) \Bigg|_{\boldsymbol{\vartheta}=\boldsymbol{\theta}^0}.$$

This is a routine in the design for nonlinear response models (Atkinson *et al.*, 2007). The sensitivity coefficients were determined using the direct-differentiation technique (Uciński, 2005). Here it consists in solving a system of ten PDEs in which one equation constitutes the original state equation (17) and the other nine equations result from its differentiation

with respect to the nine components of $\boldsymbol{\theta}$. This system of PDEs was solved using FEniCS v.2019.1.0, an open-source computing platform for solving PDEs using the finite-element method (Langtangen and Logg, 2016).

A uniform triangular mesh in space was employed (Ω was partitioned into 60×60 squares, each partitioned into a pair of triangles). The time derivative was approximated by a simple backward difference with the time step equal to 0.01 (this corresponds to the so-called implicit Euler discretization).

At each candidate point $\boldsymbol{x}_1$ and at each time instant t_ℓ at most one observation may be taken. It is also assumed that each sensor, when placed at a spatial point, is supposed to take a series of measurements at consecutive time moments t_ℓ, $\ell = 1, \ldots, L$. Thus, a weight p_i is associated with each candidate point satisfying the condition

$$0 \le p_i \le \frac{1}{r}.$$

Therefore, the framework of Section 4 applies here after setting $\boldsymbol{b} = (1/r)\mathbf{1}$.

Note that selection of best locations of r sensors from among n candidate points means that we are interested in the solutions in which the weights p_i are either 0 or $1/r$. This, together with the requirement $\boldsymbol{p}^{\mathsf{T}}\mathbf{1}$, would imply that as many as r weights would be nonzero and these would correspond to the best sites for sensor location. But Problem 1 constitutes in fact a relaxed formulation and there is no guarantee that the number of nonzero weights in its optimal solution will not exceed r. As optimal designs for uncorrelated observations usually naturally tend to be sparse (i.e., most of their components are zero) (cf. Atkinson *et al*., 2007), no special sparsity enforcing techniques are needed. Instead, a pre-defined number, N_{rand}, of feasible sensor configurations are drawn at random based on the optimal weights produced by Algorithm 1. Specifically, for each configuration, the locations of individual sensors are drawn without replacement from the set of points with nonzero weights in the optimal solution, with the probabilities being merely the values of the corresponding weights. The configuration yielding the smallest value of the criterion $J(\boldsymbol{p})$ is chosen as the ultimate sensor configuration at this stage. Here we set $N_{\text{rand}} = 100$.

All the algorithms were implemented in Python 3.7 using `NumPy`, its fundamental package for linear algebra, and `SciPy`, its core library used for scientific computing. From the latter, the real Schur decomposition (`scipy.linalg.schur`) was used to produce systems of orthonormal eigenvectors of the current information matrices, the trust-region constrained NLP solver (`scipy.optimize.minimize` with `method = 'trust-constr'`) was used to produce solutions to the RMP problem along with the vector of the Lagrange multipliers $\boldsymbol{\mu}$ needed in Step 3 of Algorithm 2, and the interior-point LP solver (`scipy.optimize.linprog`) was employed to solve the dual LP problem associated with Step 2 of Algorithm 2.

Table 1. Performance of Algorithm 1 and the procedure by Burclová and Pázman (2016) in computing continuous $\mathrm{E}_k^{\mathrm{inv}}$- and E_k-optimum designs, respectively.

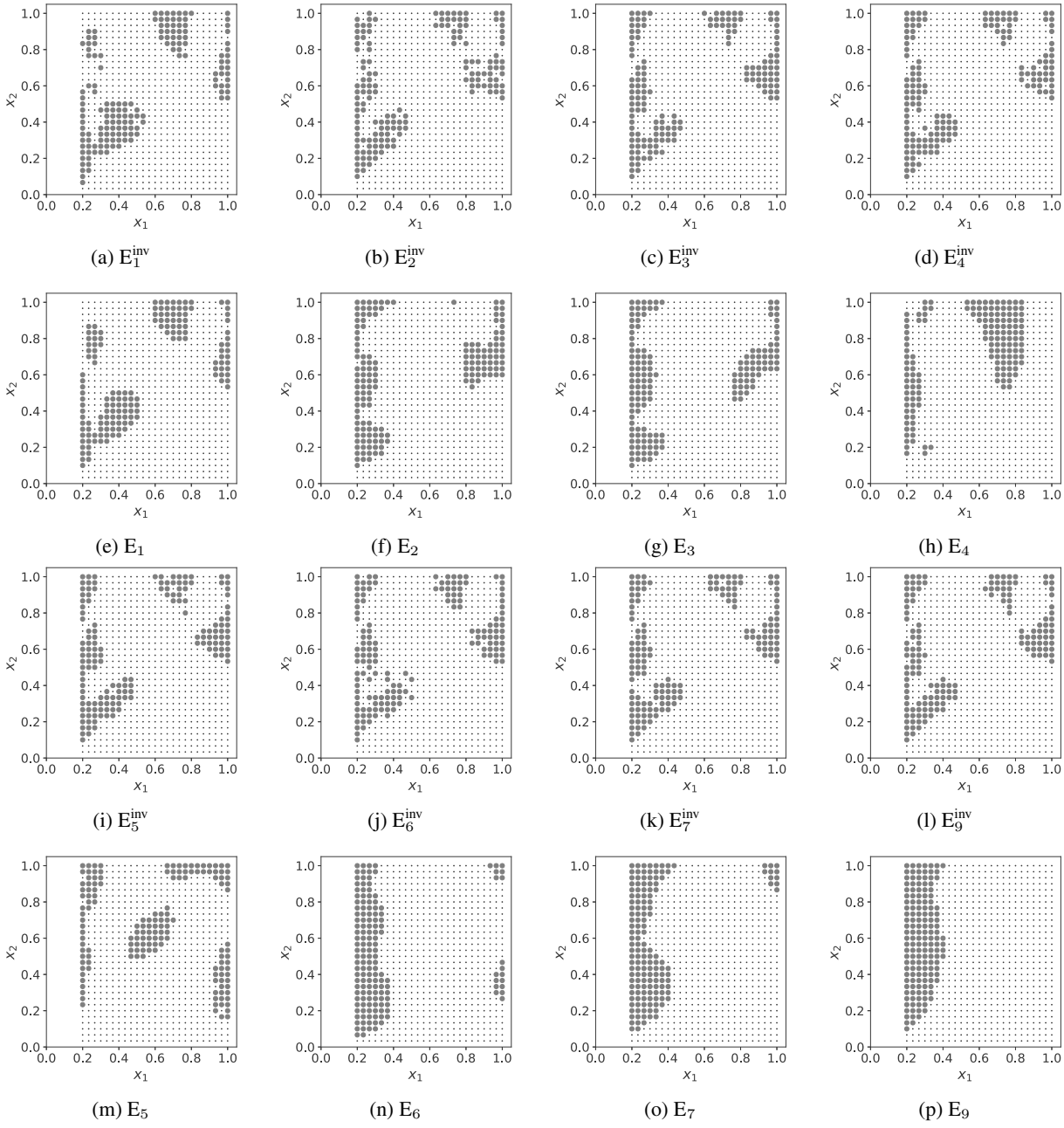
k	CPU time [s]		$\mathrm{eff}_{\mathrm{E}_k^{\mathrm{inv}}}$	# of iterations	
	$\mathrm{E}_k^{\mathrm{inv}}$	E_k		$\mathrm{E}_k^{\mathrm{inv}}$	E_k
1	3.52	30.55	0.99	4	26
2	7.91	22.11	0.82	8	20
3	1.76	10.07	0.81	2	12
4	1.82	9.96	0.60	2	12
5	1.81	15.50	0.59	2	24
6	1.70	3.72	0.55	2	6
7	0.82	3.08	0.60	1	6
8	0.90	1.46	0.31	1	3
9	0.84	1.21	0.23	1	2

The ultimate program was run with the open-source Anaconda distribution 2019.10 under Windows 10 on a laptop equipped with an Intel Core i7–6700HQ CPU, 2.60 GHz, 24 GB RAM.

The value of $\epsilon = 10^{-4}$ was set in the termination condition of Step 3 in Algorithm 1. The penalty coefficient preventing the loss of identifiability was set as $\beta = 0.001$, but it turned out that this value had only a negligible effect on the values of the design criteria. (The contribution of the penalty term to the final value of the design criterion varied from 1.2 to 1.9%.)

Figure 4 displays optimum sensor configurations for both $\mathrm{E}_k^{\mathrm{inv}}$- and E_k-optimality criteria. The latter have been obtained by implementing the LP-based procedure by Burclová and Pázman (2016) using the interior-point solver `scipy.optimize.linprog`. Sensors form a dense cluster along the trajectory of the source and in the area covered by the spreading contaminating plume. Logically, their measurements will be likely to be beneficial for estimation of the initial source position (z_1, z_2) and velocity v. But some part of experimental effort must be also spent on estimation of the diffusion coefficient. This is reflected by the location of a few sensors in the top-right part of Ω, where the diffusion coefficient is lower and its sensitivity to changes in the values of coefficients a_i is more pronounced.

The $\mathrm{E}_k^{\mathrm{inv}}$-optimum sensor configurations keep on slightly varying for moderate values of k but then their changes are only minor. This results from the influence made on the criterion by several dominating eigenvalues of the inverted FIM. Inclusion of smaller eigenvalues to the sum defining this criterion has negligible impact. For the E_k-optimality criterion the number of eigenvalues included in the criterion influences



(a) $\mathrm{E}_1^{\mathrm{inv}}$ (b) $\mathrm{E}_2^{\mathrm{inv}}$ (c) $\mathrm{E}_3^{\mathrm{inv}}$ (d) $\mathrm{E}_4^{\mathrm{inv}}$

(e) E_1 (f) E_2 (g) E_3 (h) E_4

(i) $\mathrm{E}_5^{\mathrm{inv}}$ (j) $\mathrm{E}_6^{\mathrm{inv}}$ (k) $\mathrm{E}_7^{\mathrm{inv}}$ (l) $\mathrm{E}_9^{\mathrm{inv}}$

(m) E_5 (n) E_6 (o) E_7 (p) E_9

Fig. 4. $\mathrm{E}_k^{\mathrm{inv}}$ and E_k-optimum sensor configurations for different k.

the optimal configurations much stronger. It is to be noted, however, that the greater k, the more sensors are clustered along the left boundary of the set of candidate points. But an increase in k is accompanied by much poorer quality of the produced sensor configurations in terms of the $\mathrm{E}_k^{\mathrm{inv}}$-optimality criterion. This is not surprising, since for $k = 1$ both the criteria yield the same optimum design, but for increasing k they start concentrating on completely different aspects of the FIM.

In Table 1 the performance parameters for the algorithms producing continuous $\mathrm{E}_k^{\mathrm{inv}}$ (Algorithm 1) and E_k-optimum (the procedure by Burclová and Pázman (2016)) designs are included. Additionally, the $\mathrm{E}_k^{\mathrm{inv}}$-efficiency of E_k-optimum designs is listed. It is defined as

$$\mathrm{eff}_{\mathrm{E}_k^{\mathrm{inv}}} = \frac{J(\boldsymbol{p}^\star_{\mathrm{E}_k^{\mathrm{inv}}})}{J(\boldsymbol{p}^\star_{\mathrm{E}_k})},$$

where $\boldsymbol{p}^\star_{\mathrm{E}_k^{\mathrm{inv}}}$ and $\boldsymbol{p}^\star_{\mathrm{E}_k}$ stand for $\mathrm{E}_k^{\mathrm{inv}}$- and E_k-optimum designs, respectively. It expresses how close $\boldsymbol{p}^\star_{\mathrm{E}_k}$ is to $\boldsymbol{p}^\star_{\mathrm{E}_k^{\mathrm{inv}}}$ in terms of the $\mathrm{E}_k^{\mathrm{inv}}$-optimality criterion. The larger its value (it is always between 0 and 1), the more $\boldsymbol{p}^\star_{\mathrm{E}_k}$ is

valuable as quantified by the $\mathrm{E}_k^{\mathrm{inv}}$-optimality criterion.

Comparison of CPU times of the algorithms for both the criteria indicates that the procedure for the $\mathrm{E}_k^{\mathrm{inv}}$-optimum designs is substantially faster in spite of the fact that it uses an NLP solver, while the procedure for E_k-optimality exploits only an LP solver. The reason lies in the striking differences in the dimensionalities of these NLP and LP problems. Extremely great efficiency gains result from using GSD, as the number of variables in the NLP problems in the simulations never exceeded ten. In turn, in the LP problem all 750 variables have to take part in numerical optimization. This also makes the version of the algorithm for $\mathrm{E}_k^{\mathrm{inv}}$-optimality an attractive alternative for determining E-optimum designs.

A common observation for both the criteria is that the number of loops in the method of outer approximations decreases as k increases (this may mean that the $\mathrm{E}_k^{\mathrm{inv}}$- and E_k-optimality criteria are easier to optimize for large k).

8. Conclusions

The proposed $\mathrm{E}_k^{\mathrm{inv}}$-optimality criterion and the attendant computational procedure, embodied as Algorithm 1, constitutes a viable alternative to the E_k-optimality criterion and the LP-based method of outer approximations discussed by Harman (2004) as well as Burclová and Pázman (2016). The criterion possesses a much clearer interpretation in terms of the shape of the asymptotic covariance matrix for the estimates. What is more, in spite of a more involved implementation, the gains resulting from combining the method of outer approximations and generalized simplicial decomposition make the scheme outperform the much simpler scheme for E_k-optimality.

There is still room for some improvements. Observe that the sets $\mathcal{Q}^{(\kappa)}$ in Algorithm 1 are ever expanding. As a result, solution of the NLP problem in Step 1 of Algorithm 2 becomes more and more computer-intensive. Thus, the appropriate techniques of dropping least valuable elements of these sets are badly needed. Such techniques exist for SIP problems (see, e.g., Zhang *et al*., 2010), and with some additional effort they seem to be adaptable here, but there are no such methods available for generalized simplicial decomposition. Inclusion of such schemes is of utmost importance for large-scale problems. Their adaptation or elaboration will constitute the main research direction for future research.

Acknowledgment

The author wishes to express his gratitude to the anonymous reviewers for careful reading of the paper and valuable comments that helped to improve the presentation of the paper.

References

Atkinson, A.C., Donev, A.N. and Tobias, R.D. (2007). *Optimum Experimental Designs, with SAS*, Oxford University Press, Oxford.

Beddiaf, S., Autrique, L., Perez, L. and Jolly, J.-C. (2016). Heating source localization in a reduced time, *International Journal of Applied Mathematics and Computer Science* **26**(3): 623–640, DOI: 10.1515/amcs-2016-0043.

Bernstein, D.S. (2005). *Matrix Mathematics. Theory, Facts, and Formulas with Application to Linear Systems Theory*, Princeton University Press, Princeton, NJ.

Bertsekas, D.P. (1999). *Nonlinear Programming*, 2nd Edn, Optimization and Computation Series, Athena Scientific, Belmont, MA.

Bertsekas, D.P. (2015). *Convex Optimization Algorithms*, Athena Scientific, Belmont, MA.

Bertsekas, D. and Yu, H. (2011). A unifying polyhedral approximation framework for convex optimization, *SIAM Journal on Optimization* **21**(1): 333–360.

Böhning, D. (1986). A vertex-exchange-method in D-optimal design theory, *Metrika* **33**(12): 337–347.

Botkin, N.D. and Stoer, J. (2005). Minimization of convex functions on the convex hull of a point set, *Mathematical Methods of Operations Research* **62**(2): 167–18.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*, Cambridge University Press, Cambridge.

Burclová, K. and Pázman, A. (2016). Optimal design of experiments via linear programming, *Statistical Papers* **57**(4): 893–910.

Chepuri, S.P. and Leus, G. (2015). Sparsity-promoting sensor selection for non-linear measurement models, *IEEE Transactions on Signal Processing* **63**(3): 684–698.

Coll, C. and Sánchez, E. (2019). Parameter identification and estimation for stage-structured population models, *International Journal of Applied Mathematics and Computer Science* **29**(2): 327–336, DOI: 10.2478/amcs-2019-0024.

Cook, D. and Fedorov, V. (1995). Constrained optimization of experimental design, *Statistics* **26**: 129–178.

Djelassi, H., Glass, M. and Mitsos, A. (2019). Discretization-based algorithms for generalized semi-infinite and bilevel programs with coupling equality constraints, *Journal of Global Optimization* **75**(2): 341–392.

Duarte, B.P.M., Granjo, J.F.O. and Wong, W.K. (2020). Optimal exact designs of experiments via mixed integer nonlinear programming, *Statistics and Computing* **30**(1): 93–112.

Duarte, B.P.M. and Wong, W.K. (2014). A semi-infinite programming based algorithm for finding minimax optimal designs for nonlinear models, *Statistics and Computing* **24**(6): 1063–1080.

Esteban-Bravo, M., Leszkiewicz, A. and Vidal-Sanz, J.M. (2017). Exact optimal experimental designs with constraints, *Statistics and Computing* **27**(3): 845–863.

Fedorov, V.V. (1989). Optimal design with bounded density: Optimization algorithms of the exchange type, *Journal of Statistical Planning and Inference* **22**: 1–13.

Fedorov, V.V. and Leonov, S.L. (2014). *Optimal Design for Nonlinear Response Models*, CRC Press, Boca Raton, FL.

Harman, R. (2004). Minimal efficiency of designs under the class of orthogonally invariant information criteria, *Metrika* **60**(2): 137–153.

Harman, R. and Benková, E. (2017). Barycentric algorithm for computing d-optimal size- and cost-constrained designs of experiments, *Metrika* **80**(2): 201–225.

Harman, R., Filová, L. and Richtárik, P. (2020). A randomized exchange algorithm for computing optimal approximate designs of experiments, *Journal of the American Statistical Association* **115**(529): 348–361.

Harman, R. and Pronzato, L. (2007). Improvements on removing nonoptimal support points in d-optimum design algorithms, *Statistics & Probability Letters* **77**(1): 90–94.

Harville, D.A. (1997). *Matrix Algebra From a Statistician's Perspective*, Springer-Verlag, New York, NY.

Herzog, R., Riedel, I. and Uciński, D. (2018). Optimal sensor placement for joint parameter and state estimation problems in large-scale dynamical systems with applications to thermo-mechanics, *Optimization and Engineering* **19**(3): 591–627.

Hettich, R. and Kortanek, K.O. (1993). Semi-infinite programming: Theory, methods and applications, *SIAM Review* **35**(3): 380–429.

Jacobson, M.Z. (1999). *Fundamentals of Atmospheric Modeling*, Cambridge University Press, Cambridge.

Joshi, S. and Boyd, S. (2009). Sensor selection via convex optimization, *IEEE Transactions on Signal Processing* **57**(2): 451–462.

Katoh, N. (2001). Combinatorial optimization algorithms in resource allocation problems, *in* C.A. Floudas and P.M. Pardalos (Eds), *Encyclopedia of Optimization*, Vol. 1, Kluwer Academic Publishers, Dordrecht, pp. 259–264.

Khapalov, A.Y. (2010). Source localization and sensor placement in environmental monitoring, *International Journal of Applied Mathematics and Computer Science* **20**(3): 445–458, DOI: 10.2478/v10006-010-0033-3.

Langtangen, H.P. and Logg, A. (2016). *Solving PDEs in Python. The FEniCS Tutorial I*, Springer-Verlag, Cham.

Larsson, T., Migdalas, A. and Patriksson, M. (2015). A generic column generation principle: Derivation and convergence analysis, *Operational Research* **15**(2): 163–198.

Larsson, T., Patriksson, M. and Strömberg, A. (1998). Ergodic convergence in subgradient optimization, *Optimization Methods and Software* **9**(1–3): 93–120.

Lu, Z. and Pong, T.K. (2013). Computing optimal experimental designs via interior point method, *SIAM Journal on Matrix Analysis and Applications* **34**(4): 1556–1580.

Maculan, N., Santiago, C.P., Macambira, E.M. and Jardim, M.H.C. (2003). An $O(n)$ algorithm for projecting a vector on the intersection of a hyperplane and a box in $\mathbb{R}^n$, *Journal of Optimization Theory and Applications* **117**(3): 553–574.

Marshall, A.W., Olkin, I. and Arnold, B.C. (2011). *Inequalities: Theory of Majorization and Its Applications*, 2nd Edn, Springer-Verlag, New York, NY.

Melas, V. (2006). *Functional Approach to Optimal Experimental Design*, Springer-Verlag, New York, NY.

Patan, M. and Kowalów, D. (2018). Distributed scheduling of measurements in a sensor network for parameter estimation of spatio-temporal systems, *International Journal of Applied Mathematics and Computer Science* **28**(1): 39–54, DOI: 10.2478/amcs-2018-0003.

Patan, M. and Uciński, D. (2008). Configuring a sensor network for fault detection in distributed parameter systems, *International Journal of Applied Mathematics and Computer Science* **18**(4): 513–524, DOI: 10.2478/v10006-008-0045-4.

Patan, M. and Uciński, D. (2019). Generalized simplicial decomposition for optimal sensor selection in parameter estimation of spatiotemporal processes, *2019 American Control Conference (ACC), Philadelphia, PA, USA*, pp. 2546–2551.

Patriksson, M. (2001). Simplicial decomposition algorithms, *in* C.A. Floudas and P.M. Pardalos (Eds), *Encyclopedia of Optimization*, Vol. 5, Kluwer Academic Publishers, Dordrecht, pp. 205–212.

Pázman, A. (1986). *Foundations of Optimum Experimental Design*, Mathematics and Its Applications, D. Reidel Publishing Company, Dordrecht.

Polak, E. (1987). On the mathematical foundations of nondifferentiable optimization in engineering design, *SIAM Review* **29**(1): 21–89.

Polak, E. (1997). *Optimization. Algorithms and Consistent Approximations*, Applied Mathematical Sciences, Springer-Verlag, New York, NY.

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

Pronzato, L. and Pàzman, A. (2013). *Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties*, Springer-Verlag, New York, NY.

Pronzato, L. and Zhigljavsky, A.A. (2014). Algorithmic construction of optimal designs on compact sets for concave and differentiable criteria, *Journal of Statistical Planning and Inference* **154**: 141–155.

Pukelsheim, F. (1993). *Optimal Design of Experiments*, Probability and Mathematical Statistics, John Wiley & Sons, New York, NY.

Reemtsen, R. and Görner, S. (1998). Numerical methods for semi-infinite programming: A survey, *in* R. Reemtsen and J.-J. Rückmann (Eds), *Semi-Infinite Programming*, Kluwer Academic Publishers, Boston, MA, pp. 195–275.

Sagnol, G. (2011). Computing optimal designs of multiresponse experiments reduces to second-order cone programming, *Journal of Statistical Planning and Inference* **141**(5): 1684–1708.

Sagnol, G. and Harman, R. (2015). Computing exact D-optimal designs by mixed integer second-order cone programming, *The Annals of Statistics* **43**(5): 2198–2224.

Sahm, M. and Schwabe, R. (2001). A note on optimal bounded designs, *in* A. Atkinson *et al.* (Eds), *Optimum Design 2000*, Kluwer Academic Publishers, Dordrecht, Chapter 13, pp. 131–140.

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

Shimizu, K. and Aiyoshi, E. (1980). Necessary conditions for min-max problems and algorithms by a relaxation procedure, *IEEE Transactions on Automatic Control* **AC-25**(1): 62–66.

Silvey, S.D., Titterington, D.M. and Torsney, B. (1978). An algorithm for optimal designs on a finite design space, *Communications in Statistics—Theory and Methods* **14**: 1379–1389.

Torsney, B. and Mandal, S. (2001). Construction of constrained optimal designs, *in* A. Atkinson *et al.* (Eds), *Optimum Design 2000*, Kluwer Academic Publishers, Dordrecht, Chapter 14, pp. 141–152.

Uciński, D. (2005). *Optimal Measurement Methods for Distributed-Parameter System Identification*, CRC Press, Boca Raton, FL.

Uciński, D. (2012). Sensor network scheduling for identification of spatially distributed processes, *International Journal of Applied Mathematics and Computer Science* **22**(1): 25–40, DOI: 10.2478/v10006-012-0002-0.

Uciński, D. (2015). An algorithm for construction of constrained D-optimum designs, *in* A. Steland *et al.* (Eds), *Stochastic Models, Statistics and Their Applications*, Springer Proceedings in Mathematics & Statistics, Springer-Verlag, Cham, pp. 461–468.

Uciński, D. (2020). D-optimal sensor selection in the presence of correlated measurement noise, *Measurement* **164**: 107873.

Uciński, D. and Patan, M. (2007). D-optimal design of a monitoring network for parameter estimation of distributed systems, *Journal of Global Optimization* **39**(2): 291–322.

Wu, C.-F. (1978). Some algorithmic aspects of the theory of optimal designs, *The Annals of Statistics* **6**(6): 1286–1301.

Yu, Y. (2010). Monotonic convergence of a general algorithm for computing optimal designs, *The Annals of Statistics* **38**(3): 1593–1606.

Yu, Y. (2011). D-optimal designs via a cocktail algorithm, *Statistics and Computing* **21**(3): 475–481.

Zarrop, M.B. and Goodwin, G.C. (1975). Comments on "Optimal inputs for system identification", *IEEE Transactions on Automatic Control* **AC-20**(2): 299–300.

Zhang, L., Wu, S.-Y. and López, M.A. (2010). A new exchange method for convex semi-infinite programming, *SIAM Journal on Optimization* **20**(6): 2959–2977.

Dariusz Uciński was born in 1965. He received his MSc degree in electrical engineering from the Higher College of Engineering in Zielona Góra, Poland, in 1989, and his PhD and DSc degrees in automatic control and robotics from the Wrocław University of Science and Technology, Poland, in 1992 and 2000, respectively. In 2007 he was conferred the full professorial title, the highest scientific degree in Poland. He is currently a professor at the University of Zielona Góra, Poland. His research interests are in the area of measurement optimization for distributed parameter systems. He authored the book entitled *Optimal Measurement Methods for Distributed Parameter System Identification* (CRC Press, 2005). Other areas of his expertise include optimum experimental design, algorithmic optimal control, parallel computing, data analysis and machine learning.

Appendix A

Minimum principle

The following bounds to the sum of largest eigenvalues can be found, e.g., as Fact 8.14.17, p. 328, in the work of Bernstein (2005) or Thm. 21.12.5, p. 556, in that of Harville (1997).

Theorem A1. *Let $\boldsymbol{A} \in \mathbb{S}^m$ with (not necessarily distinct) eigenvalues ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$. Then for any $\boldsymbol{Q} \in \mathbb{R}^{m\times k}$ such that $\boldsymbol{Q}^\mathsf{T}\boldsymbol{Q} = \boldsymbol{I}_k$ (i.e., with orthonormal columns), where $k \leq m$,*

$$\sum_{\ell=m-k+1}^{m} \lambda_\ell \leq \operatorname{trace}(\boldsymbol{Q}^\mathsf{T}\boldsymbol{A}\boldsymbol{Q}) \leq \sum_{\ell=1}^{k} \lambda_\ell.$$

The above two bounds are tight, i.e., equalities hold if the columns of $\boldsymbol{Q}$ are orthonormal eigenvectors of $\boldsymbol{A}$ corresponding to $\lambda_{m-k+1}, \ldots, \lambda_m$ and $\lambda_1, \lambda_2, \ldots, \lambda_k$, respectively.

Appendix B

Proof of Theorem 1

With no loss of generality, we may restrict our attention to the cone $\mathbb{S}^m_{++}$. The extension of these results to $\mathbb{S}^m_{+}$ amounts to incorporating singular matrices, which involves $+\infty$ as the corresponding values of $\mathrm{E}_k^{\mathrm{inv}}$ and is straightforward.

Property (a). Assume that $\boldsymbol{M}_1 \preceq \boldsymbol{M}_2$. Given $\boldsymbol{Q} \in \mathbb{R}^{m\times k}$, the function $\boldsymbol{X} \mapsto \operatorname{trace}(\boldsymbol{Q}^\mathsf{T}\boldsymbol{X}\boldsymbol{Q}) = \operatorname{trace}(\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\boldsymbol{X})$ is matrix nondecreasing on $\mathbb{S}^m$ since $\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T} \succeq \boldsymbol{0}$ (Boyd and Vandenberghe, 2004, p. 109). Therefore, as matrix inversion is matrix-decreasing (Marshall *et al.*, 2011, Fact E.3.b, p. 672), we get

$$\operatorname{trace}(\boldsymbol{Q}^\mathsf{T}\boldsymbol{M}_1^{-1}\boldsymbol{Q}) \geq \operatorname{trace}(\boldsymbol{Q}^\mathsf{T}\boldsymbol{M}_2^{-1}\boldsymbol{Q}).$$

Taking the maximum over all $\boldsymbol{Q} \in \mathcal{Q}$ (note that the compactness of $\mathcal{Q}$ guarantees the existence of the corresponding maxima), we conclude that $\Psi(\boldsymbol{M}_1) \geq \Psi(\boldsymbol{M}_2)$.

Property (b). It is immediate that, given $\boldsymbol{Q} \in \mathcal{Q}$ and $\alpha > 0$, we have $\operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}(\alpha\boldsymbol{M})^{-1}\boldsymbol{Q}) = (1/\alpha)\operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}\boldsymbol{Q})$. Taking the maximum over all $\boldsymbol{Q} \in \mathcal{Q}$ establishes the property.

Property (c). For any fixed $\boldsymbol{Q} \in \mathcal{Q}$, the mapping $\boldsymbol{X} \mapsto \operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{Q})$ is nondecreasing on $\mathbb{S}^m$. This, taken in conjunction with the matrix convexity of the matrix inversion (Marshall *et al.*, 2011, Fact E.7.b, p. 677), implies the convexity of the composition $\boldsymbol{M} \mapsto \operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}\boldsymbol{Q})$ on $\mathbb{S}^m_{++}$. The pointwise maximum of such functions over $\mathcal{Q}$ is convex, which yields our claim.

Property (d). Let $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ be orthonormal, i.e., $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}_m$. It is easy to check that $\mathcal{Q}$ is invariant with respect to premultiplication by $\boldsymbol{U}^{\mathsf{T}}$, i.e., $\mathcal{Q} = \boldsymbol{U}^{\mathsf{T}}\mathcal{Q}$. Consequently,

$$\begin{aligned}
&\Psi(\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}})\\
&= \max_{\boldsymbol{S}^{\mathsf{T}}\boldsymbol{S}=\boldsymbol{I}_k} \operatorname{trace}(\boldsymbol{S}^{\mathsf{T}}(\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathsf{T}})^{-1}\boldsymbol{S})\\
&= \max_{\boldsymbol{S}^{\mathsf{T}}\boldsymbol{S}=\boldsymbol{I}_k} \operatorname{trace}((\boldsymbol{U}^{\mathsf{T}}\boldsymbol{S})^{\mathsf{T}}\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{S})\\
&= \max_{\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{Q}=\boldsymbol{I}_k} \operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}\boldsymbol{Q}) = \Psi(\boldsymbol{M}).
\end{aligned}$$

Appendix C

Gradients and Hessians for the RMP

Let $g(\boldsymbol{p}) = \operatorname{trace}(\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{Q}) - \beta\log\det(\boldsymbol{M}(\boldsymbol{p}))$. We have

$$\begin{aligned}
\frac{\partial g(\boldsymbol{p})}{\partial p_i} &= -\operatorname{trace}\Big(\big(\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}(\boldsymbol{p})\\
&\qquad + \beta\boldsymbol{M}^{-1}(\boldsymbol{p})\big)\boldsymbol{M}_i\Big),\\
\frac{\partial^2 g(\boldsymbol{p})}{\partial p_i \partial p_j} &= \operatorname{trace}\Big(\big(2\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{Q}\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{M}^{-1}(\boldsymbol{p})\\
&\qquad + \beta\boldsymbol{M}^{-1}(\boldsymbol{p})\big)\boldsymbol{M}_i\boldsymbol{M}^{-1}(\boldsymbol{p})\boldsymbol{M}_j\Big).
\end{aligned}$$

Received: 23 May 2020
Revised: 4 October 2020
Accepted: 13 October 2020