



# A sparse control approach to optimal sensor placement in PDE-constrained parameter estimation problems

Ira Neitzel<sup>1</sup> · Konstantin Pieper<sup>2</sup> · Boris Vexler<sup>3</sup> · Daniel Walter<sup>3</sup>

Received: 2 May 2018 / Revised: 2 May 2019 / Published online: 5 September 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

## Abstract

We present a systematic approach to the optimal placement of finitely many sensors in order to infer a finite-dimensional parameter from point evaluations of the solution of an associated parameter-dependent elliptic PDE. The quality of the corresponding least squares estimator is quantified by properties of the asymptotic covariance matrix depending on the distribution of the measurement sensors. We formulate a design problem where we minimize functionals related to the size of the corresponding confidence regions with respect to the position and number of pointwise measurements. The measurement setup is modeled by a positive Borel measure on the spatial experimental domain resulting in a convex optimization problem. For the algorithmic solution a class of accelerated conditional gradient methods in measure space is derived, which exploits the structural properties of the design problem to ensure convergence towards sparse solutions. Convergence properties are presented and the presented results are illustrated by numerical experiments.

**Mathematics Subject Classification** 35R30 · 49K20 · 62K05 · 65K05 · 49M25

---

I. Neitzel is partially supported by CRC 1060 The Mathematics of Emergent Effects funded by the Deutsche Forschungsgemeinschaft. K. Pieper acknowledges funding by the US Department of Energy Office of Science grant DE-SC0016591 and by the US Air Force Office of Scientific Research Grant FA9550-15-1-0001. D. Walter acknowledges support by the DFG through the International Research Training Group IGDK 1754 “Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures”. Furthermore, support from the TopMath Graduate Center of TUM Graduate School at Technische Universität München, Germany and from the TopMath Program at the Elite Network of Bavaria is gratefully acknowledged.

---

✉ Daniel Walter  
walter@ma.tum.de

Extended author information available on the last page of the article

## 1 Introduction

In this paper we propose a measure-valued formulation for the optimal design of a measurement setup for the identification of an unknown parameter vector entering a system of partial differential equations. Many applications in physics, medicine or chemical engineering rely on complex mathematical models as a surrogate for real-life processes. Typically the arising equations contain unknown (material) parameters which have to be identified in order to obtain a realistic model for the simulation of the underlying phenomenon. To illustrate the ideas, we consider a similar example as presented in [10]. Here, the combustion process of a single substance on a two dimensional domain  $\Omega$  is modeled by a non-linear convection-diffusion equation with an Arrhenius-type reaction term, depending on four scalar parameters  $D$ ,  $E$ ,  $d$ , and  $c$ , representing its material properties:

$$-\Delta y + \alpha \cdot \nabla y + D \exp \{-E/(d - y)\} y(c - y) = 0 \quad \text{in } \Omega, \quad (1.1)$$

together with  $y = \hat{y}$  on an inflow boundary  $\Gamma_{\text{in}} \subset \partial\Omega$  and  $\partial_n y = 0$  on  $\partial\Omega \setminus \Gamma_{\text{in}}$ . While  $c$  and  $d$  are known physical constants, the pre-exponential factor  $D$  and the activation energy  $E$  are empirical and cannot be measured directly. Therefore one often has to rely on experimental data, for instance measurements of the mole fraction  $y$ . An estimate for the true parameters is then obtained by finding a parameter vector matching the collected data, which leads to a least-squares problem constrained by a partial differential equation. However, due to errors in the measurement process the obtained estimate is biased and could be far from the value which describes the physical process most accurately. This bias has to be quantified and the measurement procedure has to be adapted to mitigate the influence of the perturbed data.

In this manuscript, we consider a general PDE-model based on a parameter-dependent weak formulation with an unknown parameter vector  $q$  in an admissible set  $Q_{ad} \subset \mathbb{R}^d$  (for instance,  $q = (D, E) \in \mathbb{R}^2$  for (1.1)). We refer to Sect. 2.1 for the precise assumptions. The parameter is estimated from point-wise observations of the solution  $y = S[q]$  of the PDE-model at points  $\{x_j\}_{j=1}^m \subset \Omega_o$ , where  $\Omega_o \subset \bar{\Omega}$  is a closed set covering the possible observation locations. We choose optimal designs according to criteria based on a linearization of the model equation. To this purpose, we define the associated sensitivities  $\{\partial_k S[\hat{q}]\}_{k=1}^n$  of  $S[\hat{q}]$  with respect to perturbations of each parameter  $q_k$ ,  $k = 1, \dots, n$  at an initial guess  $\hat{q} \in Q_{ad}$ , stemming either from prior knowledge or obtained from previous experiments. We note that optimal design approaches based on first-order approximations have been studied for and successfully applied to ordinary differential equations [5], differential-algebraic equations [8], and also partial differential equations [27]. To each measurement location  $x_j$  we assign a positive scalar  $\lambda_j$  which is proportional to the quality of the sensor at this location (or, alternatively corresponds to the number of repeated measurements performed with an identical sensor). Associated to the measurement setup is the design measure

$$\omega(x, \lambda) = \sum_{j=1}^m \lambda_j \delta_{x_j}, \quad (1.2)$$

given by a weighted sum of Dirac delta functions. To quantify the quality of a given measurement setup  $\omega$ , we introduce the Fisher information matrix  $\mathcal{I}(\omega)$  with entries

$$\mathcal{I}(\omega)_{kl} = \int_{\Omega_o} \partial_k S[\hat{q}](x) \partial_l S[\hat{q}](x) d\omega(x), \quad k, l \in \{1, \dots, n\}. \quad (1.3)$$

Furthermore, by  $\Psi$  we denote a scalar quality criterion, which is a positive, smooth, and convex functional acting on the symmetric, positive-definite matrices. Examples for possible choices of  $\Psi$  can be found in, e.g., [36, 42]; see also Sect. 3.1. We consider optimal designs given by the solutions to the optimization problem

$$\min_{x_j \in \Omega_o, \lambda_j \geq 0, j=1, \dots, m} \Psi(\mathcal{I}(\omega(x, \lambda)) + \mathcal{I}_0) + \beta \sum_{j=1}^m \lambda_j, \quad (1.4)$$

where  $\mathcal{I}_0$  is a nonnegative-definite matrix (e.g.,  $\mathcal{I}_0 = 0$ ). It can be interpreted as a priori knowledge on the distribution of the estimator, which may be obtained from previously collected data, for instance in the context of sequential optimal design; cf. [31]. Here, we would choose  $\mathcal{I}_0 = \mathcal{I}(\omega_{\text{old}})$  where the design measure  $\omega_{\text{old}}$  describes the previous experiments. Alternatively, we may adopt a Bayesian viewpoint and consider  $\mathcal{I}_0$  as the covariance matrix of a Gaussian prior. The last term involving the cost parameter  $\beta > 0$  takes into account the overall cost of the measurement process. For other optimal design approaches with sparsity promoting regularization we refer to, e.g., [2, 18, 26]. We emphasize that we neither impose any restrictions on the number of measurements nor restrict the set of candidate locations for the sensors to a finite set.

At first glance, problem (1.4) is a non-convex problem due to the parameterization in terms of the points  $x_j$ , and has a combinatorial aspect due to the unknown number of measurements  $m$ . However, we can bypass these difficulties by embedding the problem into a more general abstract formulation: introducing the set of positive Borel measures  $M^+(\Omega_o)$  on  $\Omega_o$  we determine an optimal design measure from

$$\min_{\omega \in M^+(\Omega_o)} \Psi(\mathcal{I}(\omega) + \mathcal{I}_0) + \beta \|\omega\|_{M(\Omega_o)}, \quad (P_\beta)$$

where  $\|\omega\|_{M(\Omega_o)}$  is the canonical total variation norm. While it is clear that  $(P_\beta)$  is a more general formulation than (1.4), it can be shown that it always admits solutions of the form  $\omega = \sum_{j=1}^m \lambda_j \delta_{x_j}$  for some  $n \leq m \leq n(n+1)/2$ , making both problems essentially equivalent; see Sect. 3.1. We give a derivation of (1.4) and its connection to  $(P_\beta)$  in Sect. 2.

As an alternative to the penalization term  $\beta \|\omega\|_{M(\Omega_o)}$  in  $(P_\beta)$  it is possible to consider a fixed budget for the experiment leading to

$$\min_{\omega \in M^+(\Omega_o)} \Psi(\mathcal{I}(\omega) + \mathcal{I}_0) \quad \text{subject to} \quad \|\omega\|_{M(\Omega_o)} \leq K, \quad (P^K)$$

where  $K > 0$  denotes the overall maximal cost of the measurements. Under certain conditions on  $\Psi$  it can be shown that the inequality constraint in  $(P^K)$  is attained

for every optimal design; see Proposition 3.11. This relates  $(P^K)$  closely to the concept of approximate designs introduced by Kiefer and Wolfowitz in [30] for general linear-regression, where possible experiments are modeled by the probability measures on  $\Omega_\phi$ . We refer also to [3,22,24,33,36] for the analysis of this kind of optimal design formulations. For the adaptation of this approach to parameter estimation in distributed systems we refer to [6,42]. Both formulations,  $(P_\beta)$  and  $(P^K)$ , are closely linked (see Sect. 3.1): On the one hand, in the case of no a priori knowledge on the prior covariance, i.e. for  $\mathcal{I}_0 = 0$ , the solutions of both problems coincide up to a scalar factor, depending on either  $K$  or  $\beta$ . On the other hand, incorporating a priori knowledge, both problem formulations parameterize the same solution manifold. The parameters  $\beta$  and  $K$ , respectively, provide some indirect control over the number of measurements, which is the cardinality of the support of the optimal solution, in this case.

This paper is concerned with the analysis of  $(P_\beta)$  and its efficient numerical solution. There exists a large amount of literature on the solution of  $(P^K)$  by sequentially adding new Dirac delta functions to a sparse initial design measure. A description and proofs of convergence for several variants of these kind of methods can be found in, e.g., [22,45] for the special case of  $\Psi(\cdot) = \det((\cdot)^{-1})$ . These methods correspond to a conditional gradient, or Frank–Wolfe [25], algorithm for minimizing the smooth functional  $\Psi(\mathcal{I}(\cdot))$  over the ball with radius  $K$  in  $M^+(\Omega_\phi)$ . Despite the ease of implementation the proposed methods suffer from some drawbacks. On the one hand the speed of convergence is slow. Recently, in [13] a sub-linear  $\mathcal{O}(1/k)$  rate of convergence for the error in the objective function in terms of the iteration number  $k$  was proven by using an equivalent reformulation of  $(P^K)$  and results for the classical, finite dimensional conditional gradient algorithm; see, e.g., [28]. Note, that without further assumptions on  $\Psi$  than convexity and for example Lipschitz-continuity of its gradient, no better rate than  $\mathcal{O}(1/k)$  can be expected in general; see [19,20].

On the other hand, if only point insertion steps are considered, the support points of the iterates tend to cluster around the optimal ones. To mitigate this effect and accelerate the convergence, several modified variants of the sequential point insertion have been proposed. In [4,39] it is proposed to alternate between point insertion steps and Wolfe's away steps (see [44]) to remove mass from non-optimal points. Heuristically, adjacent support points may be lumped together; see [23]. More recently, several papers suggested to combine the addition of a single Dirac-Delta in each iteration with the solution of a finite-dimensional convex optimization problem and to apply point moving [13] or vertex exchange methods [46]. However, it appears that there is no rigorous approach to guarantee the convergence of the resulting algorithms towards a finitely supported optimal design on the function space level.

In this paper we present a sequential point insertion algorithm for the (non-smooth) optimal design problem  $(P_\beta)$  and prove convergence towards a sparse minimizer of  $(P_\beta)$  comprising at most  $n(n+1)/2$  support points. To this purpose, we adapt the generalized conditional algorithm in measure space presented in [15] for the minimization of a linear-quadratic Tikhonov-regularized problem to our setting. Additionally we incorporate a post-processing step which ensures that the support size of the generated iterates stays uniformly bounded. For further sparsification and a practical acceleration of convergence we propose to alternate between inserting several Dirac delta functions

and point removal steps based on the (approximate) solution of finite-dimensional  $\ell_1$ -regularized sub-problems, which are amenable for semi-smooth Newton methods; see, e.g., [32,43]. A sublinear rate of convergence for the value of the objective function is proven for a wide class of optimality criteria  $\Psi$ ; see Theorem 4.7. Note that we do not employ acceleration strategies based on point moving [13,15], which are difficult to realize since we will employ  $C^0$ -finite elements, which are not continuously differentiable, for the discretization of the underlying PDEs.

The paper is organized as follows: In Sect. 2 we present the optimal design formulation under consideration. In Sect. 3 we introduce notation and state basic existence results for solutions to  $(P_\beta)$  as well as first order optimality conditions. In Sect. 4 the generalized conditional gradient algorithm for the algorithmic solution of  $(P_\beta)$  is proposed and analyzed. Different acceleration and sparsification strategies are presented and a (worst-case) sub-linear convergence rate for the objective functional is proven. The paper is completed by a numerical example given in Sect. 5 to illustrate the theory and show the practical efficiency of the algorithms. In particular, we investigate the effect of the described acceleration strategies.

## 2 From parameter estimation to optimal design

In this section we derive the convex optimal design formulation  $(P_\beta)$  and establish its connection to the non-convex problem (1.4). We start by defining a least-squares estimator for parameter estimation and the notion of the associated linearized confidence domains.

### 2.1 Parameter estimation

Within the scope of this work we consider the identification of a parameter  $q$  entering a weak form  $a(\cdot, \cdot)(\cdot): Q_{ad} \times \hat{Y} \times Y \rightarrow \mathbb{R}$ , which can be non-linear in its first two arguments but is linear in the last one. Here,  $Q_{ad} \subset \mathbb{R}^n$ ,  $n \in \mathbb{N}$ , denotes a set of admissible parameters,  $Y$  denotes a suitable Hilbert space of functions, and  $\hat{Y} = \hat{y} + Y$ , where the function  $\hat{y}$  allows to include non-homogeneous (Dirichlet-type) boundary conditions in the model. For every  $q \in Q_{ad}$  a function  $y = S[q] \in \hat{Y}$  is called the state corresponding to  $q$  if it is a solution to

$$a(q, y)(\varphi) = 0 \quad \forall \varphi \in Y. \quad (2.1)$$

The operator  $S: Q_{ad} \rightarrow \hat{Y}$  mapping a parameter  $q$  to the associated state is called the parameter-to-state operator. For instance, one might think of a Sobolev space defined on an open and bounded Lipschitz domain  $\Omega \subset \mathbb{R}^d$ ,  $d \in \{1, 2, 3\}$  and as  $a(\cdot, \cdot)(\cdot)$  being the weak formulation of an elliptic partial differential operator.

**Remark 2.1** Concretely, in the case of PDE (1.1), we define

$$a(q, y)(\varphi) = (\nabla y, \nabla \varphi)_{L^2} + (\alpha \nabla y, \varphi)_{L^2} + (D \exp \{-E/(d - y)\} y(c - y), \varphi)_{L^2},$$

and  $Y = \{\varphi \in H^1(\Omega) \mid \varphi|_{\Gamma_{\text{in}}} = 0\}$ . Here, the parameter vector is given by  $q = (D, E) \in \mathbb{R}^2$ .

We define a closed set  $\Omega_o \subset \bar{\Omega}$ , on which it is possible to carry out pointwise observations of the state. We make the following general regularity assumption.

**Assumption 1** For every  $q \in Q_{ad}$  there exists a unique solution  $y \in \hat{Y} \cap C(\Omega_o)$  to (2.1). The parameter-to-state mapping  $S$  with

$$S: Q_{ad} \rightarrow C(\Omega_o) \quad \text{with} \quad q \mapsto S[q] = y,$$

is continuously differentiable in a neighborhood of  $Q_{ad}$  in  $\mathbb{R}^n$ . We denote the directional derivative of  $S$  in the direction of the  $k$ th unit vector by  $\partial_k S[q] \in C(\Omega_o)$  and by  $\partial S[q] \in C(\Omega_o, \mathbb{R}^n)$  the vector of partial derivatives.

We emphasize that, under suitable differentiability assumptions on the form  $a(\cdot, \cdot)$  and Assumption 1, the  $k$ th partial derivative  $\delta y_k = \partial_k S[q] \in Y \cap C(\Omega_o)$ ,  $k = 1, \dots, n$ , is the unique solution of the sensitivity equation

$$a'_y(q, y)(\delta y_k, \varphi) = -a'_{q_k}(q, y)(\varphi), \quad \forall \varphi \in Y, \quad (2.2)$$

where  $y = S[q]$  and  $a'_y$  and  $a'_{q_k}$  denote the partial derivatives of the form  $a$  with respect to the state and the  $k$ th parameter; see, e.g., [41].

In the following the exact value of the parameter vector  $q \in \mathbb{R}^n$  appearing in (2.1) is denoted by  $q^*$ . While, for the purposes of analysis we can assume this value to be known, it will be replaced with an appropriate a priori guess in practice. To estimate the parameter  $q$  we consider measurement data  $y_d$  collected at a set of  $m$  disjoint sensor locations  $\{x_j\}_{j=1}^m \subset \Omega_o$ . To formulate the optimal design problem, we assume that the data  $y_d^j \approx S[q^*](x_j)$  is given by the measured response of the model subject to additive measurement errors that are independent and normally distributed; cf., e.g., [7]. Thus, we obtain that

$$y_d^j = S[q^*](x_j) + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, 1/\lambda_j), \quad \text{Cov}(\epsilon_j, \epsilon_i) = 0,$$

for all  $i, j = 1, \dots, m$  and  $j \neq i$ , where the diligence factor  $\lambda_j$  denotes the inverse of the variance of the measurement at the  $j$ th location. We assume that  $\lambda_j$  can be chosen arbitrarily in  $\mathbb{R}_+ \setminus \{0\}$  in the following.

**Remark 2.2** The scalar  $\lambda_j > 0$  corresponds to the reciprocal of the error variance of the measurement taken at  $x_j$ . Thus, it is part of the design model. Since the diligence factors  $\lambda_j$  are also subject to optimization, this interpretation requires some additional discussion. First, assume that all measurements are performed with a given sensor with unit error variance. Furthermore, suppose that taking  $N \in \mathbb{N}$  repeated measurements at the same location is possible. By averaging the obtained measurement data and using the linearity of the variance,  $N$  measurements can be interpreted as a single one with the improved error variance of  $1/N$ . In this light, we can interpret (1.4) as a convex relaxation of a mixed-integer optimization problem for the overall number

of different sensor sites  $m$ , the positions  $x_j$ , and the associated number  $\lambda_j \in \mathbb{N}$  of repeated measurements at this point. Another point of view is to simply assume that performing a single measurement with a given error variance  $1/\lambda_j$  for any  $\lambda_j > 0$  is possible by manufacturing or buying a suitable sensor with precisely this variance.

To emphasize that the data  $y_d$  is a random variable, conditional on the measurement errors, we will write  $y_d(\varepsilon)$  in the following and define the least squares functional

$$J(q, \varepsilon) = \frac{1}{2} \sum_{j=1}^m \lambda_j (S[q](x_j) - y_d^j(\varepsilon))^2 \quad (2.3)$$

as well as the possibly multi-valued least squares estimator

$$\tilde{q}: \mathbb{R}^m \rightarrow \mathcal{P}(\mathbb{R}^n), \quad \tilde{q}(\varepsilon) = \arg \min_{q \in \mathcal{Q}_{ad}} J(q, \varepsilon), \quad (2.4)$$

where  $\mathcal{P}(\mathbb{R}^n)$  denotes the power set of  $\mathbb{R}^n$ . Note that this is the usual maximum likelihood estimator (MLE) using the assumption on the distribution of the measurement errors  $\varepsilon_j$ .

## 2.2 Optimal design

Since the measurement errors are modelled as random variables, the uncertainty in the data is also propagated to the estimator. This means that  $\tilde{q}$  should be interpreted as a random vector. To quantify the bias in the estimation and to assess the quality of computed realizations of the estimator, one considers the non-linear confidence domain of  $\tilde{q}$  defined as

$$D(\tilde{q}, \alpha)(\epsilon) = \left\{ p \in \mathcal{Q}_{ad} \mid J(p, \epsilon) - \min_{q \in \mathcal{Q}_{ad}} J(q, \epsilon) \leq \gamma_n^2(\alpha) \right\}, \quad (2.5)$$

where  $\gamma_n^2(\alpha)$  denotes the  $(1 - \alpha)$ -quantile of the  $\chi^2$ -distribution with  $n$  degrees of freedom; see, e.g., [9, 11]. We emphasize that the confidence domain is a function of the measurement errors and therefore a random variable whose realizations are subsets of the parameter space. In this context, the confidence level  $\alpha \in (0, 1)$  gives the probability that a certain realization of  $D(\tilde{q}(\epsilon), \alpha)(\epsilon)$  contains the true parameter vector  $q^*$ .

Consequently, a good indicator for the performance of the estimator  $\tilde{q}$  is given by the size of its associated confidence domains. The smaller their size, the closer realizations of  $\tilde{q}$  will be to  $q^*$  with high probability. Given a realization  $D(\tilde{q}, \alpha)(\bar{\epsilon})$  of the non-linear confidence domain, its size only depends on the position and the number of the measurements. To obtain a more reliable estimate for the parameter vector, the experiment, i.e. positions  $x_j$  and the measurement weights  $\lambda_j$  should be chosen a priori in such a way that confidence domains of the resulting estimator are small. However, for general models and parameter-to-state mappings  $S$  the estimator

$\tilde{q}$  cannot be given in closed form. Therefore it is generally not possible to provide an exact expression for  $D(\tilde{q}, \alpha)$ .

To circumvent this problem we follow the approach proposed in, e.g., [24, 35] and consider a linearization of the original model around an a priori guess  $\hat{q}$  of  $q^*$  which can stem from historical data or previous experiments. In the following,  $\epsilon \in \mathbb{R}^m$  denotes an arbitrary vector of measurement errors, and  $x \in \mathbb{R}^{d \times m}$ ,  $x = (x_1, \dots, x_m)$ , with  $x_j \in \mathbb{R}^d$ ,  $j = 1, \dots, m$ , stands for the measurement locations. For abbreviation we write  $S[\hat{q}](x) \in \mathbb{R}^m$  for the vector of observations with  $S[\hat{q}](x)_j = S[\hat{q}](x_j)$ ,  $j = 1, \dots, m$ . Moreover the matrices  $X \in \mathbb{R}^{m \times n}$  and  $\Sigma^{-1} \in \mathbb{R}^{m \times m}$  are defined as

$$X_{jk} = \partial_k S[\hat{q}](x_j), \quad \Sigma_{ij}^{-1} = \delta_{ij} \lambda_i, \quad i, j = 1, \dots, m, \quad k = 1, \dots, n,$$

and are assumed to have full rank. We arrive at the linearized least-squares functional

$$J_{\text{lin}}(q, \epsilon) = \frac{1}{2} \sum_{j=1}^m \lambda_j (S[\hat{q}](x_j) + \partial S[\hat{q}](x_j)^\top (q - \hat{q}) - y_d^j(\epsilon))^2,$$

which can be equivalently written as

$$J_{\text{lin}}(q, \epsilon) = \frac{1}{2} \|X(q - \hat{q}) + S[\hat{q}](x) - y_d(\epsilon)\|_{\Sigma^{-1}}^2,$$

where  $\|v\|_{\Sigma^{-1}} = v^\top \Sigma^{-1} v$  for  $v \in \mathbb{R}^m$ . In contrast to the estimator  $\tilde{q}$  (2.4), the associated linearized estimator

$$\tilde{q}_{\text{lin}}: \mathbb{R}^m \rightarrow \mathbb{R}^n, \quad \tilde{q}_{\text{lin}}(\epsilon) = \arg \min_{q \in \mathbb{R}^n} J_{\text{lin}}(q, \epsilon), \quad (2.6)$$

is single-valued and its realizations can be calculated explicitly (see, e.g., [40]), as

$$\tilde{q}_{\text{lin}}(\epsilon) = \hat{q} + (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} (y_d(\epsilon) - S[\hat{q}](x)). \quad (2.7)$$

Due to the assumptions on the noise  $\epsilon$  the estimator  $\tilde{q}_{\text{lin}}$  is a Gaussian random variable with  $\tilde{q}_{\text{lin}} \sim \mathcal{N}(\tilde{q}_{\text{lin}}(0), (X^\top \Sigma^{-1} X)^{-1})$ . The associated realizations of its confidence domain (see, e.g., [11]) are thus given by

$$D(\tilde{q}_{\text{lin}}, \alpha)(\epsilon) = \left\{ \tilde{q}_{\text{lin}}(\epsilon) + (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1/2} \xi, \mid \xi \in \mathbb{R}^m, \|\xi\|_{\mathbb{R}^m} \leq \gamma_n(\alpha) \right\}, \quad (2.8)$$

where  $\|\cdot\|_{\mathbb{R}^m}$  denotes the Euclidean norm. We point out that the linearized confidence domains are ellipsoids in the parameter space centered around  $\tilde{q}_{\text{lin}}$ . Their half axes are given by the eigenvectors of the Fisher-information matrix  $\mathcal{I} = X^\top \Sigma^{-1} X$  with lengths proportional to the associated eigenvalues. Their sizes depend only on the a priori guess  $\hat{q}$  and the setup of the experiment, i.e. the position and total number of measurements, but not on the concrete realization of the measurement noise. Consequently we can



improve the estimator by minimizing the linearized confidence domains as a function of the measurement setup, which leads to (1.4).

To establish the connection to the sparse optimal design approach we observe that the entries of the Fisher-information matrix can be written alternatively as

$$(X^T \Sigma^{-1} X)_{kl} = \sum_{j=1}^m \partial_k S[\hat{q}](x_j) \partial_l S[\hat{q}](x_j) \lambda_j = \int_{\Omega_o} \partial_k S[\hat{q}] \partial_l S[\hat{q}] d\omega = \mathcal{I}(\omega)_{kl}, \quad (2.9)$$

with the design measure  $\omega = \sum_{j=1}^m \lambda_j \delta_{x_j}$ . Furthermore we note that for such a design measure there holds  $\|\omega\|_{M(\Omega_o)} = \sum_{j=1}^m \lambda_j$ . Consequently, for some design criterion  $\Psi$  and prior knowledge  $\mathcal{I}_0$ , the optimal design problem (1.4) can be equivalently expressed as

$$\min_{\omega \in \text{cone}\{\delta_x \mid x \in \Omega_o\}} \Psi(\mathcal{I}(\omega) + \mathcal{I}_0) + \beta \|\omega\|_{M(\Omega_o)}, \quad (2.10)$$

where we minimize the objective functional over all non-negative linear combinations of Dirac delta functions corresponding to points in the observational domain. A priori it is however unclear if this reformulation admits an optimal solution, since the admissible set is not closed in the weak\* topology on  $M(\Omega_o)$ . For a rigorous analysis one therefore has to pass to the closure  $\overline{\text{cone}\{\delta_x \mid x \in \Omega_o\}}^* = M^+(\Omega_o)$ ; see, e.g., [16, Problem 24.C]. As (2.9) suggests, the definition of  $\mathcal{I}$  can be extended to the set of positive regular Borel measures  $M^+(\Omega_o)$ , resulting in the more general problem formulation  $(P_\beta)$ .

**Remark 2.3** In view of Remark 2.2, it may seem reasonable to incorporate upper bounds on the coefficients  $\lambda_j$  into the formulation. This could be motivated either by restricting the maximum number of repeated measurements at the same location (in case the problem arises from a problem with identical sensors and integer  $\lambda_j$  representing the number of measurements) or correspond to a restriction on the variance provided by the best available sensor (e.g., due to manufacturing constraints). Let us briefly discuss this issue. Without restriction, we impose the restriction  $0 \leq \lambda_j \leq 1$ , thus replacing the cone of Dirac delta functions in (2.10) by the set

$$M_{\text{const}}^+(\Omega_o) = \left\{ \omega = \sum_{j=1}^m \lambda_j \delta_{x_j} \mid x_j \neq x_i \text{ for } i \neq j, 0 \leq \lambda_j \leq 1, m \in \mathbb{N} \right\} \\ \subset M^+(\Omega_o).$$

We distinguish two cases: First, let  $\Omega_o$  be the closure of a bounded domain. In this case,  $M_{\text{const}}^+(\Omega_o)$  is not weak\* closed. Indeed, it is straightforward to argue that  $\text{cone}\{\delta_x \mid x \in \Omega_o\} \subset \overline{M_{\text{const}}^+(\Omega_o)}^*$  and consequently  $\overline{M_{\text{const}}^+(\Omega_o)}^* = M^+(\Omega_o)$ , i.e. we again arrive at  $(P_\beta)$ . This stems back to the assumption that measurements at different locations are pairwise uncorrelated. Thus, a measurement with arbitrarily

small variance at a point  $x$  can be approximated by a number of independent measurements with unit variance at distinct points located in a small neighborhood of  $x$ . Second, in the case that  $\Omega_o$  is a collection of a finite number of isolated points, replacing  $M^+(\Omega_o)$  by  $M_{\text{const}}^+(\Omega_o)$  is possible, since the latter is weak\* closed. However, for such  $\Omega_o$  the problem  $(P_\beta)$  can be rewritten as a simpler finite dimensional optimization problem (cf. Sect. 4.2). We do not specifically discuss this case in the following.

### 3 Analysis of the optimal design problem

In the following, we fix the general notation for the remainder of the paper. We consider an observation set  $\Omega_o$  in which we allow measurements to be carried out. It is assumed to be a closed subset of  $\bar{\Omega}$ , which is the closure of the bounded spatial domain  $\Omega \subset \mathbb{R}^d$ . On  $\Omega_o$  we define the space of regular Borel measures  $M(\Omega_o)$  as the topological dual of  $C(\Omega_o)$ , the space of continuous and bounded functions (see, e.g., [21]), with associated duality pairing  $\langle \cdot, \cdot \rangle$ . The norm on  $M(\Omega_o)$  is given by

$$\|\omega\|_{M(\Omega_o)} = \sup_{y \in C(\Omega_o), \|y\|_{C(\Omega_o)} \leq 1} \langle y, \omega \rangle,$$

where  $\|\cdot\|_{C(\Omega_o)}$  is the supremum norm on  $C(\Omega_o)$ . By  $M^+(\Omega_o)$  we refer to the set of positive Borel measures on  $\Omega_o$  (see, e.g., [38, Def. 1.18]),

$$M^+(\Omega_o) = \{ \omega \in M(\Omega_o) \mid \langle y, \omega \rangle \geq 0, \forall y \in C(\Omega_o), y \geq 0 \},$$

with convex indicator function  $I_{\omega \geq 0}$ . For  $\omega \in M(\Omega_o)$  the support is defined as usual by

$$\text{supp } \omega = \Omega_o \setminus \left\{ \bigcup B \in \mathcal{B}(\Omega_o) \mid B \text{ open, } \omega(B) = 0 \right\},$$

where  $\mathcal{B}(\Omega_o)$  are the Borel subsets of  $\Omega_o$ . Note that the support is a closed set. In case the support is a finite set, we denote its cardinality (or counting measure) by  $\#\text{supp } \omega \in \mathbb{N}$ .

A sequence  $\{\omega_k\} \subset M(\Omega_o)$  is called convergent with respect to the weak\*-topology with limit  $\omega \in M(\Omega_o)$  if  $\langle y, \omega_k \rangle \rightarrow \langle y, \omega \rangle$  for  $k \rightarrow \infty$  for all  $y \in C(\Omega_o)$  indicated by  $\omega_k \rightharpoonup^* \omega$ . Additionally we define the usual Lebesgue spaces of integrable and square integrable functions  $L^1(\Omega_o)$  and  $L^2(\Omega_o)$ , respectively, as well as the usual Sobolev space  $H_0^1(\Omega_o)$  with associated (semi-)norm and inner product; see, e.g., [1]. Furthermore we denote by  $\text{Sym}(n)$ ,  $\text{NND}(n)$ , and  $\text{PD}(n)$  the sets of symmetric, symmetric non-negative definite (also, positive semi-definite), and symmetric positive definite matrices, respectively. On the set of symmetric matrices we consider the inner product  $(A, B)_{\text{Sym}(n)} = \text{Tr}(AB^\top)$  for  $A, B \in \text{Sym}(n)$ , where  $\text{Tr}$  denotes the trace, and the Löwner partial order

$$0 \leq_L A \Leftrightarrow A \text{ is positive semidefinite.}$$

Last, for  $\phi: M(\Omega_o) \rightarrow \mathbb{R} \cup \{\infty\}$  and a convex set  $M \subset M(\Omega_o)$  we define the domain of  $\phi$  over  $M$  as

$$\text{dom}_M \phi = \{\omega \in M \mid \phi(\omega) < \infty\},$$

where the index is omitted when  $M = M(\Omega_o)$ .

We consider design criteria of the form  $\Psi(\cdot + \mathcal{I}_0)$ , where  $\mathcal{I}_0 \in \text{NND}(n)$  (e.g.  $\mathcal{I}_0 = 0$ ) incorporates prior knowledge, as described in the introduction. Concerning the function  $\Psi$  the following assumptions are made.

**Assumption 2** The function  $\Psi: \text{Sym}(n) \rightarrow \mathbb{R} \cup \{+\infty\}$  satisfies:

**A1** There holds  $\text{dom } \Psi = \text{PD}(n)$ .

**A2**  $\Psi$  is continuously differentiable for every  $N \in \text{PD}(n)$ .

**A3**  $\Psi$  is non-negative on  $\text{NND}(n)$ .

**A4**  $\Psi$  is lower semi-continuous and convex on  $\text{NND}$ .

**A5**  $\Psi$  is monotone with respect to the Löwner ordering on  $\text{NND}(n)$ , i.e. there holds

$$N_1 \leq_L N_2 \Rightarrow \Psi(N_1) \geq \Psi(N_2) \quad \forall N_1, N_2 \in \text{NND}(n).$$

While Assumptions **(A1)** to **(A4)** are important for the existence of optimal designs and the derivation of first order optimality conditions, Assumption **(A5)** is related to the size of the linearized confidential domains (2.8). Given two design measures  $\omega_1, \omega_2 \in M^+(\Omega_o)$  with  $\mathcal{I}(\omega_1), \mathcal{I}(\omega_1) \in \text{PD}(n)$  and  $\mathcal{I}(\omega_1) \leq_L \mathcal{I}(\omega_2)$  it holds

$$\mathcal{E}_2 = \{\delta q \in \mathbb{R}^n \mid \delta q^\top \mathcal{I}(\omega_2) \delta q \leq r\} \subset \mathcal{E}_1 = \{\delta q \in \mathbb{R}^n \mid \delta q^\top \mathcal{I}(\omega_1) \delta q \leq r\}$$

for any  $r > 0$ . Thus, **(A5)** ensures that  $\Psi$  is a scalar criterion for the size of the linearized confidence ellipsoids (2.8) that is compatible with the inclusion of sets. For a similar set of conditions we refer to [42, p. 41]. The given assumptions can be verified for a large class of classical optimality criteria, among them the A and D criterion

$$\Psi_A(N) = \begin{cases} \text{Tr}(N^{-1}), & N \in \text{PD}(n), \\ \infty, & \text{else,} \end{cases} \quad \Psi_D(N) = \begin{cases} \det(N^{-1}), & N \in \text{PD}(n), \\ \infty, & \text{else,} \end{cases}$$

corresponding to the combined length of the half axis and the volume of the confidence ellipsoids. Additionally, one may also use weighted versions of the design criteria: for instance  $\Psi_A^w(N) = \text{Tr}(WN^{-1}W)$  allows to put special emphasis on particular parameters by virtue of the weight matrix  $W \in \text{NND}(n)$ . However, we emphasize that the results presented in this paper cannot be applied to other non-differentiable popular criteria such as the  $E$  criterion defined by

$$\Psi_E(N) = \begin{cases} \max_i \{\lambda_i(N^{-1})\}, & N \in \text{PD}(n), \\ \infty, & \text{else.} \end{cases}$$

describing the length of the longest half axis and the length of the longest side of the smallest box containing the confidence ellipsoid. In this case, one can for instance resort to smooth approximations of the design criteria.

### 3.1 Existence of optimal solutions to $(P_\beta)$ and optimality conditions

In this section we prove the existence of solutions as well as first order necessary and sufficient optimality conditions for the optimal design problem  $(P_\beta)$ . Additionally, results on the sparsity pattern of optimal designs are derived. First, as canonical extension of (2.9), we introduce the linear and continuous Fisher-operator  $\mathcal{I}$  by

$$\mathcal{I}: M(\Omega_o) \rightarrow \text{Sym}(n), \quad \text{with} \quad \mathcal{I}(\omega)_{k,l} = \langle \partial_k S[\hat{q}] \partial_l S[\hat{q}], \omega \rangle \quad \forall k, l \in \{1, \dots, n\}.$$

It is readily verified that it is the Banach space adjoint of the operator

$$\mathcal{I}^*: \text{Sym}(n) \rightarrow C(\Omega_o), \quad \text{with} \quad \mathcal{I}^*(A) = \varphi_A,$$

where  $\varphi_A \in C(\Omega_o)$  is the continuous function given for  $A \in \text{Sym}(n)$  by

$$\varphi_A(x) = \text{Tr} \left( \partial S[\hat{q}](x) \partial S[\hat{q}](x)^\top A \right) = \partial S[\hat{q}](x)^\top A \partial S[\hat{q}](x) \quad \forall x \in \Omega_o. \quad (3.1)$$

Now, we formulate the reduced design problem  $(P_\beta)$  as

$$\min_{\omega \in M^+(\Omega_o)} F(\omega) = \psi(\omega) + \beta \|\omega\|_{M(\Omega_o)},$$

where  $\psi(\omega) = \Psi(\mathcal{I}(\omega) + \mathcal{I}_0)$ . In the following proposition we collect some properties of the reduced functional.

**Proposition 3.1** *Let Assumptions (A1)–(A5) be fulfilled and let  $\mathcal{I}_0 \in \text{NND}(n)$  be given. The operator  $\mathcal{I}$  and the functional  $\psi$  satisfy:*

1. *For every  $\omega \in M^+(\Omega_o)$  there holds  $\mathcal{I}(\omega) \in \text{NND}(n)$ .*
2. *There holds  $\text{dom}_{M^+(\Omega_o)} \psi = \{ \omega \in M^+(\Omega_o) \mid \mathcal{I}(\omega) + \mathcal{I}_0 \in \text{PD}(n) \}$ .*
3.  *$\psi$  is differentiable with derivative  $\psi'(\omega) = \mathcal{I}^*(\Psi'(\mathcal{I}(\omega) + \mathcal{I}_0)) \in C(\Omega_o)$  for every  $\omega \in \text{dom}_{M^+(\Omega_o)} \psi$ . The derivative can be identified with the continuous function*

$$[\psi'(\omega)](x) = \partial S[q](x)^\top \Psi'(\mathcal{I}(\omega) + \mathcal{I}_0) \partial S[q](x) \leq 0 \quad \forall x \in \Omega_o. \quad (3.2)$$

*Moreover the gradient  $\psi': \text{dom}_{M^+(\Omega_o)} \psi \rightarrow C(\Omega_o)$  is weak\*-to-strong continuous.*

4.  *$\psi$  is non-negative on  $\text{dom}_{M^+(\Omega_o)} \psi$ .*
5.  *$\psi$  is weak\* lower semi-continuous and convex on  $M^+(\Omega_o)$ .*
6.  *$\psi$  is monotone in the sense that*

$$\mathcal{I}(\omega_1) \leq_L \mathcal{I}(\omega_2) \Rightarrow \psi(\omega_1) \geq \psi(\omega_2) \quad \forall \omega_1, \omega_2 \in M^+(\Omega_o).$$

**Proof** To prove the first claim we observe that there holds

$$z^\top \mathcal{I}(\omega)z = \langle (\partial S[\hat{q}]^\top z)^2, \omega \rangle \geq 0 \quad \forall z \in \mathbb{R}^n \quad (3.3)$$

for an arbitrary  $\omega \in M^+(\Omega_o)$ , thus  $\mathcal{I}(\omega) \in \text{NND}(n)$ . Statement 2. follows directly with (A1). For  $\omega \in \text{dom}_{M^+(\Omega_o)} \psi$  the differentiability of  $\psi$  follows from assumption (A2) using the chain rule. We obtain the derivative  $\psi'(\omega) \in M(\Omega_o)^*$  characterized by

$$\langle \psi'(\omega), \delta\omega \rangle_{M^*, M} = \text{Tr}(\Psi'(\mathcal{I}(\omega) + \mathcal{I}_0) \mathcal{I}(\delta\omega)) = \langle \mathcal{I}^* (\Psi'(\mathcal{I}(\omega) + \mathcal{I}_0)), \delta\omega \rangle_{M^*, M},$$

for every  $\delta\omega \in M(\Omega_o)$ , where  $\langle \cdot, \cdot \rangle_{M^*, M}$  denotes the duality pairing between  $M(\Omega_o)$  and its topological dual space. Using the adjoint expression for  $\mathcal{I}$  given in (3.1) we can identify  $\psi'(\omega)$  with the continuous function (3.2). Due to the monotonicity of  $\Psi$  there holds

$$\begin{aligned} \psi'(\omega)(x) &= \langle \psi'(\omega), \delta_x \rangle = \lim_{\tau \rightarrow 0^+} \frac{1}{\tau} [\psi(\omega + \tau \delta_x) - \psi(\omega)] \\ &= \lim_{\tau \rightarrow 0^+} \frac{1}{\tau} \left[ \Psi(\mathcal{I}(\omega) + \mathcal{I}_0 + \tau \partial S[\hat{q}](x) \partial S[\hat{q}](x)^\top) - \Psi(\mathcal{I}(\omega) + \mathcal{I}_0) \right] \\ &\leq 0 \quad \forall x \in \Omega_o. \end{aligned}$$

For a sequence  $\{\omega_k\} \subset \text{dom}_{M^+(\Omega_o)} \psi$  with  $\omega_k \rightharpoonup^* \omega$  it follows from the definition of  $\mathcal{I}$  that  $\mathcal{I}(\omega_k) \rightarrow \mathcal{I}(\omega)$  for  $k \rightarrow \infty$ . Using (3.2), it now follows  $\psi'(\omega_k) \rightarrow \psi'(\omega)$  in  $C(\Omega_o)$ , which shows the continuity of  $\psi'$ . Statements 4., 5., and 6. can be derived directly from Assumptions (A2), (A4), and (A5) using again the continuity of  $\mathcal{I}$ .  $\square$

**Proposition 3.2** Assume that  $\text{dom}_{M^+(\Omega_o)} \psi \neq \emptyset$  and  $\beta > 0$ . Then there exists at least one optimal solution  $\bar{\omega}_\beta$  to (P $_\beta$ ). Moreover the set of optimal solutions is bounded. If  $\Psi$  is strictly convex on PD(n) then the optimal Fisher-information matrix  $\mathcal{I}(\bar{\omega}_\beta)$  is unique.

**Proof** The proof follows standard arguments, using the direct method in variational calculus, using the estimate  $\|\omega\|_{M(\Omega_o)} \leq F(\omega)/\beta$ , the sequential version of the Banach-Alaoglu theorem, and the facts that  $F$  is proper and weak\* lower-semicontinuous. The boundedness of the set of optimal solutions is another direct consequence. Additionally, uniqueness of the optimal Fisher information matrix can be deduced from strict convexity of  $\Psi$  by a direct contradiction argument.  $\square$

**Remark 3.3** The A and D criterion introduced above are strictly convex.

Next we give conditions for the domain of  $\psi$  to be non-empty.

**Proposition 3.4** Assume that  $\beta > 0$  and

$$\mathbb{R}^n = \text{span} \left( \text{Ran } \mathcal{I}_0 \cup \{ \partial S[\hat{q}](x) \mid x \in \Omega_o \} \right).$$

Then there exists at least one optimal solution of (P $_\beta$ ). Furthermore, every design measure  $\omega \in \text{dom}_{M^+(\Omega_o)} \psi$  consists of at least  $n_0 = n - \text{rank } \mathcal{I}_0$  support points.

**Proof** According to Proposition 3.2 we have to show that there exists an admissible design measure. By assumption we can choose a set of  $n - \text{rank } \mathcal{I}_0$  distinct points  $x_j \in \Omega_o$  such that

$$\mathbb{R}^n = \text{span} \left( \text{Ran } \mathcal{I}_0 \cup \left\{ \partial S[\hat{q}](x_j) \mid j = 1, \dots, n - \text{rank } \mathcal{I}_0 \right\} \right).$$

Consequently, setting  $\omega = \sum_{j=1}^{n_0} \delta_{x_j} \in M^+(\Omega_o)$ , we obtain

$$\mathcal{I}(\omega) + \mathcal{I}_0 = \sum_{j=1}^{n_0} \partial S[\hat{q}](x_j) \partial S[\hat{q}](x_j)^\top + \mathcal{I}_0 \in \text{PD}(n),$$

by straightforward arguments. For the last statement we simply observe that for a measure  $\omega$  with less than  $n_0 = n - \text{rank } \mathcal{I}_0$  support points, the associated information matrix  $\mathcal{I}(\omega) + \mathcal{I}_0$  has a non-trivial kernel.  $\square$

By standard results from convex analysis the following necessary and sufficient optimality conditions can be obtained.

**Proposition 3.5** *Let  $\bar{\omega}_\beta \in \text{dom}_{M^+(\Omega_o)} \psi$  be given. Then  $\bar{\omega}_\beta$  is an optimal solution to  $(P_\beta)$  if and only if holds:*

$$\langle -\psi'(\bar{\omega}_\beta), \omega - \bar{\omega}_\beta \rangle + \beta \|\bar{\omega}_\beta\|_{M(\Omega_o)} \leq \beta \|\omega\|_{M(\Omega_o)} \quad \forall \omega \in M^+(\Omega_o). \quad (3.4)$$

**Proof** Since  $F$  is convex, a given  $\bar{\omega}_\beta$  is optimal if and only if

$$0 \in \partial \left( F(\bar{\omega}_\beta) + I_{\omega \geq 0}(\bar{\omega}_\beta) \right),$$

where the expression on the right denotes the subdifferential of  $F + I_{\omega \geq 0}$  at  $\bar{\omega}_\beta$  in  $M(\Omega_o)^*$ . Due to the convexity of  $\beta \|\cdot\|_{M(\Omega_o)} + I_{\omega \geq 0}$  and since  $\psi$  is convex and differentiable at  $\bar{\omega}_\beta$  there holds

$$0 \in \partial F(\bar{\omega}_\beta) = \psi'(\bar{\omega}_\beta) + \partial \left( \beta \|\cdot\|_{M(\Omega_o)} + I_{\omega \geq 0}(\cdot) \right) (\bar{\omega}_\beta),$$

which is equivalent to (3.4).  $\square$

Since the norm as well as the indicator function are positively homogeneous, the subdifferential of  $\beta \|\cdot\|_{M(\Omega_o)} + I_{\omega \geq 0}$  can be characterized further. This yields an equivalent characterization of optimality relating the support points of an optimal design to the set of minimizers of the gradient of  $\psi$  in the optimum.

**Lemma 3.6** *Let  $\bar{\omega}_\beta$  be an optimal solution to  $(P_\beta)$ . Condition (3.4) is equivalent to*

$$-\psi'(\bar{\omega}_\beta) \leq \beta, \quad \text{supp } \bar{\omega}_\beta \subset \{x \in \Omega_o \mid \psi'(\bar{\omega}_\beta)(x) = -\beta\}. \quad (3.5)$$

**Proof** We only give a brief sketch of the proof. Set  $g = \beta \|\cdot\|_{M(\Omega_o)} + I_{\omega \geq 0}$ . Clearly, there holds  $g(\lambda\omega) = \lambda g(\omega)$  for all  $\omega \in M(\Omega_o)$  and  $\lambda \geq 0$ . As a consequence, we obtain from  $-\psi'(\bar{\omega}_\beta) \in \partial g(\bar{\omega}_\beta)$  that

$$-\psi'(\bar{\omega}_\beta) \in \partial g(0) \quad \text{and} \quad \beta \|\bar{\omega}_\beta\|_{M(\Omega_o)} = \langle -\psi'(\bar{\omega}_\beta), \bar{\omega}_\beta \rangle.$$

Due to the non-negativity of  $-\psi'(\bar{\omega}_\beta)$ , the first condition can be equivalently expressed as

$$\langle -\psi'(\bar{\omega}_\beta), \omega \rangle \leq \beta \|\omega\|_{M(\Omega_o)} \quad \forall \omega \in M^+(\Omega_o) \Leftrightarrow -\psi'(\bar{\omega}_\beta)(x) \leq \beta \quad \forall x \in \Omega_o.$$

The condition on the support of  $\bar{\omega}_\beta$  in (3.5) now follows with similar arguments as in [15, Proposition 3].  $\square$

**Remark 3.7** For  $(P^K)$  a similar optimality condition can be derived. A measure  $\bar{\omega}^K \in \text{dom}_{M^+(\Omega_o)} \psi$  is an optimal solution of  $(P^K)$  if and only if

$$\text{supp } \bar{\omega}^K \subset \left\{ x \in \Omega_o \mid \psi'(\bar{\omega}^K)(x) = \arg \min_{x \in \Omega_o} \psi'(\bar{\omega}^K)(x) \right\}, \quad (3.6)$$

where the condition on the support of  $\bar{\omega}^K$  is equivalent to

$$-\langle \psi'(\bar{\omega}^K), \bar{\omega}^K \rangle + \min_{x \in \Omega_o} \psi'(\bar{\omega}^K)(x) \|\bar{\omega}^K\|_{M(\Omega_o)} = 0. \quad (3.7)$$

As for the norm regularized case, we give a short proof of this result. We only derive (3.7). The equivalence to (3.6) then again follows as in [15, Proposition 3]. The measure  $\bar{\omega}^K \in \text{dom}_{M^+(\Omega_o)} \psi$  is optimal for  $(P^K)$  if and only if  $\|\bar{\omega}^K\|_{M(\Omega_o)} \leq K$  and

$$\langle \psi'(\bar{\omega}^K), \bar{\omega}^K \rangle \leq \langle \psi'(\bar{\omega}^K), \omega \rangle \quad \forall \omega \in M^+(\Omega_o), \quad \|\omega\|_{M(\Omega_o)} \leq K.$$

Clearly, since  $\psi'(\bar{\omega}^K)$  is non-positive, this holds if and only if

$$\langle \psi'(\bar{\omega}^K), \bar{\omega}^K \rangle = \inf_{\substack{\omega \in M^+(\Omega_o) \\ \|\omega\|_{M(\Omega_o)} \leq K}} \langle \psi'(\bar{\omega}^K), \omega \rangle = \min_{x \in \Omega_o} \psi'(\bar{\omega}^K)(x) \|\bar{\omega}^K\|_{M(\Omega_o)}.$$

This finishes the proof. Moreover, if  $\min_{x \in \Omega_o} \psi'(\bar{\omega}^K)(x) \neq 0$ , we have  $\|\bar{\omega}^K\|_{M(\Omega_o)} = K$  and optimality of  $\bar{\omega}^K$  is equivalent to

$$0 = \langle \psi'(\bar{\omega}^K), \bar{\omega}^K \rangle - \min_{x \in \Omega_o} \psi'(\bar{\omega}^K)(x) K \leq \langle \psi'(\omega), \omega \rangle - \min_{x \in \Omega_o} \psi'(\omega)(x) K \quad (3.8)$$

for all  $\omega \in \text{dom}_{M^+(\Omega_o)} \psi$  with  $\|\omega\|_{M(\Omega_o)} \leq K$ . For  $K = 1$  the three statements in (3.6), (3.7) and (3.8) yield the well-known Kiefer–Wolfowitz equivalence theorem; see [29,30] and [42, Theorem 3.2].

Since the Fisher-operator  $\mathcal{I}$  is a finite rank operator, uniqueness of the optimal solution is usually not guaranteed. However, the existence of at least one solution with the practically desired sparsity structure is addressed in the following theorem.

**Theorem 3.8** *Let  $\omega \in M^+(\Omega_o)$  be given. Then there exists  $\tilde{\omega} \in M^+(\Omega_o)$  with*

$$\mathcal{I}(\omega) = \mathcal{I}(\tilde{\omega}), \quad \|\tilde{\omega}\|_{M(\Omega_o)} \leq \|\omega\|_{M(\Omega_o)}, \quad \#\text{supp } \tilde{\omega} \leq n(n+1)/2.$$

*Additionally, if there exists an optimal solution to  $(P_\beta)$ , then there exists an optimal solution  $\bar{\omega}_\beta$  with  $\#\text{supp } \bar{\omega}_\beta \leq n(n+1)/2$ .*

In order to prove this statement, we first provide some auxiliary results. For  $m \in \mathbb{N}$  define the cone of measures supported on at most  $m$  points as

$$M_m^+(\Omega_o) = \{\omega \in M^+(\Omega_o) \mid \#\text{supp } \omega \leq m\}.$$

**Lemma 3.9** *Let  $m \in \mathbb{N}$  be given. The set  $M_m^+(\Omega_o)$  is weak\* closed.*

**Proof** Let a weak\* convergent subsequence  $\{\omega_k\} \subset M_m^+(\Omega_o)$  with  $\omega_k \rightharpoonup^* \bar{\omega} \in M^+(\Omega_o)$  be given. For each  $k \in \mathbb{N}$  there exist  $\lambda_j^k \in \mathbb{R}_+$  and  $x_j^k \in \Omega_o$ ,  $j = 1, \dots, m$  with

$$\omega_k = \sum_{j=1}^m \lambda_j^k \delta_{x_j^k} \quad \text{and} \quad \|\omega_k\|_{M(\Omega_o)} = \sum_{j=1}^m \lambda_j^k \leq c$$

with  $c > 0$  independent of  $k \in \mathbb{N}$ . Introducing  $\lambda^k = (\lambda_1^k, \dots, \lambda_m^k)^\top \in \mathbb{R}_+^m$  and  $\mathbf{x}^k = (x_1^k, \dots, x_m^k)^\top \in \Omega_o^m$  there exist a convergent subsequence of  $\{(\mathbf{x}^k, \lambda^k)\}$ , denoted by the same symbol, as well as elements with  $(\mathbf{x}^k, \lambda^k) \rightarrow (\mathbf{x}, \lambda)$ . Define the measure

$$\omega = \sum_{j=1}^m \lambda_j \delta_{x_j} \in M_m^+(\Omega_o) \quad \text{where} \quad \mathbf{x} = (x_1, \dots, x_m)^\top, \quad \lambda = (\lambda_1, \dots, \lambda_m)^\top.$$

Given  $\varphi \in C(\Omega_o)$  there holds  $\varphi(x_j^k) \rightarrow \varphi(x_j)$  as well as  $\lambda_j^k \rightarrow \lambda_j$ ,  $j = 1, \dots, m$ . Since  $\varphi$  is arbitrary, we conclude  $\langle \varphi, \omega^k \rangle \rightarrow \langle \varphi, \omega \rangle$ . Thus there holds  $\omega = \bar{\omega}$  since the weak\* limit is unique. This proves the statement.  $\square$

We require the following lemma, which is a variant of the Carathéodory lemma.

**Lemma 3.10** *Let  $\omega \in M_m^+(\Omega_o)$  for some  $m \in \mathbb{N}$  be given. Furthermore assume that the set  $\{\mathcal{I}(x) \mid x \in \text{supp } \omega\}$  is linearly dependent. Then there exists  $\tilde{\omega} \in M^+(\Omega_o)$  with*

$$\mathcal{I}(\omega) = \mathcal{I}(\tilde{\omega}), \quad \|\tilde{\omega}\|_{M(\Omega_o)} \leq \|\omega\|_{M(\Omega_o)}, \quad \#\text{supp } \tilde{\omega} \leq \#\text{supp } \omega - 1. \quad (3.9)$$

*In particular, given any measure  $\omega \in M_m^+(\Omega_o)$ ,  $m \in \mathbb{N}$ , there is  $\tilde{\omega} \in M^+(\Omega_o)$  fulfilling  $\mathcal{I}(\omega) = \mathcal{I}(\tilde{\omega})$ ,  $\|\tilde{\omega}\|_{M(\Omega_o)} \leq \|\omega\|_{M(\Omega_o)}$  and  $\#\text{supp } \omega \leq n(n+1)/2$ .*



**Proof** Let  $\omega = \sum_{j=1}^m \lambda_j \delta_{x_j}$  be given. Without restriction, assume that  $\lambda_j > 0$  for  $j = 1, \dots, m$ . Define  $\mathcal{I}_j = \mathcal{I}(\delta_{x_j}) \in \text{Sym}(n)$ . By assumption, the set  $\{\mathcal{I}_j\}_{j=1}^m$  is linearly dependent. Thus, we find a nontrivial solution  $\gamma$  of the system of equations  $\sum_{j=1, \dots, m} \gamma_j \mathcal{I}_j = 0$ . By possibly taking the negative of  $\gamma$  we can ensure that  $\sum_{j=1, \dots, m} \gamma_j \geq 0$ . Set

$$\mu = \max_{j=1, \dots, m} \frac{\gamma_j}{\lambda_j} > 0.$$

We define

$$\tilde{\omega} = \omega - \frac{1}{\mu} \sum_{j=1}^m \gamma_j \delta_{x_j} = \sum_{j=1}^m \left(1 - \frac{\gamma_j}{\mu \lambda_j}\right) \lambda_j \delta_{x_j}.$$

The coefficients of the new measure  $\tilde{\omega} = \sum_{j=1}^m \tilde{\lambda}_j \delta_{x_j}$  are given as  $\tilde{\lambda}_j = [1 - \gamma_j/(\mu \lambda_j)] \lambda_j \in \mathbb{R}_+$  since  $\gamma_j/\mu \leq \lambda_j$ . Moreover, we have  $\mathcal{I}(\omega) = \mathcal{I}(\tilde{\omega})$  as well as

$$\|\tilde{\omega}\|_{M(\Omega_o)} = \|\omega\|_{M(\Omega_o)} - \sum_{j=1, \dots, m} \gamma_j/\mu \leq \|\omega\|_{M(\Omega_o)}.$$

The proof of (3.9) is finished with the observation that

$$\tilde{\lambda}_j = 0 \quad \text{for } j \in \arg \max_{j=1, \dots, m} \frac{\gamma_j}{\lambda_j}.$$

For the last statement, we recall that for any  $\omega \in M(\Omega_o)$  it holds  $\mathcal{I}(\omega) \in \text{Sym}(n) \simeq \mathbb{R}^{n(n+1)/2}$ . Thus, if  $\#\text{supp } \omega > n(n+1)/2$ , the set  $\{\mathcal{I}(x) \mid x \in \text{supp } \omega\}$  is linearly dependent. The result can now be proven by induction over the number of support points.  $\square$

**Proof of Theorem 3.8** Let  $\omega \in M^+(\Omega_o)$  be given. There exist sequences  $\{\omega_k\} \subset M^+(\Omega_o)$  with  $\#\text{supp } \omega_k < \infty$ ,  $\omega_k \rightarrow^* \omega$ , and  $\|\omega_k\|_{M(\Omega_o)} \leq \|\omega\|_{M(\Omega_o)}$ . Invoking Lemma 3.10 now yields the existence of a sequence  $\{\tilde{\omega}_k\} \subset M_m^+(\Omega_o)$ , where  $\tilde{m} = n(n+1)/2$ , with  $\mathcal{I}(\tilde{\omega}_k) = \mathcal{I}(\omega_k)$  and  $\|\tilde{\omega}_k\|_{M(\Omega_o)} \leq \|\omega_k\|_{M(\Omega_o)}$  for all  $k \in \mathbb{N}$ . Consequently  $\tilde{\omega}_k$  admits a weak\* convergent subsequence, denoted by the same symbol, with limit  $\tilde{\omega}$ . Moreover,

$$\|\tilde{\omega}_k\|_{M(\Omega_o)} = \langle 1, \tilde{\omega}_k \rangle \rightarrow \langle 1, \tilde{\omega} \rangle = \|\tilde{\omega}\|_{M(\Omega_o)}.$$

Similarly there holds  $\lim_{k \rightarrow \infty} \|\omega_k\|_{M(\Omega_o)} = \|\omega\|_{M(\Omega_o)}$ . Combining these observations we obtain

$$\mathcal{I}(\omega) = \lim_{k \rightarrow \infty} \mathcal{I}(\omega_k) = \lim_{k \rightarrow \infty} \mathcal{I}(\tilde{\omega}_k) = \mathcal{I}(\tilde{\omega}), \quad \|\tilde{\omega}\|_{M(\Omega_o)} \leq \|\omega\|_{M(\Omega_o)}.$$

Since  $\tilde{\omega} \in M_m^+(\Omega_o)$  (see Lemma 3.9) and  $F(\tilde{\omega}) \leq F(\omega)$ , this finishes the proof.  $\square$

In the last part of this section we will further discuss structural properties of solutions to  $(P_\beta)$ , mainly focusing on their connection to  $(P^K)$  and their behaviour for  $\beta \rightarrow \infty$ . In the following, we call a criterion  $\Psi$  strictly monotone with respect to the Löwner ordering, if

$$N_2 - N_1 \in \text{PD}(n) \Rightarrow \Psi(N_1) > \Psi(N_2) \quad \forall N_1, N_2 \in \text{PD}(n).$$

In particular, this is true for the A and D criterion introduced above.

**Proposition 3.11** *The problems  $(P^K)$  and  $(P_\beta)$  are equivalent in the following sense: Given  $K > 0$ , there exists a  $\beta(K) \geq 0$  (not necessarily unique), such that any optimal solution to  $(P^K)$  is an optimal solution to  $(P_\beta)$  and vice-versa.*

Furthermore, assuming that  $\Psi$  is strictly monotone with respect to the Löwner ordering, we additionally obtain the following:

1. We have  $\|\bar{\omega}^K\|_{M(\Omega_o)} = K$  for each optimal solution  $\bar{\omega}^K$  to  $(P^K)$ .
2. There exists a uniquely defined function

$$\beta: \mathbb{R}_+ \setminus \{0\} \rightarrow \mathbb{R}_+ \setminus \{0\}, \quad K \mapsto \beta(K),$$

such that each optimal solution  $\bar{\omega}^K$  to  $(P^K)$  is a minimizer of  $(P_{\beta(K)})$ .

**Proof** We will derive the first result as a consequence of Lagrange duality. Define the Lagrangian as

$$\begin{aligned} L: M(\Omega_o) \times \mathbb{R} &\rightarrow \mathbb{R} \cup \{+\infty\}, \\ L(\omega, \beta) &= \psi(\omega) + I_{\omega \geq 0}(\omega) + \beta (\|\omega\|_{M(\Omega_o)} - K). \end{aligned}$$

Since a Slater condition holds for  $(P^K)$ —there exists a  $\omega \geq 0$  with  $\psi(\omega) < +\infty$  and  $\|\omega\|_{M(\Omega_o)} < K$ —the following strong duality holds (see [12, Theorem 2.165]):

$$\min (P^K) = \min_{\omega \in M(\Omega)} \sup_{\beta \geq 0} L(\omega, \beta) = \max_{\beta \geq 0} \inf_{\omega \in M(\Omega)} L(\omega, \beta) = \max_{\beta \geq 0} [\inf (P_\beta) - \beta K]$$

Therefore, by Lagrange duality (see, e.g., [12, Sect. 2.5.2]), the set of saddle points of the Lagrangian is given precisely by  $(\bar{\omega}_\beta^K, \beta(K))$ , where  $\bar{\omega}_\beta^K$  solves  $(P^K)$  and  $\beta(K)$  solves the dual problem given above. Clearly, saddle points of  $L$  give solutions of  $(P_\beta)$  with  $\beta = \beta(K)$ . Together, this proves the first statement.

Assume that  $\Psi$  is strictly monotone. Let  $\bar{\omega}^K$  be an arbitrary optimal solution to  $(P^K)$  with  $\|\bar{\omega}^K\|_{M(\Omega_o)} < K$ . Using the strict monotonicity of  $\Psi$  we deduce that  $\bar{\omega}^K \neq 0$ . Defining  $\tilde{\omega} = (K/\|\bar{\omega}^K\|_{M(\Omega_o)})\bar{\omega}^K$  there holds  $\psi(\tilde{\omega}) < \psi(\bar{\omega}^K)$  since  $(K/\|\bar{\omega}^K\|_{M(\Omega_o)}) > 1$ . This gives a contradiction and  $\|\bar{\omega}^K\|_{M(\Omega_o)} = K$ . It remains to show that for a given  $K$  the associated Lagrange multiplier denoted by  $\beta(K)$  is positive, unique, and  $\beta(K_1) \leq \beta(K_2)$  if  $K_2 > K_1$ . To prove the positivity, assume that  $\beta(K) = 0$ . Then we obtain

$$L(\bar{\omega}^K, \beta(K)) = \inf_{\omega \in M^+(\Omega_o)} L(\omega, \beta(K)) = \inf_{\omega \in M^+(\Omega_o)} \psi(\omega).$$

Given  $\omega \in \text{dom}_{M^+(\Omega_o)} \psi$ , we have  $\psi(2\omega) < \psi(\omega)$  and consequently the infimum in the equality above is not attained, yielding a contradiction. Assume that  $\beta(K)$  is not unique, i.e. there exist  $\beta_1(K), \beta_2(K) > 0$  such that each optimal solution  $\bar{\omega}^K$  of  $(P^K)$  is also a minimizer of  $L(\cdot, \beta_1(K))$  and  $L(\cdot, \beta_2(K))$  over  $M^+(\Omega_o)$ . First we note again that  $0 \in M^+(\Omega_o)$  is not an optimal solution to  $(P^K)$  due to the strict monotonicity of  $\Psi$ . Additionally it holds  $\|\bar{\omega}^K\|_{M(\Omega_o)} = K$ . Without loss of generality assume that  $\beta_1(K) < \beta_2(K)$ . From the necessary optimality conditions for  $(P_{\beta_1(K)})$  and  $(P_{\beta_2(K)})$ , see (3.5), we then obtain

$$-\psi'(\bar{\omega}^K) \leq \beta_1(K) < \beta_2(K), \quad \text{supp } \bar{\omega}^K \subset \{x \in \Omega_o \mid -\psi'(\bar{\omega}^K)(x) = \beta_2(K)\},$$

implying  $\bar{\omega}^K = 0$  which gives a contradiction.  $\square$

Many commonly used optimality criteria  $\Psi$  are positively homogeneous in the sense that there exists a convex, strictly decreasing, and positive function  $\gamma$  fulfilling

$$\Psi(rN) = \gamma(r) \Psi(N) \quad \forall r > 0, N \in \text{PD}(n); \quad (3.10)$$

cf. also [23, p. 26]. For example, both the A and the D-criterion fulfill this homogeneity with  $\gamma_A$  and  $\gamma_D$  given by

$$\gamma_A(r) = r^{-1}, \quad \gamma_D(r) = r^{-n}.$$

The following lemma illustrates the findings of the previous result, provided that  $\mathcal{I}_0 = 0$ . It turns out that solutions to  $(P^K)$  can be readily obtained by scaling optimal solutions to  $(P_\beta)$ .

**Proposition 3.12** *Assume that  $\mathcal{I}_0 = 0$  and  $\Psi$  is positive homogeneous in the sense of (3.10). Let  $\bar{\omega}_\beta$  be a solution to  $(P_\beta)$  for some fixed  $\beta > 0$ . Then*

$$K \bar{\omega}_\beta / \|\bar{\omega}_\beta\|_{M(\Omega)} \text{ solves } (P^K). \quad (3.11)$$

**Proof** First we note that under the stated assumptions every optimal solution  $\bar{\omega}^K$  to  $(P^K)$  fulfills  $\|\bar{\omega}^K\|_{M(\Omega_o)} = K$ . Clearly, we have

$$\min (P^K) = \min_{\substack{\omega \in M^+(\Omega_o), \\ \|\omega\| = K}} \psi(\omega) = \min_{\substack{\omega' \in M^+(\Omega_o), \\ \|\omega'\| = 1}} \psi(K\omega') = \gamma(K) \min(P^1),$$

by using the positive homogeneity of  $\Psi$ . Thus, the solutions of  $(P^K)$  are given by  $K\omega^1$ , where  $\omega^1$  are solutions of  $(P^1)$ . Now, using the fact that

$$\min (P_\beta) = \min_{K \geq 0} \left[ \min_{\omega' \in M^+(\Omega_o), \|\omega'\| = 1} \psi(K\omega') + \beta K \right] = \min_{K \geq 0} \left[ \gamma(K) \min(P^1) + \beta K \right]$$

the solutions  $\bar{\omega}_\beta$  of  $(P_\beta)$  can be computed as  $\bar{\omega}_\beta = K\omega^1$ , where  $K$  minimizes the above expression and  $\omega^1 \in \arg \min(P^1)$ . Together, this directly implies (3.11).  $\square$

As we have shown in the case  $\mathcal{I}_0 = 0$ , i.e. in the absence of a priori knowledge, the optimal locations of the sensors  $x$  are independent of the cost parameter  $\beta$  (resp.  $K$ ), which only affects the scaling of the coefficients  $\lambda$ . However for  $\mathcal{I}_0 \neq 0$  this is generally not the case. Loosely speaking, if the a priori information is relatively good (i.e.  $\mathcal{I}_0 \in \text{PD}(n)$ ) and the cost per measurement is too high, the optimal design is given by the zero function, i.e. the experiment should not be carried out at all.

**Proposition 3.13** *Let  $\mathcal{I}_0 \in \text{PD}(n)$ . Then the zero function  $\bar{\omega} = 0$  is an optimal solution to  $(P_\beta)$  if and only if  $\beta > \beta_0 = -\min_{x \in \Omega_o} \psi'(0)$ .*

**Proof** We first note that  $0 \in \text{dom } \psi$  and  $\beta_0 = -\min_{x \in \Omega_o} \psi'(0) < \infty$ . Clearly, for  $\beta \geq \beta_0$ , the zero function fulfills the optimality conditions from Lemma 3.6. Thus, it is a solution to  $(P_\beta)$ . Conversely, for  $\beta < \beta_0$ , the optimality conditions are violated.  $\square$

## 4 Algorithmic solution

In this section we will elaborate on the solution of  $(P_\beta)$ . We consider two different approaches. First, we present an algorithm relying on finitely supported iterates and the sequential insertion of single Dirac Delta functions based on results for a linear-quadratic optimization problem in [14] and [15]. We derive all necessary results to prove convergence of the generated sequence of measures towards a minimizer of  $(P_\beta)$  together with a sub-linear convergence rate of the objective function value. Additionally we propose to alternate between point insertion and point deletion steps to benefit the sparsity of the iterates and to speed up the convergence of the algorithm in practice. These sparsification steps are based on the approximate solution of finite dimensional optimization problems in every iteration. As an example we give two explicit realizations for the point removal and discuss the additional computational effort in comparison to an algorithm solely based on point insertion steps. Moreover, we propose a sparsification procedure based on the proof of Theorem 3.8, which ensures that the support size of all iterates is uniformly bounded and guarantees a sparse structure of the computed optimal design.

### 4.1 A generalized conditional gradient method

For the direct solution of  $(P_\beta)$  on the admissible set  $M^+(\Omega_o)$  we adapt the numerical procedure presented in [15], which relies on finitely supported iterates. A general description of the method is given in Algorithm 1. For convenience of the reader we give a detailed description of the individual steps and their derivation. The basic idea behind the procedure relies on a point insertion process (steps 2.–4. in Algorithm 1) related to a generalized conditional gradient method. More precisely, they consist of conditional gradient steps for a surrogate optimization problem with the same optimal solutions, in which the sublinear total variation norm is replaced by a coercive cost term for designs of very large norm. Additionally, we consider the minimization of the finite dimensional subproblem that arises from restriction of the design measure to the active support of the current iterate (in step 5.). This is motivated on the one

hand by the desire to potentially remove non-optimal support points by setting the corresponding coefficient to zero, and on the other hand by the desire to obtain an accelerated convergence behavior in practice.

---

**Algorithm 1** Successive point insertion
 

---

1. Choose  $\omega^1 \in \text{dom}_{M^+(\Omega_o)} \psi$ ,  $\# \text{supp } \omega^1 \leq n(n+1)/2$ . Set  $M_0 = F(\omega^1)/\beta$ .
  - while**  $\Phi(\omega^k) \geq \text{TOL}$  **do**
  2. Compute  $\psi'_k = \psi'(\omega^k)$ . Determine  $\hat{x}^k \in \arg \min_{x \in \Omega_o} \psi'_k(x)$ .
  3. Set  $v^k = \theta^k \delta_{\hat{x}^k}$  with  $\theta^k = \begin{cases} 0, & \psi'_k(\hat{x}^k) \geq -\beta, \\ -(M_0/\beta) \psi'_k(\hat{x}^k), & \text{else} \end{cases}$
  4. Select a step size  $s^k \in (0, 1]$  and set  $\omega^{k+1/2} = (1 - s^k)\omega^k + s^k v^k$ .
  5. Find  $\omega^{k+1}$  with  $\text{supp } \omega^{k+1} \subseteq \text{supp } \omega^{k+1/2}$  and  $F(\omega^{k+1}) \leq F(\omega^{k+1/2})$ .
  - end while**
- 

This section is structured as follows: First, we focus on the point insertion step and its descent properties. By a suitable choice of the step size  $s^k$  in each step of the procedure we are able to prove a sub-linear convergence rate for the objective functional value. Secondly, we consider concrete examples for the point removal step 5.

#### 4.1.1 Convergence analysis

As already pointed out, Algorithm 1 relies on a coercive surrogate design problem which admits the same optimal solutions as  $(P_\beta)$ . Given a constant  $M_0 > 0$ , we start by introducing the auxiliary function  $\varphi_{M_0} : \mathbb{R}_+ \rightarrow \mathbb{R}$  as

$$\varphi_{M_0}(t) = \begin{cases} t, & t \leq M_0, \\ (1/(2M_0)) [t^2 + M_0^2], & \text{else,} \end{cases}$$

and consider the modified problem

$$\min_{\omega \in M^+(\Omega_o)} F_{M_0}(\omega) = \psi(\omega) + \beta \varphi_{M_0}(\|\omega\|_{M(\Omega_o)}) \quad (P_\beta^{M_0})$$

for the special choice of  $M_0 = F(\omega^1)/\beta$ , with arbitrary but fixed  $\omega^1 \in \text{dom}_{M^+(\Omega_o)} \psi$ . Note that, for all  $\omega \in M^+(\Omega_o)$  with  $F(\omega) \leq F(\omega^1)$ , there holds  $\|\omega\|_{M(\Omega_o)} \leq M_0$  and consequently  $F(\omega) = F_{M_0}(\omega)$ . We additionally point out that

$$\varphi_{M_0}(\|\bar{\omega}_\beta\|_{M(\Omega_o)}) = \|\bar{\omega}_\beta\|_{M(\Omega_o)} \quad (4.1)$$

for every optimal solution  $\bar{\omega}_\beta$  of  $(P_\beta^{M_0})$ . Connected to this auxiliary problem we additionally define the primal-dual gap  $\Phi : \text{dom } \psi \rightarrow [0, \infty)$  by

$$\Phi(\omega) = \sup_{v \in M^+(\Omega_o)} [\langle \psi'(\omega), \omega - v \rangle + \beta \|\omega\|_{M(\Omega_o)} - \beta \varphi_{M_0}(\|\omega\|_{M(\Omega_o)})].$$

Note that the value of  $\Phi$  is finite for every  $v \in \text{dom } \psi$ , which follows with the coercivity of  $\varphi_{M_0}(\cdot)$ . In the next proposition we collect several results to establish the connection between the optimal design problems  $(P_\beta)$  and  $(P_\beta^{M_0})$ .

**Proposition 4.1** *Let  $\omega^1 \in \text{dom}_{M^+(\Omega_o)} \psi$  be arbitrary but fixed and set  $M_0 = F(\omega^1)/\beta$ . Given  $\bar{\omega}_\beta \in \text{dom}_{M^+(\Omega_o)} \psi$  the following three statements are equivalent:*

1. *The measure  $\bar{\omega}_\beta$  is a minimizer of  $(P_\beta)$ .*
2. *The measure  $\bar{\omega}_\beta$  is a minimizer of  $(P_\beta^{M_0})$ .*
3. *The measure  $\bar{\omega}_\beta$  fulfils  $\Phi(\bar{\omega}_\beta) = 0$ .*

Furthermore there holds

$$\Phi(\omega) \geq F(\omega) - F(\bar{\omega}_\beta) =: r_F(\omega), \quad (4.2)$$

for all  $\omega \in \text{dom}_{M^+(\Omega_o)} \psi$ ,  $\|\omega\|_{M(\Omega_o)} \leq M_0$  and all minimizers  $\bar{\omega}_\beta$  of  $(P_\beta^{M_0})$ .

**Proof** The equivalence between the first two statements can be proven as in [15]. We only prove the third one. Similar to the proof of (3.4) (see Proposition 3.5) a given  $\bar{\omega}_\beta \in \text{dom}_{M^+(\Omega_o)} \psi$  is a minimizer of  $(P_\beta^{M_0})$  if and only if it fulfills

$$-\langle \psi'(\bar{\omega}_\beta), \omega - \bar{\omega}_\beta \rangle + \beta \varphi_{M_0}(\|\bar{\omega}_\beta\|_{M(\Omega_o)}) \leq \beta \varphi_{M_0}(\|\omega\|_{M(\Omega_o)}) \quad \forall \omega \in M^+(\Omega_o).$$

By reordering and taking the minimum over all  $\omega \in M^+(\Omega_o)$  this can be equivalently written as

$$\sup_{\omega \in M^+(\Omega_o)} [\langle \psi'(\bar{\omega}_\beta), \bar{\omega}_\beta - \omega \rangle + \beta \varphi_{M_0}(\|\bar{\omega}_\beta\|_{M(\Omega_o)}) - \beta \varphi_{M_0}(\|\omega\|_{M(\Omega_o)})] = 0.$$

Utilizing (4.1) we find  $\Phi(\bar{\omega}_\beta) = 0$  if and only if  $\bar{\omega}_\beta$  is a minimizer of  $F_{M_0}$ . It remains to prove (4.2). Given  $\omega \in \text{dom}_{M^+(\Omega_o)} \psi$  with  $\|\omega\|_{M(\Omega_o)} \leq M_0$  and a minimizer  $\bar{\omega}_\beta$  we obtain

$$F(\omega) - F(\bar{\omega}_\beta) \leq \beta \|\omega\|_{M(\Omega_o)} - \beta \|\bar{\omega}_\beta\|_{M(\Omega_o)} + \langle \psi'(\omega), \omega - \bar{\omega}_\beta \rangle, \quad (4.3)$$

by convexity of  $\psi$ . Noting that

$$-[\beta \|\bar{\omega}_\beta\|_{M(\Omega_o)} + \langle \psi'(\omega), \bar{\omega}_\beta \rangle] \leq -\inf_{v \in M^+(\Omega_o)} [\langle \psi'(\omega), v \rangle + \beta \varphi_{M_0}(\|v\|_{M(\Omega_o)})],$$

the right-hand side in (4.3) is estimated by  $\Phi(\omega)$ , which concludes the proof.  $\square$

With the result of the last proposition we may consider a minimization algorithm for  $(P_\beta^{M_0})$  in order to compute optimal solutions to  $(P_\beta)$ . Additionally, the result suggests the use of  $\Phi$  as a convergence criterion, since it gives an upper bound for the residual error in the objective function value. As can be seen below, the evaluation of  $\Phi$  can be easily computed as a by-product of steps 2.–3. in Algorithm 1.

The algorithm operates on finitely supported iterates  $\omega^k = \sum_{i=1}^{m_k} \lambda_i^k \delta_{x_i^k}$  with distinct support points  $x_i^k \in \Omega_o$  and positive coefficients  $\lambda_i^k$ ,  $i \in \{1, \dots, m_k\}$ ,  $m_k \in \mathbb{N}$ . In steps 2.–4. the intermediate iterate  $\omega^{k+1/2}$  is obtained as a convex combination between the previous iterate  $\omega^k$  and a scaled Dirac delta function  $\theta^k \delta_{\hat{x}^k}$  inserted at the global minimum of the gradient  $\psi'(\omega^k)$ . The initial coefficient  $\theta^k$  is determined by the current maximal violation of the lower bound on the gradient of  $\psi$ ; see (3.5). In the following lemma we relate this definition to the computation of a descent direction in the context of a generalized conditional gradient method (cf. [14,15,37]) for the auxiliary problem  $(P_\beta^{M_0})$ .

**Lemma 4.2** *Let  $\omega^k \in \text{dom}_{M^+(\Omega_o)} \psi$  be given. Then the measure  $v^k = \theta^k \delta_{\hat{x}^k}$  with  $\hat{x}^k \in \Omega_o$  and  $\theta^k \geq 0$  as defined in steps 2.–3. of Algorithm 1 is a minimizer of*

$$\min_{v \in M^+(\Omega_o)} \langle \psi'(\omega^k), v \rangle + \beta \varphi_{M_0}(\|v\|_{M(\Omega_o)}). \quad (P_\beta^{\text{lin}})$$

Moreover,  $v^k$  realizes the supremum in the definition of the primal-dual gap: it holds  $\Phi(\omega^k) = \langle \psi'(\omega^k), \omega^k - v^k \rangle + \beta \|\omega^k\|_{M(\Omega_o)} - \beta \varphi_{M_0}(\|v^k\|_{M(\Omega_o)})$ .

**Proof.** We note that  $(P_\beta^{\text{lin}})$  can be equivalently expressed as

$$\min_{\theta \in [0, \infty)} \theta \min_{\substack{\tilde{v} \in M^+(\Omega_o), \\ \|\tilde{v}\|_{M(\Omega_o)} = 1}} \langle \psi'(\omega^k), \tilde{v} \rangle + \beta \varphi_{M_0}(\theta) \quad (4.4)$$

Due to  $\psi'(\omega^k) \leq 0$  and  $\theta \geq 0$ , a solution to the inner minimization problem is given by  $\tilde{v}^k = \delta_{\hat{x}^k}$  with  $\hat{x}^k \in \arg \min_{x \in \Omega_o} \psi'(\omega^k)(x)$ . In fact we have

$$\langle \psi'(\omega^k), \tilde{v} \rangle \geq \langle \psi'(\omega^k), \tilde{v}^k \rangle = \min_{x \in \Omega_o} \psi'(\omega^k)(x) \quad \forall \tilde{v} \in M^+(\Omega_o), \|\tilde{v}\|_{M(\Omega_o)} = 1.$$

Thus problem (4.4) reduces to

$$\min_{\theta \in [0, \infty)} \theta \min_{x \in \Omega_o} \psi'(\omega^k)(x) + \beta \varphi_{M_0}(\theta).$$

By straightforward calculations, we verify that  $\theta^k$  as defined in step 2. of Algorithm 1 is a minimizer of this problem. We conclude that  $v^k = \theta^k \tilde{v}^k$  is a solution of  $(P_\beta^{\text{lin}})$ . This finishes the proof of the first statement. Moreover, the second statement follows due to

$$\begin{aligned} \Phi(\omega^k) &= \sup_{v \in M^+(\Omega_o)} \left[ \langle \psi'(\omega^k), \omega - v \rangle + \beta \|\omega^k\|_{M(\Omega_o)} - \beta \varphi_{M_0}(\|v\|_{M(\Omega_o)}) \right] \\ &= \langle \psi'(\omega^k), \omega \rangle + \beta \|\omega^k\|_{M(\Omega_o)} - \min_{v \in M^+(\Omega_o)} \langle \psi'(\omega^k), v \rangle + \beta \varphi_{M_0}(\|v\|_{M(\Omega_o)}) \\ &= \langle \psi'(\omega^k), \omega^k - v^k \rangle + \beta \|\omega^k\|_{M(\Omega_o)} - \beta \varphi_{M_0}(\|v^k\|_{M(\Omega_o)}). \end{aligned}$$

□

**Remark 4.3** At this point, replacing  $(P_\beta)$  by the equivalent formulation  $(P_\beta^{M_0})$  is crucial. In fact, the partially linearized problem corresponding to the original problem

$$\min_{v \in M^+(\Omega_o)} \langle \psi'(\omega), v \rangle + \beta \|v\|_{M(\Omega_o)},$$

is either unbounded or has an unbounded solution set in the case that  $\min_{x \in \Omega_o} \psi'(\omega) \leq -\beta$ .

Note that, as a by-product of the last result, the convergence criterion  $\Phi(\omega^k)$  can be evaluated cheaply, once the current gradient  $\psi'(\omega^k)$  and its minimum point are calculated.

We form the intermediate iterate as convex combination between the old iterate and the new sensor i.e.,  $\omega^{k+1/2} = (1 - s^k)\omega^k + s^k v^k$ , where  $s^k \in (0, 1]$  is suitably chosen. This ensures  $\omega^{k+1/2} \in M^+(\Omega_o)$ . The step size  $s^k$  will be chosen by the following generalization of the well-known Armijo-Goldstein condition; see, e.g., [14]. This choice of the step size ensures a sufficient decrease of the objective function value in every iteration of Algorithm 1 and the overall convergence of the presented method. More precisely, for fixed  $\gamma \in (0, 1)$ ,  $\alpha \in (0, 1/2]$ , the step size is set to  $s^k = \gamma^{n_k}$ , where  $n_k$  is the smallest non-negative integer with

$$\alpha s^k \Phi(\omega^k) \leq F_{M_0}(\omega^k) - F_{M_0}(\omega^k + s^k(v^k - \omega^k)). \quad (4.5)$$

Note that given an arbitrary non-optimal  $\omega^k \in \text{dom}_{M^+(\Omega_o)} \psi$  with  $\|\omega^k\|_{M(\Omega_o)} \leq M_0$  this choice of the step size  $s^k$  is always possible since the function  $W: [0, 1] \rightarrow \mathbb{R} \cup \{-\infty\}$

$$W(s) = \frac{F_{M_0}(\omega^k) - F_{M_0}(\omega^k + s(v^k - \omega^k))}{s\Phi(\omega^k)}, \quad (4.6)$$

fulfills  $\lim_{s \rightarrow 0} W(s) \geq 1$ , similarly to [14, Remark 2]. Note that the left-hand side of (4.5) is positive if  $\omega^k$  is not optimal. Thus, the quasi-Armijo-Goldstein stepsize rule ensures a decrease of the objective function value in each iteration. In particular, we get

$$\beta \|\omega^{k+1}\|_{M(\Omega_o)} \leq F_{M_0}(\omega^{k+1}) \leq F_{M_0}(\omega^{k+1/2}) \leq F_{M_0}(\omega^k) \leq F_{M_0}(\omega^1),$$

and consequently  $F_{M_0}(\omega^k) = F(\omega^k)$  for all iterates  $\omega^k$ . To obtain quantifiable estimates for the descent in the objective function value we impose additional regularity assumptions on  $\Psi'$  until the end of this section.

**Assumption 3** Assume that  $\Psi'$  is Lipschitz-continuous on compact sets: Given a compact set  $\mathcal{N} \subset \text{dom } \Psi$  there exists  $L_{\mathcal{N}} > 0$  with



$$\sup_{N_1, N_2 \in \mathcal{N}} \frac{\|\Psi'(N_1) - \Psi'(N_2)\|}{\|N_1 - N_2\|} \leq L_{\mathcal{N}}, \quad (4.7)$$

where  $\|A\| = \|A\|_{\text{Sym}(n)} = \sqrt{\text{Tr}(AA^T)}$  is the Frobenius norm.

Note that this additional assumption is fulfilled if the design criterion  $\Psi$  is two-times continuously differentiable on its domain. This is the case for, e.g., the already mentioned A and D-criterion, see Sect. 3.1. We immediately arrive at the following proposition.

**Proposition 4.4** *Let Assumption 3 hold and let  $\omega_1 \in \text{dom}_{M^+(\Omega_o)} \psi$  be given. Define the associated sub-level set  $E_{\omega^1}$  as*

$$E_{\omega^1} = \left\{ \omega \in M^+(\Omega_o) \mid F(\omega) \leq F(\omega^1) \right\}.$$

*Then there exists  $L_{\omega^1}$  such that*

$$\sup_{\omega_1, \omega_2 \in E_{\omega^1}} \frac{\|\psi'(\omega_1) - \psi'(\omega_2)\|_{C(\Omega_o)}}{\|\omega_1 - \omega_2\|_{M(\Omega_o)}} \leq L_{\omega^1}. \quad (4.8)$$

**Proof** First we observe that  $E_{\omega^1}$  is convex, bounded, and weak\* closed. Consequently the set of associated information matrices

$$\mathcal{I}(E_{\omega^1}) = \left\{ \mathcal{I}(\omega) + \mathcal{I}_0 \mid \omega \in E_{\omega^1} \right\}$$

is compact. For  $\omega_1, \omega_2 \in E_{\omega^1}$  we obtain

$$\begin{aligned} & \|\psi'(\omega_1) - \psi'(\omega_2)\|_{C(\Omega_o)} \\ &= \|\mathcal{I}^* \Psi'(\mathcal{I}(\omega_1) + \mathcal{I}_0) - \mathcal{I}^* \Psi'(\mathcal{I}(\omega_2) + \mathcal{I}_0)\| \\ &\leq \|\mathcal{I}^*\|_{\text{Sym}(n) \rightarrow C(\Omega_o)} \|\Psi'(\mathcal{I}(\omega_1) + \mathcal{I}_0) - \Psi'(\mathcal{I}(\omega_2) + \mathcal{I}_0)\| \\ &\leq L_{\mathcal{I}(E_{\omega^1})} \|\mathcal{I}^*\|_{\text{Sym}(n) \rightarrow C(\Omega_o)} \|\mathcal{I}(\omega_1) - \mathcal{I}(\omega_2)\| \\ &\leq L_{\mathcal{I}(E_{\omega^1})} \|\mathcal{I}^*\|_{\text{Sym}(n) \rightarrow C(\Omega_o)} \|\mathcal{I}\|_{M^+(\Omega_o) \rightarrow \text{Sym}(n)} \|\omega_1 - \omega_2\|_{M(\Omega_o)}, \end{aligned}$$

completing the proof.  $\square$

Using this additional local regularity we obtain the following estimate on the growth behavior of the function  $F$  at  $\omega^k$  in the search direction.

**Lemma 4.5** *Let Assumption 3 hold. Let  $\omega^k \in \text{dom}_{M^+(\Omega_o)} \psi$  with  $\|\omega^k\|_{M(\Omega_o)} \leq M_0$  and  $v^k$  as in Lemma 4.2 be given. Moreover, let  $\omega_s^{k+1/2} = (1-s)\omega^k + sv^k$  with  $s \in [0, 1]$  and  $\omega_s^{k+1/2} \in E_{\omega^k}$  be given. Then there holds*

$$F_{M_0}(\omega_s^{k+1/2}) - F_{M_0}(\omega^k) \leq -s\Phi(\omega^k) + \frac{L_{\omega^k}}{2} s^2 \|v^k - \omega^k\|_{M(\Omega_o)}^2,$$

where  $L_{\omega^k}$  denotes the Lipschitz constant of  $\psi'$  on  $E_{\omega^k}$ .

**Proof** By assumption there holds  $F_{M_0}(\omega_s^{k+1/2}) \leq F_{M_0}(\omega^k)$  and consequently  $\omega_s^{k+1/2} \in E_{\omega^k}$ . Therefore we obtain

$$\begin{aligned} F_{M_0}(\omega_s^{k+1/2}) - F_{M_0}(\omega^k) &= -s \langle \psi'(\omega^k), \omega^k - v^k \rangle \\ &\quad + \beta \varphi_{M_0}(\|\omega_s^{k+1/2}\|_{M(\Omega_o)}) - \beta \varphi_{M_0}(\|\omega^k\|_{M(\Omega_o)}) \\ &\quad + \int_0^s \langle \psi'(\omega_\sigma) - \psi'(\omega^k), v^k - \omega^k \rangle d\sigma, \end{aligned}$$

with  $\omega_\sigma = \omega^k + \sigma(v^k - \omega^k)$  for  $\sigma \in [0, 1]$ . Using the convexity of  $\varphi_{M_0}(\|\cdot\|_{M(\Omega_o)})$  we obtain

$$\begin{aligned} &-s \langle \psi'(\omega^k), \omega^k - v^k \rangle + \beta \varphi_{M_0}(\|\omega_s^{k+1/2}\|_{M(\Omega_o)}) - \beta \varphi_{M_0}(\|\omega^k\|_{M(\Omega_o)}) \\ &\leq -s \left( \langle \psi'(\omega^k), \omega^k - v^k \rangle + \beta \varphi_{M_0}(\|\omega^k\|_{M(\Omega_o)}) - \beta \varphi_{M_0}(\|v^k\|_{M(\Omega_o)}) \right), \end{aligned}$$

where the right-hand side simplifies to  $-s\Phi(\omega^k)$ . Due to the Lipschitz continuity of  $\psi'$  on  $E_{\omega^k}$  we get

$$\begin{aligned} &\int_0^s \langle \psi'(\omega_\sigma) - \psi'(\omega^k), v^k - \omega^k \rangle d\sigma \\ &\leq \|v^k - \omega^k\|_{M(\Omega_o)} \int_0^s \|\psi'(\omega_\sigma) - \psi'(\omega^k)\|_{C(\Omega_o)} d\sigma \\ &\leq L_{\omega^k} \|v^k - \omega^k\|_{M(\Omega_o)}^2 \int_0^s \sigma d\sigma = \frac{L_{\omega^k} s^2}{2} \|v^k - \omega^k\|_{M(\Omega_o)}^2. \end{aligned}$$

Combining both estimates yields the result.  $\square$

In order to prove the main result we additionally need the following technical lemma.

**Lemma 4.6** *Let  $\omega^k \in \text{dom}_{M^+(\Omega_o)} \psi$  with  $\Phi(\omega^k) > 0$  be given. The function*

$$W: (0, 1] \rightarrow \mathbb{R} \cup \{-\infty\}$$

*from (4.6) is continuous on  $(0, 1)$ . Furthermore, denoting by  $s^k$  the step size from (4.5), there exists  $\hat{s}^k \in [s^k, s^k/\gamma]$  with  $W(\hat{s}^k) = \alpha$  if  $s^k < 1$ .*

**Proof** First, note that for  $s \in [0, 1]$  we have  $\omega_s = (1-s)\omega^k + s v^k \in \text{dom}_{M^+(\Omega_o)} \psi$  due to  $\mathcal{I}(\omega_s) + \mathcal{I}_0 = (1-s)\mathcal{I}(\omega^k) + s\theta_k \partial S[\hat{q}](\hat{x}^k) \partial S[\hat{q}](\hat{x}^k)^\top + \mathcal{I}_0 \in \text{PD}(n)$ . Furthermore, using Assumption 2 it can be verified that

$$W(s) = (F_{M_0}(\omega_0) - F_{M_0}(\omega_s)) / (s \Phi(\omega_0))$$

is continuous on  $s \in (0, 1)$ . Additionally, with lower semi-continuity of  $\Psi$ , we verify that  $W(s) \rightarrow -\infty$  for  $s \rightarrow 1$  in case that  $\mathcal{I}(v^k) \notin \text{dom } \Psi$ . We conclude the proof

by applying the mean value theorem on  $[s^k, s^k/\gamma] \subset (0, 1]$ , taking into account that  $W(s^k) \geq \alpha > W(s^k/\gamma)$ .  $\square$

Combining the previous results we are able to prove sub-linear convergence of the presented algorithm.

**Theorem 4.7** *Let the sequence  $\omega^k$  be generated by Algorithm 1 using the quasi-Armijo-Goldstein condition (4.5). Then there exists at least one weak\* accumulation point  $\bar{\omega}_\beta$  of  $\omega^k$  and every such point is an optimal solution to  $(P_\beta)$ . Additionally there holds*

$$r_F(\omega^k) \leq \frac{r_F(\omega^1)}{1 + q(k-1)} \quad (4.9)$$

with

$$q = \alpha \min \left\{ \frac{c_1}{L_{\omega^1}(M_0 + c_2)^2}, 1 \right\}, \quad (4.10)$$

where  $L_{\omega^1}$  is the Lipschitz-constant of  $\psi'$  on  $E_{\omega^1}$ ,  $M_0 = F(\omega^1)/\beta$ ,  $c_1 = 2\gamma(1 - \alpha)r_F(\omega_1)$  and a constant  $c_2 > 0$  with  $\|v^k\|_{M(\Omega_o)} \leq c_2$  for all  $k$ .

**Proof** Assume without restriction that  $\Phi(\omega^k) > 0$ , i.e. the algorithm does not terminate after finitely many steps. By construction and the choice of  $s^k$  there holds  $\omega^k \in E_{\omega^1}$  and consequently  $\|\omega^k\|_{M(\Omega_o)} \leq M_0$ ,  $F_{M_0}(\omega^k) = F(\omega^k)$  for all  $k$ . The same can be proven for  $\omega^{k+1/2}$ . Note that  $\omega^k$  is bounded and  $\psi'$  is weak\*-to-strong continuous. Therefore, there exists  $c_2 > 0$  with  $\|v^k\|_{M(\Omega_o)} \leq c_2$  for all  $k$ .

By the definition of the step size  $s^k$  as well as (4.2) there holds

$$\alpha s^k r_F(\omega^k) \leq \alpha s^k \Phi(\omega^k) \leq r_F(\omega^k) - r_F(\omega^{k+1/2}),$$

which yields

$$r_F(\omega^{k+1/2}) \leq (1 - \alpha s^k) r_F(\omega^k). \quad (4.11)$$

Since  $\Phi(\omega^k) > 0$  we obtain  $s^k \neq 0$  for all  $k$ . Two cases have to be distinguished. If  $s^k$  is equal to one we immediately arrive at

$$r_F(\omega^{k+1/2}) \leq (1 - \alpha) r_F(\omega^k) \leq r_F(\omega^k) - \alpha \frac{r_F(\omega^k)^2}{r_F(\omega^1)}.$$

In the second case, if  $s^k < 1$ , there exists  $\hat{s}^k \in [s^k, s^k/\gamma]$  with

$$\alpha = \frac{F(\omega^k) - F(\omega^k + \hat{s}^k(v^k - \omega^k))}{\hat{s}^k \Phi(\omega^k)},$$

using Lemma 4.6. Consequently  $\omega^k + s(v^k - \omega^k) \in E_{\omega^1}$  for all  $0 \leq s \leq \hat{s}^k$  due to the convexity of  $F$ . Because of the Lipschitz-continuity of  $\psi'$  on  $E_{\omega^1}$ , Lemma 4.5 can be applied and, defining  $\delta\omega^k = v^k - \omega^k$ , there holds

$$\begin{aligned}\alpha &= \frac{F(\omega^k) - F(\omega^k + \hat{s}^k \delta\omega^k)}{\hat{s}^k \Phi(\omega^k)} \geq 1 - \frac{L_{\omega^1} \hat{s}^k \|\delta\omega^k\|_{M(\Omega_o)}^2}{2 \Phi(\omega^k)} \\ &\geq 1 - \frac{L_{\omega^1} s^k \|\delta\omega^k\|_{M(\Omega_o)}^2}{2\gamma \Phi(\omega^k)}.\end{aligned}$$

The last estimate is true because of  $\hat{s}^k \leq s^k/\gamma$ . Reordering and using (4.2) yields

$$1 \geq s^k \geq 2\gamma(1 - \alpha) \frac{\Phi(\omega^k)}{L_{\omega^1} \|v^k - \omega^k\|_{M(\Omega_o)}^2} \geq 2\gamma(1 - \alpha) \frac{r(\omega^k)}{L_{\omega^1} \|v^k - \omega^k\|_{M(\Omega_o)}^2}.$$

Combining the estimates in both cases and using  $r_F(\omega^{k+1}) \geq r_F(\omega^{k+1/2})$ , the inequality

$$0 \leq \frac{r_F(\omega^{k+1})}{r_F(\omega^1)} \leq \frac{r_F(\omega^{k+1/2})}{r_F(\omega^1)} \leq \frac{r_F(\omega^k)}{r_F(\omega^1)} - q_k \left( \frac{r_F(\omega^k)}{r_F(\omega^1)} \right)^2 \quad \forall k \in \mathbb{N} \quad (4.12)$$

holds, where the constant  $q_k$  is given by

$$q_k = r_F(\omega^1) \alpha \min \left\{ \frac{2\gamma(1 - \alpha)}{L_{\omega^1} \|v^k - \omega^k\|_{M(\Omega_o)}^2}, \frac{1}{r_F(\omega^k)} \right\},$$

if  $s^k < 1$  and  $q_k = \alpha$  otherwise. We estimate

$$q_k \geq \alpha \min \left\{ \frac{2\gamma(1 - \alpha)r_F(\omega^1)}{L_{\omega^1} (M_0 + c_2)^2}, 1 \right\} =: q.$$

The claimed convergence rate (4.9) now follows directly from the recursion formula (4.12); see [20, Lemma 3.1]. Consequently, each subsequence of  $\omega^k$  is a minimizing sequence. Since  $\omega^k$  is bounded, it admits at least one weak\* accumulation point. Due to the derived convergence rate and the weak\* lower semi-continuity of  $F$  each weak\* accumulation point  $\tilde{\omega}_\beta$  is a minimizer of  $(P_\beta)$ .  $\square$

## 4.2 Acceleration and sparsification strategies

As we have seen in the previous section, an iterative application of steps 2.–4. in Algorithm 1 is sufficient to obtain weak\* convergence of the iterates  $\omega^k$ , as well as a sublinear convergence rate for the objective function. However, it is obvious that the support size of the iterates  $\omega^k$  grows monotonically in every iteration unless the current gradient is bounded from below by  $-\beta$  or, more unlikely, the step size  $s^k$  is chosen as

1. Therefore, while the implementation of steps 2.–4. is fairly easy, an algorithm only consisting of point insertion steps will likely yield iterates with undesirable sparsity properties, e.g., a clusterization of the intermediate support points around the support points of a minimizer to  $(P_\beta)$ . In the following we mitigate those effects by augmenting the point insertion steps by point removal steps, where we incorporate ideas from [13, 15]. For  $\{x_j\}_{j=1}^{m_k} = \text{supp } \omega^{k+1/2}$ , we define the parameterization:

$$\omega(\lambda) := \sum_{j=1}^{m_k} \lambda_j \delta_{x_j} \quad \forall \lambda \in \mathbb{R}^{m_k}. \quad (4.13)$$

Now, we set  $\omega^{k+1} = \omega(\lambda^{k+1})$ , where the improved vector  $\lambda^{k+1} \in \mathbb{R}^{m_k}$  is chosen as an approximate solution to the (finite dimensional) coefficient optimization problem

$$\min_{\lambda \in \mathbb{R}^{m_k}, \lambda \geq 0} F(\omega(\lambda)) = \psi(\omega(\lambda)) + \beta \|\lambda\|_1, \quad (4.14)$$

that fulfills  $F(\omega^{k+1}) \leq F(\omega^{k+1/2})$ . In this manuscript, we focus on two special instances of this removal step, which are detailed below.

In the first strategy, the new coefficient vector  $\lambda^{k+1} = \lambda^{k+1}(\sigma_k)$  is obtained by

$$\lambda^{k+1}(\sigma_k)_j = \max \left\{ \lambda_i^{k+1/2} - \sigma_k \left[ \psi'(\omega^{k+1/2})(x_j) + \beta \right], 0 \right\} \quad \forall j \in \{1, \dots, m_k\}, \quad (4.15)$$

where  $\sigma_k > 0$  is a suitably chosen step size that avoids ascent in the objective function value. This corresponds to performing one step of a projected gradient method on (4.14) using the previous coefficient vector  $\lambda^{k+1/2}$  as a starting point. Thus, step 5. in Algorithm 1 subtracts or adds mass at support point  $x_j$  for  $-\psi'(\omega^{k+1/2})(x_j) < \beta$  or  $-\psi'(\omega^{k+1/2})(x_j) > \beta$ , respectively. Furthermore, the new coefficient  $\lambda_j^{k+1}$  of the Dirac delta function  $\delta_{x_j}$  is set to zero if

$$\lambda_j^{k+1/2} - \sigma_k \left[ \psi'(\omega^{k+1/2})(x_j) + \beta \right] \leq 0,$$

removing the point measure from the iterate.

Secondly, we suppose that the finite-dimensional sub-problems (4.14) can be solved exactly and choose

$$\lambda^{k+1} \in \arg \min_{\lambda \in \mathbb{R}^{m_k}, \lambda \geq 0} F(\omega(\lambda)). \quad (4.16)$$

In this case, the conditions

$$\text{supp } \omega^{k+1} \subset \text{supp } \omega^{k+1/2}, \quad F(\omega^{k+1}) \leq F(\omega^{k+1/2})$$

are trivially fulfilled. If all finite dimensional sub-problems are solved exactly, the method can be interpreted as a method operating on a set of active points  $\mathcal{A}_k =$

$\text{supp } \omega^k$ ; cf. [34]: In each iteration, the minimizer  $\hat{x}^k$  of the current gradient  $\psi'_k$  is added to the support set to obtain  $\mathcal{A}_{k+1/2} = \mathcal{A}_k \cup \{\hat{x}^k\}$ . Then, the problem (4.16) is solved on the new support set (i.e. with  $\text{supp } \omega^{k+1/2}$  replaced by  $\mathcal{A}_{k+1/2}$  in the definition of (4.13)) to obtain the next iterate  $\omega^{k+1}$ . Note that the next active set is given by  $\mathcal{A}_{k+1} = \text{supp } \omega^{k+1}$ , which automatically removes support points corresponding to zero coefficients in each iteration.

Finally, the proof of Lemma 3.10 leads to an implementable sparsifying procedure which, given an arbitrary finitely supported positive measure, finds a sparse measure choosing a subset of at most  $n(n+1)/2$  support points and yielding the same information matrix at a smaller cost. The procedure is summarized in Algorithm 2. Applying this method to the intermediate iterate  $\omega^{k+1/2}$  in step 5. of Algorithm 1 guarantees the a priori bound  $\#\text{supp } \omega^k \leq n(n+1)/2$  as well as the convergence of the presented procedure towards a sparse minimizer of  $(P_\beta)$ .

**Proposition 4.8** *Let  $\omega = \sum_{j=1}^m \lambda_j \delta_{x_j}$  be given and assume that  $\{\mathcal{I}(\delta_{x_j})\}_{j=1}^m$  is linearly dependent. Denote by  $\omega_{\text{new}} = \sum_{\{j \mid \lambda_{\text{new},j} > 0\}} \lambda_{\text{new},j} \delta_{x_j}$  the measure that is obtained after one execution of the loop in Algorithm 2. Then there holds*

$$F(\omega_{\text{new}}) \leq F(\omega), \quad \#\text{supp } \omega_{\text{new}} \leq \#\text{supp } \omega - 1.$$

**Proof.** We point to the proof of Lemma 3.10 which gives

$$\mathcal{I}(\omega_{\text{new}}) = \mathcal{I}(\omega), \quad \|\omega_{\text{new}}\|_{M(\Omega_o)} \leq \|\omega\|_{M(\Omega_o)}, \quad \#\text{supp } \omega_{\text{new}} \leq \#\text{supp } \omega - 1.$$

□

---

#### Algorithm 2 Support-point removal

---

1. Let  $\omega = \sum_{j=1}^m \lambda_j \delta_{x_j}$  be given.
  - while**  $\{\mathcal{I}(\delta_{x_j})\}_{j=1}^m$  linearly dependent **do**
    2. Find  $0 \neq \gamma$  with  $0 = \sum_{j=1}^m \gamma_j \mathcal{I}(\delta_{x_j})$  and  $\sum_{j=1}^m \gamma_j \geq 0$  (see Sect. 4.3).
    3. Set  $\mu = \max_j \{\gamma_j / \lambda_j\}$ ,  $\lambda_{\text{new},j} = \lambda_j - \gamma_j / \mu$ .
    4. Update  $\omega_{\text{new}} = \sum_{\{j \mid \lambda_{\text{new},j} > 0\}} \lambda_{\text{new},j} \delta_{x_j}$ .
  - end while**
- 

**Proposition 4.9** *Assume that  $\#\text{supp } \omega^1 \leq n(n+1)/2$  and let  $\omega^{k+1}$  be obtained by applying Algorithm 2 to  $\omega^{k+1/2}$  in each iteration of Algorithm 1. Then the results of Theorem 4.7 hold. Furthermore we obtain  $\#\text{supp } \omega^k \leq n(n+1)/2$  for all  $k \in \mathbb{N}$  and consequently  $\#\text{supp } \bar{\omega}_\beta \leq n(n+1)/2$  for every weak\* accumulation point  $\bar{\omega}_\beta$  of  $\omega^k$ .*

**Proof** The statement for the support of  $\omega^k$  readily follows from an inductive application of Proposition 4.8 by noting that  $\mathcal{I}(\delta_{x_j}) \in \text{Sym}(n)$  and  $\dim \text{Sym}(n) = n(n+1)/2$ . The sparsity statement for every accumulation point  $\bar{\omega}$  follows then directly from Lemma 3.9. □

We emphasize that the sparsifying procedure from Algorithm 2 can be readily combined with the previously presented point removal steps in a straightforward fashion. In practical computations we optimize the coefficients of the Dirac delta functions in the current support either by (4.15) or (4.16) obtaining an intermediate iterate  $\omega^{k+3/4}$ . Subsequently we apply Algorithm 2. Since in both cases, the number of support points cannot increase, the statements of the last proposition remain true.

### 4.3 Computation of the sparsification steps

It remains to comment on the computational aspects of the point removal steps presented in this section. First, we address the approximate solution of the finite dimensional subproblems. If  $\lambda^k$  is determined from (4.16), we have to solve a finite-dimensional convex optimization problem in every iteration. Since the most common choices for the optimal design criterion  $\Psi$  are twice continuously differentiable, we choose to implement a semi-smooth Newton method; see, e.g., [32]. To benefit from the fast local convergence behavior for this class of methods we warm-start the algorithm using the coefficient vector  $\lambda^{k+1/2}$  of the intermediate iterate  $\omega^{k+1/2}$ . This choice of the starting point often gives a good initial guess for  $\lambda^{k+1}$ . However, we note that essentially any algorithm for smooth convex problems with non-negativity constraints on the optimization variables can be employed instead.

Finally, we consider the application of Algorithm 2, given a sparse input measure  $\omega$  with  $\text{supp } \omega = \{x_i\}_{i=1}^m$ . Step 1. amounts to the computation of the symmetric rank one matrices  $\{\mathbf{I}(\delta_{x_i})\}_{i=1}^m \subset \text{NND}(n)$ , which we identify with vectors  $\{\mathbf{I}(\delta_{x_i})\}_{i=1}^m \subset \mathbb{R}^{n(n+1)/2}$ . Additionally, in each execution of the loop step 2. has to be executed, which requires to compute a vector  $\gamma$  in the kernel of the matrix  $\mathbf{I}(\omega) \in \mathbb{R}^{n(n+1)/2 \times m}$ , defined by

$$[\mathbf{I}(\omega)]_{i,j} = \mathbf{I}(\delta_{x_j})_i, \quad i = 1, \dots, n(n+1)/2, \quad j = 1, \dots, m.$$

This can be done efficiently employing either a SVD-decomposition or a rank-revealing QR-decomposition. Since  $\gamma$  is only determined up to a scalar multiple, it can be chosen with  $\sum_{j=1}^m \gamma_j \geq 0$ . Furthermore, assuming that Algorithm 2 is applied to  $\omega^{k+1/2}$  for every  $k$ , this loop will run at most once in each iteration. This follows since each iteration starts with a support set such that  $\mathbf{I}(\omega^k)$  is of full rank, and the point insertion step either maintains full rank, or adds a linearly dependent vector to  $\mathbf{I}(\omega^{k+1/2})$ . In the latter case the removal of at least one support point yields again full rank in the next iteration.

## 5 Numerical example

We end this paper with the study of a numerical example. In the following, we consider the unit square  $\bar{\Omega} = \Omega_o = [0, 1]^2$  and a family  $\{\mathcal{T}_h\}_{h>0}$  of uniform triangulations of  $\Omega_o$ , where  $h$  denotes the maximal diameter of a cell  $K \in \mathcal{T}_h$ . The set of associated grid nodes is called  $\mathcal{N}_h$ . Concretely, we consider a sequence of successively refined grids

with  $h_k = \sqrt{2}/2^k, k \in \{1, 2, \dots, 9\}$ . The state and sensitivity equations, respectively, are discretized by linear finite elements on  $\mathcal{T}_h$  and the solutions to the discretized sensitivity equations are denoted by  $\{\partial_k S^h[\hat{q}]\}_{k=1}^n$ . Moreover,  $M^+(\Omega_o)$  is replaced by positive linear combinations of nodal Dirac delta function

$$M_h^+ := \{\omega_h \in M^+(\Omega_o) \mid \text{supp } \omega_h \subset \mathcal{N}_h\} = M^+(\mathcal{N}_h).$$

The discrete design problem is now stated as

$$\begin{aligned} \min_{\omega \in M_h^+} \quad & \Psi(\mathcal{I}_h(\omega) + \mathcal{I}_0) + \beta \|\omega\|_{M(\Omega_o)}, \\ \text{where } \mathcal{I}_h(\omega) = \quad & \int_{\Omega_o} \partial S^h[\hat{q}](x) \partial S^h[\hat{q}](x)^\top d\omega(x). \end{aligned} \quad (5.1)$$

A solution  $\bar{\omega}_{\beta,h} \in M_h^+$  to (5.1) is computed by the different variants of Algorithm 1 where the search for the new position  $\hat{x}^k$  in step 2. is restricted to  $\mathcal{N}_h$ . For abbreviation we again define the reduced design criterion  $\psi_h(\omega) = \Psi(\mathcal{I}_h(\omega))$ .

Our aim in this section is twofold. First, we want to numerically illustrate the theoretical results. Secondly, we want to study the practical performance of the proposed algorithms according to various criteria including the computational time, the evolution of the sparsity pattern throughout the iterations and the influence of the fineness of the triangulation. Concretely, we consider the A-optimal design problem, i.e.  $\Psi(N) = \text{Tr}(N^{-1})$  and the discrete state and the associated sensitivities  $\partial S^h[\hat{q}]$  are computed for a fixed  $\hat{q}$  once at the beginning. During the execution of the different variants of Algorithms 1 no additional PDEs need to be solved. Moreover, the gradient of the reduced cost functional is given by

$$\begin{aligned} [\psi'_h(\omega)](x) &= -\text{Tr}(\mathcal{I}_h(\omega)^{-1} \mathcal{I}_h(\delta_x) \mathcal{I}_h(\omega)^{-1}) \\ &= -\|\mathcal{I}_h(\omega)^{-1} \partial S^h[\hat{q}](x)\|_{\mathbb{R}^n}^2 \quad \forall x \in \Omega, \end{aligned}$$

which relates the pointwise value of the gradient directly to the corresponding sensitivity vector  $\partial S^h[\hat{q}](x) \in \mathbb{R}^n$ . A corresponding computation on the discrete level allows for an efficient implementation based on a single Cholesky-decomposition of  $\mathcal{I}^h(\omega)$  in each iteration. Moreover an expression for the Hessian-vector-product  $[\psi''_h(\omega)(\delta\omega)](x)$  for  $\delta\omega \in M(\Omega)$  can be derived by differentiating the above expression.

**Remark 5.1** It is possible to show that every solution  $\bar{\omega}_{\beta,h} \in M^+(\Omega_o) \cap M_h$  to (5.1) is also a minimizer of the semi-discrete problem

$$\min_{\omega \in M^+(\Omega_o)} \psi_h(\omega) + \beta \|\omega\|_{M(\Omega_o)} \quad (5.2)$$

where the space of possible designs is not discretized. This corresponds to the variational discretization paradigm; cf [17]. In particular, proceeding as for the fully



continuous problem, a measure  $\bar{\omega}_{\beta,h} \in M^+(\Omega_o)$  is an optimal solution to (5.1) if and only if

$$-\psi'_h(\bar{\omega}_{\beta,h}) \leq \beta, \quad \text{supp } \bar{\omega}_{\beta,h} \subset \{x \in \Omega_o \mid \psi'_h(\bar{\omega}_{\beta,h})(x) = -\beta\}.$$

Since the main focus of the present paper lies on the description of the sparse sensor placement problem and its efficient solution, we postpone a detailed discussion of the discretization to a follow-up paper.

## 5.1 Estimation of convection and diffusion parameters

As an example for the state equation (2.1), we take a convection-diffusion process where for a given  $q \in \mathcal{Q}_{ad} = \{q \in \mathbb{R}^3 \mid 0.25 \leq q_1 \leq 5\}$  the associated state  $y = S[q] \in H_0^1(\Omega) \cap C(\Omega_o)$  is the unique solution to

$$a(q, y)(\varphi) = \int_{\Omega} \left[ q_1 \nabla y \cdot \nabla \varphi + q_2 \varphi \frac{\partial y}{\partial x_1} + q_3 \varphi \frac{\partial y}{\partial x_2} \right] dx = \int_{\Omega} f \varphi dx, \quad (5.3)$$

for all  $\varphi \in H_0^1(\Omega)$ . The forcing term  $f$  is chosen as  $\exp(3(x_1^2 + x_2^3))$ . This corresponds to the linear elliptic equation

$$-q_1 \Delta y + \begin{pmatrix} q_2 \\ q_3 \end{pmatrix} \cdot \nabla y = f \quad \text{in } \Omega,$$

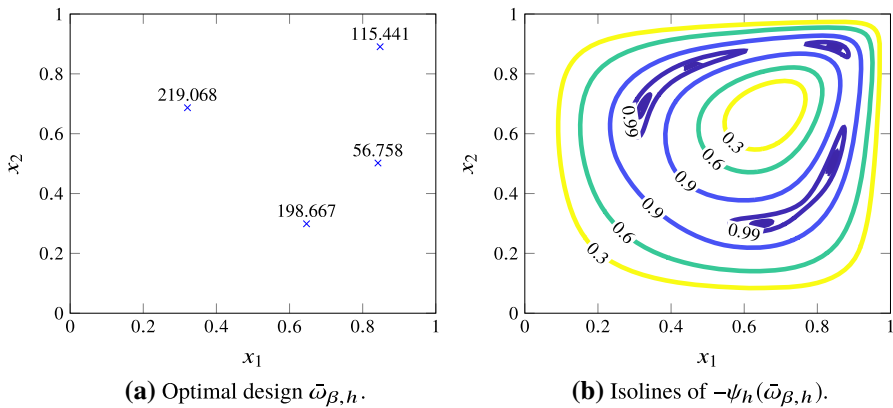
together with homogeneous Dirichlet boundary conditions on  $\partial\Omega$ . Here, the parameter  $q$  contains the scalar diffusion and convection coefficients of the elliptic operator. As a priori guess for the parameter we choose  $\hat{q} = (3, 0.5, 0.25)^\top$ . Note that while (5.3) is a linear equation, the state  $y \in H_0^1(\Omega) \cap C(\Omega_o)$  depends non-linearly but differentiably on  $q$ . For each  $k \in \{1, 2, 3\}$  the sensitivity  $\delta y_k = \partial_k S[\hat{q}] \in H_0^1(\Omega) \cap C(\Omega_o)$  can be computed from (2.2). Due to the tri-linearity of the form  $a(\cdot, \cdot)(\cdot)$  it fulfills

$$a(\hat{q}, \delta y_k)(\varphi) = -a(\mathbf{e}_k, \hat{y})(\varphi) \quad \forall \varphi \in H_0^1(\Omega),$$

where  $\hat{y} = S[\hat{q}]$  and  $\mathbf{e}_k \in \mathbb{R}^3$  denotes the  $k$ th canonical unit vector.

### 5.1.1 First order optimality condition

In this section we numerically illustrate the first-order necessary and sufficient optimality conditions from Proposition 3.12. Therefore we compute an A-optimal design for Example 1 on grid level nine  $\mathcal{T}_{h_9}$  for  $\beta = 1$  and  $\mathcal{I}_0 = 0$ . For the computation we use Algorithm 1 (together with Algorithm 2 and a full resolution of the arising finite-dimensional subproblems), until the residual is zero (up to machine precision). We obtain a discrete optimal design  $\bar{\omega}_{\beta,h}$  in  $M^+(\Omega_o) \cap M_h$  with five support points. By closer inspection we observe that two of the computed support points are located in adjacent nodes of the triangulation. For a better visualization of the computed result,



**Fig. 1** Optimal design and isolines of the gradient

the corresponding Dirac delta functions are replaced by a single one placed at the center of mass. The coefficient of this new Dirac delta function is given by the combined mass of the original ones; see Fig. 1a. Alongside we plot the isolines of the nodal interpolant of  $-\psi'_h(\bar{\omega}_{\beta,h})$ . Note that the values of  $-\psi'_h(\bar{\omega}_{\beta,h})$  in  $N_h$  are bounded from above by the cost parameter  $\beta = 1$  and the support points of  $\bar{\omega}_{\beta,h}$  align themselves with those points in which this upper bound is achieved; see Fig. 1b.

### 5.1.2 Confidence domains of the optimal estimator

Given the optimal design  $\bar{\omega}_h$  from Fig. 1a, and  $K > 0$  we note that the measure  $\bar{\omega}_h^K = (K/\|\bar{\omega}_{\beta,h}\|_{M(\Omega_o)})\bar{\omega}_{\beta,h}$  is an optimal solution to

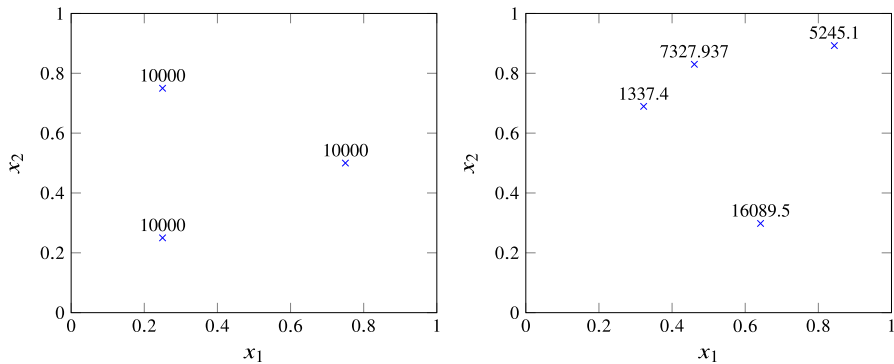
$$\min_{\omega \in M^+(\Omega_o)} \text{Tr}(\mathcal{I}_h(\omega_h)^{-1}) \quad \text{subject to } \|\omega_h\|_{M(\Omega_o)} \leq K,$$

since the A-optimal design criterion is positive homogeneous; see Proposition 3.12. In this section we compute the linearized confidence domains (2.8) of the least-squares estimator  $\tilde{q}$  from (2.4) corresponding to  $\bar{\omega}_h^K$  for  $K = 3 \cdot 10^4$ .

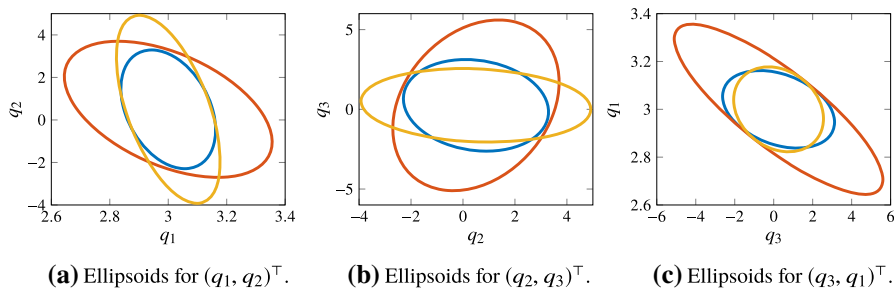
Note that, given a sparse design measure  $\omega$ , and the associated linearized estimator  $\tilde{q}_{\text{lin}} = (\tilde{q}_{\text{lin}}^1, \tilde{q}_{\text{lin}}^2, \tilde{q}_{\text{lin}}^3)^\top$ , see (2.7), there holds  $\text{Cov}[\tilde{q}_{\text{lin}}, \tilde{q}_{\text{lin}}] = \mathcal{I}_h(\omega)^{-1}$ ; see the discussion in Sect. 2. Consequently we have

$$\mathcal{I}_h(\omega)_{kk}^{-1} = \text{Var}[\tilde{q}_{\text{lin}}^k], \quad k \in \{1, 2, 3\} \quad \text{and} \quad \text{Tr}(\mathcal{I}_h(\omega)^{-1}) = \sum_{k=1}^3 \text{Var}[\tilde{q}_{\text{lin}}^k].$$

As a comparison, we also consider the estimators corresponding to two reference designs of the same norm. The first measure  $\omega^1$  is chosen as a linear combination of three Dirac delta functions with equal coefficients while the second measure  $\bar{\omega}_h^{K,W}$  is



**Fig. 2** Reference measures  $\omega_1$  (left) and  $\bar{\omega}_h^{K,W}$  (right)



**Fig. 3** Confidence ellipsoids for the estimators associated to  $\bar{\omega}_h^K$  (blue),  $\omega_1$  (red) and  $\bar{\omega}_h^{K,W}$  (yellow)

**Table 1** Trace and diagonal entries of  $\mathcal{I}_h(\omega)^{-1}$

$\omega$	$\mathcal{I}_h(\omega)_{11}^{-1}$	$\mathcal{I}_h(\omega)_{22}^{-1}$	$\mathcal{I}_h(\omega)_{33}^{-1}$	$\text{Tr}(\mathcal{I}_h(\omega)^{-1})$
$\bar{\omega}_h^K$	0.019	5.627	5.955	11.601
$\omega_1$	0.091	7.388	20.678	28.157
$\bar{\omega}_h^{K,W}$	0.023	14.12	3.831	17.974

a solution to

$$\min_{\omega \in M^+(\Omega_o)} \text{Tr}(W \mathcal{I}_h(\omega_h)^{-1} W) \quad \text{subject to } \|\omega_h\|_{M(\Omega_o)} \leq K, \quad (5.4)$$

where  $W = \text{diag}(1, 1, 4)$ , i.e. we place more weight on the variance for the estimation of  $q_3$ . The designs  $\omega_1$  and  $\bar{\omega}_h^{K,W}$  are depicted in Fig. 2.

For a better visualization we plot the 50%-linearized confidence domains of the obtained estimators for the two dimensional parameter vectors  $(q_1, q_2)^\top$ ,  $(q_2, q_3)^\top$ , and  $(q_3, q_1)^\top$  in Fig. 3. Additionally, for each design we report  $\text{Tr}(\mathcal{I}_h(\omega)^{-1})$  as well as the diagonal entries of  $\mathcal{I}_h(\omega)^{-1}$  in Table 1.

As expected, since  $\bar{\omega}_{\beta,h}$  is chosen by the A-optimal design criterion, we observe that

$$\text{Tr}(\mathcal{I}_h(\bar{\omega}_h^K)^{-1}) \leq \text{Tr}(\mathcal{I}_h(\bar{\omega}_h^{K,W})^{-1}) \leq \text{Tr}(\mathcal{I}_h(\omega_1)^{-1}). \quad (5.5)$$

Moreover we note that  $\mathcal{I}_h(\bar{\omega}_h^K)^{-1}_{kk} < \mathcal{I}_h(\omega_1)^{-1}_{kk}$  for all  $k$ , i.e. the optimal estimator estimates all unknown parameters with a smaller variance than the estimator associated to the reference design  $\omega_1$ . As a consequence, the linearized confidence domains of the optimal estimator are contained in those of the one corresponding to  $\omega_1$ ; see Fig. 3. In contrast, considering  $\omega_2$ , we have  $\mathcal{I}_h(\bar{\omega}_h^{K,W})^{-1}_{33} < \mathcal{I}_h(\bar{\omega}_h^K)^{-1}_{33}$  and  $\mathcal{I}_h(\bar{\omega}_h^K)^{-1}_{kk} < \mathcal{I}_h(\mathcal{I}_h(\bar{\omega}_h^{K,W})^{-1})_{kk}$  for  $k = 1, 2$ , i.e. the third parameter is estimated more accurately by choosing the measurement locations and weights according to  $\omega_2$  while the variance for the estimation of the other parameters is larger. This is a consequence of the different weighting of the matrix entries in (5.5). On the one hand, the obtained results show the efficiency of an optimally chosen measurement design at least for the linearized model. On the other hand, they also highlight that the properties of the obtained optimal estimators crucially depend on the choice of the optimal design criterion  $\Psi$ .

### 5.1.3 Comparison of point insertion algorithms

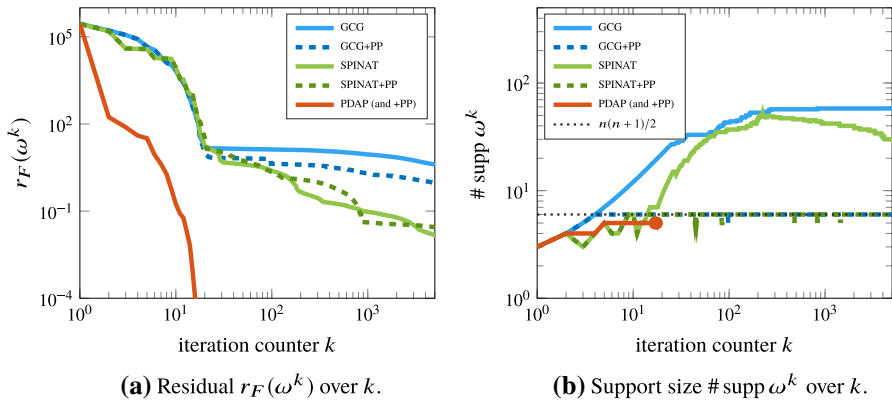
In this section we investigate the performance of the successive point insertion algorithm presented in Sect. 4.1. We consider the same setup as in Sect. 5.1.1, i.e. we solve the A-optimal design problem for Example 1 on grid level nine with  $\beta = 1$  and  $\mathcal{I}_0 = 0$ . The step size parameters  $\alpha$  and  $\gamma$  in (4.5) are both chosen as  $1/2$  throughout the experiments and the iteration is terminated if either  $\Phi(\omega^k) \leq 10^{-9}$  or if the iteration number  $k$  exceeds  $2 \cdot 10^4$ . The aim of this section is to confirm the theoretical convergence results for Algorithm 1 and to demonstrate the necessity of additional point removal steps. Additionally we want to highlight the differences between the three presented choices of the new coefficient vector  $\lambda^{k+1}$  concerning the sparsity of the iterates and the practically achieved acceleration of the convergence. Specifically, we consider the following implementations of step 4. in Algorithm 1:

**GCG** In the straightforward implementation of the GCG algorithm we set  $\lambda^{k+1} = \lambda^{k+1/2}$ , i.e. only steps 1. to 4. are performed.

**SPINAT** Here, we employ the procedure suggested in [15], termed “Sequential Point Insertion and Thresholding”. In step 5.,  $\lambda^{k+1}$  is determined from a proximal gradient iteration (4.15). The step size is chosen as  $\sigma_k = (1/2)^n \sigma_{0,k}$ , where  $\sigma_{0,k} > 0$  for the smallest  $n \in \mathbb{N}$  giving  $F(\omega(\lambda^{k+1}(\sigma_k))) \leq F(\omega(\lambda^{k+1/2}))$ . In particular, given  $\omega^{k+1/2} = \sum_i \lambda_i^{k+1/2} \delta_{x_i}$ , we choose  $\sigma_{0,k}$  as

$$\sigma_{0,k} = \max \left\{ 100, -2 \min_i \left\{ \frac{\lambda_i}{-\psi'(\omega^{k+1/2})(x_i) - \beta} \right\} \right\}.$$

Note that by this choice of  $\sigma_{0,k}$ , the coefficients of all points  $x \in \text{supp } \omega^{k+1/2}$  with  $-\psi'(\omega^{k+1/2})(x) < \beta$  are set to zero in the first trial step (i.e. for  $n = 0$ ).



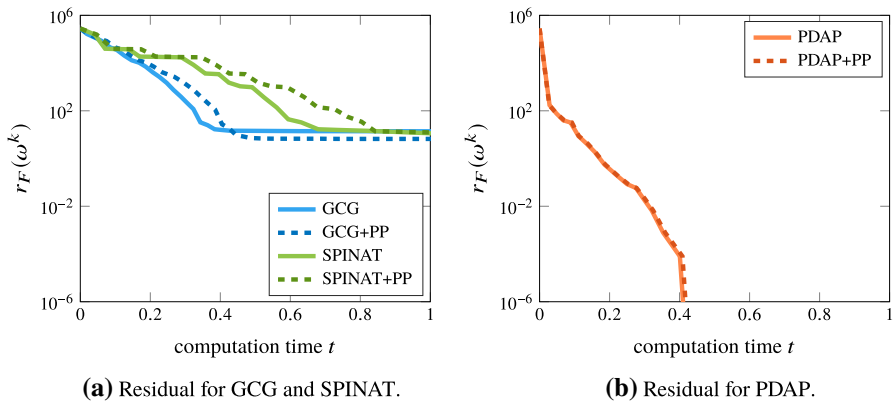
**Fig. 4** Residual and support size plotted over iteration number  $k$ . The results for PDAP and PDAP+PP are identical; a line with a dot denotes termination of the algorithm within machine tolerance

**PDAP** Here, the coefficient vector  $\lambda^{k+1}$  is chosen as in (4.16) by solving the finite dimensional sub-problem (4.14) up to machine precision in each iteration. For the solution we use a semi-smooth Newton method with a globalization strategy based on a backtracking line-search. The convergence criterion for the solution of the sub-problems is based on the norm of the Newton-residual. Since, this method can be interpreted as a method operating on a set of active points  $\mathcal{A}_k = \text{supp } \omega^k$  (see Sect. 4.2), we reference it by the name: “Primal-Dual Active Point”.

All three versions of the algorithm are also considered with an application of the sparsification step Algorithm 2 applied at the end of each iteration of Algorithm 1. In the following this will be denoted by an additional “+PP”.

In Fig. 4a we plot the residual  $r_F(\omega^k)$  for all considered algorithms over the iteration counter  $k$ . For GCG as well as SPINAT we observe a rapid decay of the computed residuals in the first few iterations. However, asymptotically both admit a sub-linear convergence rate, suggesting that the convergence result derived in Theorem 4.7 is sharp in this instance. The additional application of Algorithm 2 has no significant impact on the convergence behavior. We additionally note that both GCG and SPINAT terminate only since the maximum number of iterations is exceeded while the computed residuals  $r_F(\omega^k)$  and thus also the primal-dual gap  $\Phi(\omega^k)$  remain above  $10^{-3}$ . In contrast, PDAP terminates after few iterations within the tolerance. The results clearly indicate a better convergence rate than the one derived in Theorem 4.7.

Next, we study the influence of the different point removal steps on the sparsity pattern of the obtained iterates in Fig. 4b. For GCG we notice that the number of support points increases monotonically up to approximately 60. This suggests a strong clusterization of the intermediate support points around those of  $\bar{\omega}_{\beta,h}$  which is possibly caused by the small curvature of  $-\psi'_h(\bar{\omega}_{\beta,h})$  (see Fig. 1b) in the vicinity of its global maxima. A similar behavior can be observed for the iterates obtained through SPINAT. However, compared to GCG the support size for SPINAT grows slower due to the additional projected gradient step in every iteration. Concerning the application of



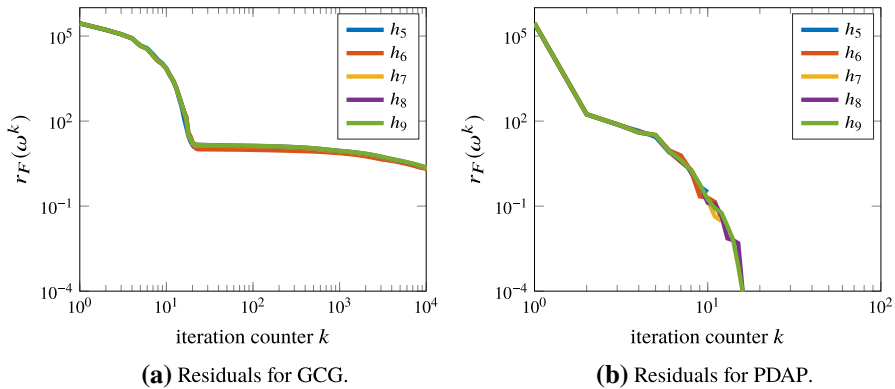
**Fig. 5** Residual  $r_F(\omega^k)$  plotted over the first second of the running time

Algorithm 2, we observe that the support remains bounded for all implementations with “+PP” by  $n(n+1)/2 = 6$  as predicted by Proposition 4.9. We note that this upper bound is achieved in almost all but the first few iterations for GCG and SPINAT. In contrast, PDAP yields iterates comprising less than six support points independently of the additional post-processing. A closer inspection reveals that the loop in Algorithm 2 is not carried out in any iteration, i.e. the sparsity of the iterates is fully provided by the exact solution of the finite-dimensional sub-problems.

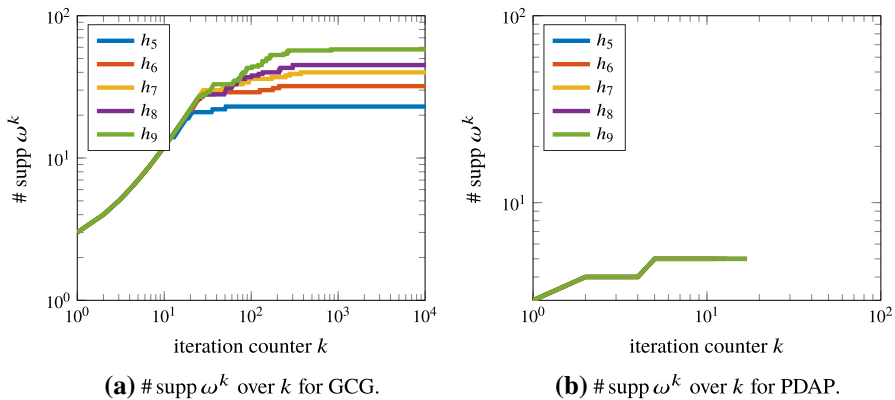
Last, we report on the computational time for the setup considered before, in order to account for the numerical effort of the additional point removal steps. The evolution of the residuals in the first second of the running time for GCG and SPINAT can be found in Fig. 5a. We observe that neither the additional projected gradient steps nor the additional application of Algorithm 2 lead to a significant increase of the computational time. For PDAP, the measurement times and residuals for all iterations are shown in Fig. 5b. We point out that PDAP converges after 12 iterations computed in approximately 0.4 seconds in this example. This is comparable to the elapsed computation time for computing 25 iterations of the GCG method. The small average time for a single iteration of PDAP is on the one hand a consequence of the uniformly bounded, low dimension of the sub-problem (4.16). On the other hand, using the intermediate iterate  $\omega^{k+1/2}$  to warm-start the semi-smooth Newton method greatly benefits its convergence behavior, restricting the additional numerical effort in of PDAP in comparison to GCG to the solution of a few low-dimensional Newton systems in each iteration. These results again underline the practical efficiency of the presented acceleration strategies.

### 5.1.4 Mesh-independence

To finish our numerical studies on Example 1 we examine the influence of the mesh-size  $h$  on the performance of Algorithm 1. We again consider the A-optimal design problem for  $\beta = 1$  and  $\mathcal{I}_0 = 0$  on consecutively refined meshes  $\mathcal{T}_{h_l}$ ,  $l = 5, \dots, 9$ . On each refinement level  $l$  the optimal design problem is solved using GCG and PDAP,



**Fig. 6** Evolution of residuals  $r_F(\omega^k)$  over iterations  $k$  on different refinement levels



**Fig. 7** Evolution of the support size on different refinement levels

respectively. The computed residuals are shown in Fig. 6. For both versions we observe that the convergence rate of the objective function value is stable with respect to mesh-refinement. We point out that this indicates a better convergence behavior of PDAP also on the continuous level. A theoretical investigation of this improved rate is beyond the scope of this work but will be given in a future manuscript. Additionally, in Fig. 7, we plot the support size over the iteration counter for each refinement level. For GCG we observe a monotonic growth of the support size up to a certain threshold. Note that the upper bound on the support size seems to depend on the spatial discretization: the finer the grid, the more clusterization around the true support points can be observed. In contrast, for PDAP, the evolution of the support size admits a mesh-independent behavior in this example.

## References

1. Adams, R.A.: Sobolev Spaces, Pure and Applied Mathematics. Academic Press, New York (1978)

2. Alexanderian, A., Petra, N., Stadler, G., Ghattas, O.: A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized  $\ell_0$ -sparsification. *SIAM J. Sci. Comput.* **36**, A2122–A2148 (2014)
3. Atkinson, A.C., Donev, A.N., Tobias, R.D.: *Optimum Experimental Designs*, with SAS. Oxford Statistical Science Series, vol. 34, Oxford University Press, Oxford (2007)
4. Atwood, C.L.: Sequences converging to  $D$ -optimal designs of experiments. *Ann. Stat.* **1**, 342–352 (1973)
5. Avery, M., Banks, H.T., Basu, K., Cheng, Y., Eager, E., Khasawinah, S., Potter, L., Rehm, K.L.: Experimental design and inverse problems in plant biological modeling. *J. Inverse Ill-Posed Probl.* **20**, 169–191 (2012)
6. Banks, H.T., Rehm, K.L.: Experimental design for vector output systems. *Inverse Probl. Sci. Eng.* **22**, 557–590 (2014)
7. Bates, D.M., Watts, D.G.: *Nonlinear Regression Analysis and Its Applications*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Wiley, New York (1988)
8. Bauer, I., Bock, H.G., Körkel, S., Schlöder, J.P.: Numerical methods for optimum experimental design in DAE systems. *J. Comput. Appl. Math.*, **120**, 1–25 (2000). SQP-based direct discretization methods for practical optimal control problems
9. Beale, E.M.L.: Confidence regions in non-linear estimation. *J. R. Stat. Soc. Ser. B* **22**, 41–88 (1960)
10. Becker, R., Braack, M., Vexler, B.: Parameter identification for chemical models in combustion problems. *Appl. Numer. Math.* **54**, 519–536 (2005)
11. Bock, H.G.: *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*, vol. 183 of *Bonner Mathematische Schriften* [Bonn Mathematical Publications], Universität Bonn, Mathematisches Institut, Bonn, 1987. Dissertation, Rheinische Friedrich-Wilhelms-Universität, Bonn (1985)
12. Bonnans, J.F., Shapiro, A.: *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, New York (2000)
13. Boyd, N., Schiebinger, G., Recht, B.: The alternating descent conditional gradient method for sparse inverse problems. *SIAM J. Optim.* **27**, 616–639 (2017)
14. Bredies, K., Lorenz, D.A., Maass, P.: A generalized conditional gradient method and its connection to an iterative shrinkage method. *Comput. Optim. Appl.* **42**, 173–193 (2009)
15. Bredies, K., Pikkarainen, H.K.: Inverse problems in spaces of measures. *ESAIM Control Optim. Calc. Var.* **19**, 190–218 (2013)
16. Brezis, H.: *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext, Springer, New York (2011)
17. Casas, E., Clason, C., Kunisch, K.: Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM J. Control Optim.* **50**, 1735–1752 (2012)
18. Chung, M., Haber, E.: Experimental design for biological systems. *SIAM J. Control Optim.* **50**, 471–489 (2012)
19. Dunn, J.C.: Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM J. Control Optim.* **17**, 187–211 (1979)
20. Dunn, J.C.: Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM J. Control Optim.* **18**, 473–487 (1980)
21. Elstrodt, J.: *Maß- und Integrationstheorie*. Springer-Lehrbuch, Springer, Berlin Heidelberg (2013)
22. Fedorov, V.V.: *Theory of Optimal Experiments*. Academic Press, New York-London (1972). Translated from the Russian and edited by Studden, W.J., Klimko, E.M., *Probability and Mathematical Statistics*, No. 12
23. Fedorov, V.V., Hackl, P.: *Model-oriented design of experiments*. Lecture Notes in Statistics, vol. 125. Springer, New York (1997)
24. Fedorov, V.V., Leonov, S.L.: *Optimal Design for Nonlinear Response Models*. Chapman & Hall/CRC Biostatistics Series, CRC Press, Boca Raton (2014)
25. Frank, M., Wolfe, P.: An algorithm for quadratic programming. *Naval Res. Logist. Quart.* **3**, 95–110 (1956)
26. Haber, E., Horesh, L., Tenorio, L.: Numerical methods for experimental design of large-scale linear ill-posed inverse problems. *Inverse Probl.* **24**, 055012 (2008)
27. Herzog, R., Riedel, I.: Sequentially optimal sensor placement in thermoelastic models for real time applications. *Optim. Eng.* **16**, 737–766 (2015)



28. Jaggi, M.: Revisiting Frank–Wolfe: projection-free sparse convex optimization. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, ICML'13, JMLR.org, pp. 1-427-1-435 (2013)
29. Kiefer, J.: General equivalence theory for optimum designs (approximate theory). *Ann. Stat.* **2**, 849–879 (1974)
30. Kiefer, J., Wolfowitz, J.: Optimum designs in regression problems. *Ann. Math. Stat.* **30**, 271–294 (1959)
31. Körkel, S., Bauer, I., Bock, H.G., Schlöder, J.: A sequential approach for nonlinear optimum experimental design in DAE systems. *Sci. Comput. Chem. Eng.* **II**(2), 338–345 (1999)
32. Milzarek, A., Ulbrich, M.: A semismooth Newton method with multidimensional filter globalization for  $l_1$ -optimization. *SIAM J. Optim.* **24**, 298–333 (2014)
33. Pázman, A.: Foundations of Optimum Experimental Design, vol. 14 of Mathematics and its Applications (East European Series). D. Reidel Publishing Co., Dordrecht (1986). Translated from the Czech
34. Pieper, K., Tang, B.Q., Trautmann, P., Walter, D.: Inverse point source location with the Helmholtz equation on a bounded domain (2018). [arXiv:1805.03310](https://arxiv.org/abs/1805.03310)
35. Pronzato, L.: Removing non-optimal support points in  $D$ -optimum design algorithms. *Stat. Probab. Lett.* **63**, 223–228 (2003)
36. Pukelsheim, F.: Optimal design of experiments, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, A Wiley-Interscience Publication, Wiley, New York (1993)
37. Rakotomamonjy, A., Flamary, R., Courty, N.: Generalized conditional gradient: analysis of convergence and applications. *ArXiv e-prints* (2015)
38. Rudin, W.: Real and Complex Analysis, 3rd edn. McGraw-Hill Book Co., New York (1987)
39. John, R.C., St. Draper, N.R.:  $D$ -optimality for regression designs: a review. *Technometrics* **17**, 15–23 (1975)
40. Tarantola, A.: Inverse Problem Theory and Methods for Model Parameter Estimation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2005)
41. Tröltzsch, F.: Optimal Control of Partial Differential Equations, vol. 112 of Graduate Studies in Mathematics. American Mathematical Society, Providence (2010). Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels
42. Uciński, D.: Optimal Measurement Methods for Distributed Parameter System Identification. Systems and Control Series, CRC Press, Boca Raton (2005)
43. Ulbrich, M.: Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.* **13**(2002), 805–842 (2003)
44. Wolfe, P.: Convergence Theory in Nonlinear Programming. North-Holland, Amsterdam (1970)
45. Wynn, H.P.: The sequential generation of  $D$ -optimum experimental designs. *Ann. Math. Stat.* **41**, 1655–1664 (1970)
46. Yu, Y.:  $D$ -optimal designs via a cocktail algorithm. *Stat. Comput.* **21**, 475–481 (2011)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Ira Neitzel<sup>1</sup> · Konstantin Pieper<sup>2</sup> · Boris Vexler<sup>3</sup> · Daniel Walter<sup>3</sup>

Ira Neitzel  
neitzel@ins.uni-bonn.de

Konstantin Pieper  
kpieper@fsu.edu

Boris Vexler  
vexler@ma.tum.de

<sup>1</sup> Institut für Numerische Simulation, Universität Bonn, Wegelerstr. 6, 53115 Bonn, Germany

- 
- <sup>2</sup> Department of Scientific Computing, Florida State University, 400 Dirac Science Library, Tallahassee, FL 32306, USA
- <sup>3</sup> Center for Mathematical Sciences, Chair M17, Technische Universität München, Boltzmannstr. 3, 85748 Garching bei München, Germany